Barcelona School of Economics

Master Program in Data Science for Decision Making

"Corpus Construction and Social Media Analysis about Immigration in Chile"

Andrés Couble, Mathias Schindler and Kalliope Stassinos

Supervisors: Jesús Cerquides and Hannes Mueller

Date: June 2022

ABSTRACT IN ENGLISH

This thesis presents a general-purpose corpus construction methodology with Twitter data for a given political topic in a given country. It applies the methodology to immigration in Chile from November 2021 to April 2022, resulting in a corpus with 573,999 tweets. Our results indicate increasing anti-immigration views from Chilean Twitter users. Right-leaning users are more active and more anti-immigration. Left-leaning users are mostly concerned with xenophobia and racism.

Utilizing network analysis methods, we find that right-leaning users are also more influential and interconnected. The results are consistent with previous studies and the methodology is robust to other political topics such as feminism.

ABSTRACT IN SPANISH

Esta tesis presenta una metodología de construcción de corpus con datos de Twitter para un tema político dado en un país dado. Aplicamos la metodología al tema de inmigración en Chile desde noviembre de 2021 hasta abril de 2022, resultando en un corpus con 573.999 tuits. Nuestros resultados indican un aumento de las opiniones contra la inmigración de los usuarios chilenos de Twitter. Los usuarios de derecha son más activos y antinmigración. Los usuarios de tendencia izquierdista se preocupan principalmente por la xenofobia y el racismo.

Utilizando métodos de análisis de red, encontramos que los usuarios de derecha también son más influyentes e interconectados. Los resultados son consistentes con estudios previos y la metodología fue testeada para otros temas políticos de interés como el feminismo.

KEYWORDS IN ENGLISH: Twitter, Immigration, Corpus Construction

KEYWORDS IN SPANISH: Twitter, Inmigración, Construcción de Corpus

Corpus Construction and Social Media Analysis about Immigration in Chile

written by

Andrés Couble, Mathias Schindler, Kalliope Stassinos

A thesis submitted in partial fulfillment of the requirements for the degree

Master's Degree in Data Science

in the

Data Science for Decision Making Program, Class of 2022

at the

Barcelona School of Economics

June 2022

Corpus Construction and Social Media Analysis about Immigration in Chile*

Andrés Couble[†] Mathias Schindler[‡] Kalliope Stassinos[§]

Submitted: June 28, 2022

Abstract

This thesis presents a general-purpose corpus construction methodology with Twitter data for a given political topic in a given country. It applies the methodology to immigration in Chile from November 2021 to April 2022, resulting in a corpus with 573,999 tweets. Our results indicate increasing anti-immigration views from Chilean Twitter users. Right-leaning users are more active and more anti-immigration. Left-leaning users are mostly concerned with xenophobia and racism. Utilizing network analysis methods, we find that right-leaning users are also more influential and interconnected. The results are consistent with previous studies and the methodology is robust to other political topics such as feminism.

Future improvements could include more advanced classification algorithms for political affiliation and bot detection. Practitioners using our social listening tool should be aware of the general misrepresentation of Twitter users in regards to a general population.

Keywords: Twitter, Corpus Construction, Politics, Network Analysis, Reproducibility, Chile, Immigration

[*] We thank tenured researchers Jesus Cerquides (IIIA-CSIC) and Hannes Mueller (IAE-CSIC and BSE) for supervision and suggestions. We also thank Claudio Villegas Oliva from the Chilean Communication Office for giving us our thesis challenge and providing regular feedback. Finally, we thank PhD-students Elliot Motte (Universitat Pompeu Fabra) and Luis I. Menéndez García (Universitat Autònoma de Barcelona) for excellent technical assistantship. All our codes are available at our GitHub repository: `https://github.com/BSE-DSDM-2022/ChileGov`. Any potential errors that remain are our own.

[†] Barcelona School of Economics; email: andres.couble@bse.eu.

[‡] Barcelona School of Economics; email: mathias.schindler@bse.eu.

[§] Barcelona School of Economics; email: kalliope.stassinos@bse.eu.

Contents

List of Figures

List of Tables

1 Introduction

Digital communication is becoming increasingly important everyday. In fact, social media platforms such as Twitter can provide information in real-time about the opinions of a population's subgroups. Developing social listening tools that can extract and clean social media data can therefore be helpful for political actors to gauge public opinion regarding particular issues or implemented policies. They can thereby complement traditional telephone- or survey-based opinion polls at higher frequency and lower costs.

The work in this thesis is motivated by a task given to us by the Chilean Communications Office (henceforth, CCO) which was stated as follows

> *"Develop a methodology to analyze the Chilean conversation on Twitter on a specific topic, considering the political affiliation of the users. Describe the associated narratives that appear over time and the network structure of the groups involved. Apply the developed methodology using immigration as the test subject."*
>
> — *(See Appendix A.1 for original text)*

Our contribution to the CCO and the literature in general is two-fold. First, we build a general-purpose Twitter corpus construction methodology with conversations about a given topic in a specified time frame distinguishing between left- and right- wing users. Second, we apply this methodology to provide descriptive evidence as to how political affiliation shapes the online conversation about immigration in Chile in order to answer the questions: What are the main concerns of Chilean Twitter users regarding immigration? Which are the main differences between left-leaning and right-leaning discourses? Who are the most influential users in this conversation? Generally we find that over a one and a half year period until April 2022, the discourse in the Chilean Twittersphere becomes more focused on anti-immigration and that this trend is particularly prevalent among politically right-leaning Twitter users. We also find that right-leaning users are more active, have a stronger community and generally have a higher capacity to push certain talking-points on immigration in Chile. Our findings are generally consistent with previous studies, while also providing new insights.

We reach these insights by analyzing a corpus of tweets written by Chilean users from November 2020 to April 2022 consisting of 573,999 tweets. The corpus is built by adapting the general-purpose methodology, specifying keywords and hashtags pertaining to immigration in Chile. To extrapolate meaningful insights from the data, we build a custom Python library and an interactive dashboard tailored to the corpora resulting from our methodology. Utilizing these tools, we analyze metrics such as the most-used words, hashtags and bigrams as well as measures from network analysis.

First, we analyze the entire time period. By analyzing a political affiliation-labeled subsample of 216,245 tweets, we find that right-leaning users post close to thrice as many posts per user as left-leaning ones. Left-leaning users primarily push anti-xenophobia agendas while right-leaning users mostly talk about undocumented immigrants and blame these for crime. We also find that a small number of right-leaning users who aggressively try to push an agenda relating immigration with terrorism (which is mostly carried out by nationals in Chile) and blame the left for both issues.

Second, we analyze how conversations were shaped during a violent anti-immigration protest in September 2021. The pattern that right-leaning users mostly talk about illegal immigration and left-leaning users talk mostly about xenophobia and racism is still present in this period. Also we found that right-leaning users used this protest as a pretext to campaign and position their candidate in the discussion. In contrast, left-leaning users did not link the protest with their candidate's campaign and the main political issue that they pushed was to blame the previous right-wing government for the migration crisis. An interesting insight appears while looking at the group of users that are neither classified left- or right-leaning. These raised an anti-UN campaign, linking a UN agenda with the immigration crisis. We again find a small subgroup of users, pushing these agendas aggressively.

Finally, using network metrics, we show that the retweet network of left-leaning and right-leaning users differ substantially: Right-leaning users are more interconnected and more active than left-leaning users. The former right-leaning presidential candidate José Kast was the most influential account while certain media outlets were also influential. These findings are consistent with the text analysis. All together this indicates that right-leaning users dominate the Twitter conversation around immigration in Chile, led by the former candidate Kast.

We believe our work to be immensely useful for governments and political institutions in general. Despite Twitter users generally not being representative of a whole population, we believe the benefits of quickly monitoring opinions around certain topics at a low cost outweigh the disadvantage of analyzing conversations of a not representative subsample of the whole population. Knowing the most influential accounts as well as the agendas and concerns from both sides of the political spectrum can help inform to design better targeted communication strategies from political institutions.

The rest of this thesis is structured as follows: Section 2 goes through previous, relevant studies; Section 3 present our general-purpose methodology; Section 4 presents results regarding immigration in Chile; Section 5 discusses the results and future improvements. The codes for the social listening tool developed in this thesis are available at our GitHub repository under the following url: `https://github.com/BSE-DSDM-2022/ChileGov`.

2 Previous Studies

In recent years, Twitter data has received increased attention in academia. Researchers are attracted to data from the microblogging platform because of its simple structure with short and frequent interactions between users and the easy access to data that the Twitter API provides. Tufekci (2014) and Barberá and Rivero (2015) show that Twitter users are not representative of the whole population and this can potentially lead to biased results. However, as long as researchers are aware of these limitations, analyzing Twitter data can still yield relevant insights.

The literature around political conversations on Twitter is already vast. Jungherr (2016) provides an extensive systematization of the literature around this topic up to 2016. The paper finds that the use of the Twitter API is common in these studies and specifically the use of hashtags (34 of 127 studies analyzed) or keywords (26 of 127 papers) as a criteria for identifying topics of a conversation. These studies select specific events where keywords or hashtags are easy to identify: E.g. Lin et al. (2014) focus on debates or party conventions that had a commonly known hashtag to identify the event (e.g the hashtag `#debate` during presidential debates), while Himelboim et al. (2017) select a long list of different topics, where each one is represented by a single word or hashtag (e.g the hashtag `#OHSen` used to discuss the Senate race in Ohio or the word "Autism" to identify people who talk about this neurodevelopmental condition). We are providing a methodology to collect data about broad topics that need more than a single word of hashtag, in order to retain more information than the previously mentioned one-word/-hashtag studies.[1] In general, these methods require domain knowledge for choosing which words or hashtags pertain to a given topic and, to the best of our knowledge, there have not yet been developed computational algorithms that can outsource this task from human input.

Recent research with Twitter data include hate speech detection (Plaza-del Arco et al., 2021; Basile et al., 2019; Pereira-Kohatsu et al., 2019), sentiment analysis (Agarwal et al., 2011; Saif et al., 2012) and feature engineering such as extrapolating Twitter user's age, gender and political affiliation (Conover et al., 2011b; Pennacchiotti and Popescu, 2011; Kruspe et al., 2021). Identifying Twitter users' political affiliation is particularly relevant for our work because we base our political affiliation identification in part on the methodology developed in Rao et al. (2010), considering the use of hashtag as the main criteria to identify affiliation. Our methodology also contributes to the literature on feature engineering on Twitter data, as we propose a novel methodology to identify Twitter users' nationality. To the best of our knowledge, feature engineering nationality of Twitter users has not been done before, and we believe this to be fruitful in future research across multiple topics utilizing Twitter data.

In terms of network analysis, using graph theory can be relevant for analyzing information flows. Twitter data contains various types of interactions, the most prevalent ones being "retweets", "likes", "mentions" and "follows". Conover et al. (2011a) compares retweets and mentions networks, finding that political retweets exhibit a highly segregated partisan structure. On the other hand, the network of mentions is dominated by a single politically heterogeneous cluster. Following these conclusions, for analyzing information diffusion networks and identifying clusters, retweets are most appropriate. Once the network is correctly built, various metrics can be used to characterize the network. Common metrics include volume, influence and density. Maharani et al. (2014) use retweets as links between users and identify patterns among influential users using degree and eigenvector centrality. We replicate these methods in the immigration retweets network.

Studies about Chilean's general perception towards immigration are scarce. One such study is González et al. (2019) which analyzes attitudes towards immigration and their relationship with social diversity between 2002 and 2017. Some interesting findings are the differences in attitudes when considering the nationality of immigrants and also that people who self-identify with right-political positions have slightly higher anti-immigrant attitudes compared to people who self-identify with center or left-leaning political positions. The study also shows data from polls that affirm the existence of a growing concern about illegal immigration.

Gálvez et al. (2020) uses Chilean Twitter data to analyze discriminatory message against immigrants between January 2018 and August 2020. The study finds that Twitter activity and discriminatory speech are sensitive to public immigration-related events and that discriminatory speech mainly originates from far-right and nationalist users. The study also finds that immigration is used as a topic to attack certain political figures.

[1] For instance, if in our case we used only the word "immigration" we would lose information about people who do not mention this word, but only mention Venezuelans immigrating to Chile. So, we should add the word "Venezuelans" in order to lose less information.

3 Methodology

This section describes a methodology to analyze Twitter conversation on a specific political topic over a given time frame in a given country.

3.1 Twitter API and `twarc2`

Twitter generously provides access for researchers to the full collection of tweets that are currently published on the microblogging website through their *Twitter API for Academic Research* service.[2] After being accepted through the company's application procedure, researchers are provided with an allowance of up to 10 M tweets per month to download. To archive Twitter data in `.json`-file format, we utilize the `twarc2` Python-based command line tool. `twarc2` allows users to specify the corpus of tweets to download by writing queries specifying e.g. keywords to filter by, hashtags, retweets, time frames, locations, etc. After downloading, the resulting datafile also includes metadata for each tweet such as number of likes, retweets, author characteristics, among others.

Unfortunately, the usage of geotagged tweets has declined from 2012 to 2022 following an announcment in mid-2019 that Twitter would remove the option to attach precise geotagging to tweets (Kruspe et al., 2021). This complicates the task of building a country-specific corpus. Solely obtaining geotagged tweets would not yield a sample of sufficient size and raise concerns about the resulting sample being too biased.[3] To address this issue we develop a novel methodology to filter non-geotagged tweets to citizens of a given country in Step 3 of Section 3.2.

3.2 General Methodology

3.2.1 Corpus Construction

This presents our methodology in general terms which allows to construct a Twitter corpus that can be used to analyze any political topic of choice in any given country. The steps below are described briefly. For a more extensive description of the methodology see Appendix A.2. Table 1 presents a general overview of the entire methodology. This table might be of particular interest to practitioners.

Step 1: Exploring Topical Semantic Links as Search Keywords + `twarc2` Query 1

Enter the word for the topic of choice into a platform such as `http://semantic-link.com/`. Manually select the most relevant words and store the list L_{sem}. Run the first `twarc2` query using the script 1.1 of the methodology, specifying keyword list L_{sem}, time frame and geotag the country of interest.

Step 2: Adding Keywords from Twitter Contextual Data + `twarc2` Query 2

To uncover the most relevant keywords, explore the 200 most common words and hashtags in the corpus from Step 1. Manually build a list of keywords related to the topic, L_t. Manually build another list of keywords related to the country of interest (e.g. relevant cities, president's surname, relevant politicians, etc.), L_c. Run two `twarc` queries: 1) Tweets with keywords from L_t and geolocated in the country of interest; 2) Tweets with keywords from $L_t + L_c$.

[2] The methodology also works with Elevated Access Twitter API, free provided for non-academic projects.

[3] Appendix Figure A.1 shows the decline in geotagged tweets in Chile.

This way, we obtain a broader corpus that should correspond better to topic and location of interest.

Step 3: Filtering the Authors by the Country's Location

To filter the authors by location, we develop a novel methodology. We store a list of all the authors who's self-written account location or biography (or both) contains at least one of the following:

1. The country's flag as an emoji

2. The country as a regular expression – including derivations thereof such as demonyms

3. The unambiguous cities of the country either as an n-gram or unigram[4]

This way we obtain a list primarily made up of citizens of the country of interest, present in the topic studied.[5]

Step 4: Recovering Topic-Related Tweets from the Country's Citizens + `twarc2` Query 3

After obtaining the list of relevant authors from Step 3, this step downloads all tweets in relation to the topic's discussions. This way we obtain the tweets regarding the topic of interest, for the studied time frame, written by the country's nationals or individuals located in the country.

Step 5: Filtering the Final Corpus by Topic

Following Conover et al. (2011b) and Small (2011) we propose a word- and hashtag-based filtering approach to clean the corpus such that it mainly consists of tweets regarding the topic of interest.[6] Looking at tweets not related with the topic, we drop the prominent non-relevant hashtags and words that create noise in the data. The output from Step 5 is the final corpus for analysis.

Suggested Validation After Step 5 we suggest reviewing a random sample of tweets in the corpus, to ensure that it mainly contains tweets from the country of interest and related to the topic. Otherwise, try to find words or hashtags to delete noise or try new searches changing the keywords.

3.2.2 Labeling by Political Affiliation

For our political discussion analysis, distinguishing between different trending political topics between left and right-leaning Twitter users is of great interest.

[4] By 'unambiguous' we refer to city names in the country which are not also a city or country name in another country. This would disqualify cities names such as*'Florida'*, *'El Salvador'*, and so forth.

[5] We deliberately use the wording *'primarily'*, as we do not want to give the impression that we are perfectly able to distinguish the nationality or location of Twitter users. We are however, confident that our proposed methodology yields good results, given the various verification and exploration analysis we have performed while testing the methodology on immigration in Chile .

[6] This step could also have been approached in other ways such as unsupervised machine learning algorithms to determine and filter topics (Hong and Davison, 2010; Zhao et al., 2011; Cataldi et al., 2010). However, in applications this would involve running various models for each topic of interest to determine whether a model performs sufficiently well, which necessitates employing statisticians to understand and evaluate model accuracy. For this reason, we believe utilizing unsupervised learning algorithms for this step does not meet the criteria that the developed methodology should be *"easy-to-use"* nor *"low-cost"*. It would also increase computational execution time. Given the importance of this step in corpus construction, we believe our proposed solution to this step with words and hashtags is the most appropriate for this specific objective.

Step 6: Classify Hashtags by Political Affiliation into Right- And Left-Leaning

Following Rao et al. (2010), we consider political hashtags as a proxy of political affiliation. Construct a list of right-wing and left-wing politicians. Download all the politicians' tweets in an appropriate time frame (e.g. an electoral period).[7] Identify the most-used hashtags by left- and right-wing politicians, respectively.

Step 7: Label Users by Political Affiliation

To label the Twitter users from the main corpus, we recover their tweets that used political hashtags during the same period used in Step 6. Then, we construct a criterion to classify the nodes as 'left' or 'right' (and 'unlabeled' when not possible to classify) based on the frequency of hashtags found in their tweets. When higher thresholds are chosen, accuracy is higher for the left-right classification, but unlabeled becomes more heterogeneous (including neutral, as well as many left and right).[8]

This way, we obtain labels of political affiliation classification for the users in our corpus.

3.2.3 Network Construction

In graph theory, a network is a collection of nodes (e.g. individual Twitter users) connected by edges (interactions such as retweets, likes, comments, etc.). When looking at information flows between users, retweets are the most relevant interactions to study.

Step 8: Construct Retweet Network

To create the network of retweets for the selected topic we download all users' retweets in relation to the discussion, during the period studied.[9] Then, we create a directed and weighted network where each node is one user of our final data set and each edge is a retweet, going from A to B and weighted by the number of times that A retweeted B.[10]

3.2.4 Updating the Corpus

The methodology also includes scripts to update the information. (for instance, it is possible to update the information weekly, using the same keywords). The scripts repeat Steps 2, 3, 4 and 5 for the new dates, repeat Step 7 to politically label new users and Step 8 to add nodes and links with the updated list of users. The final output is the data set with tweets, the data set of labels and the updated network.

[7] We suggest to consider the use of hashtags not directly related with the topic of interest to avoid bias. For instance, implementing a affiliation classification and then comparison on the same set of hashtags is problematic.

[8] The current hashtag frequency threshold is high, to maximize accuracy of political patterns analysis. But the practitioner can easily choose a different criterion.

[9] For this, we go to `twarc` to download all the retweets from our list of users that contain one of the keywords during the selected period.

[10] For this, we first use the plugging `twarc`-network to obtain a data frame containing: user A that sent the retweet, user B that received the retweet and the total number of retweets that A gave to B in the corpus. Looping through this information, we construct a `DiGraph` object from the `NetworkX` package.

Table 1: Corpus Construction Methodology Overview

Step	Notebooks, Initial	Notebooks, Updating	Manual Input	Output
1	1.1_First_Query_to_Twarc	Not required	· Filtered keywords related with the topic of interest from http://semantic-link.com/	Exploratory corpus with tweets containing semantic linked topic-related words and geolocated in country
2	1.2_Adding_words_looking_the_Chilean_context	1.1_Downloading_tweets_for_new_period	· List of keywords related with the topic after the contextual review · List of keywords related with the country	Corpus with tweets that contain topic-related keywords (semantic link and contextual) and are geolocated or contain country keywords
3	1.3_Filter_by_Chileans	1.2_Filter_by_Chileans	· Emoji code of the countries' flag · Country name as a regular expression · List of cities of the country (code snippet to download already pre-programmed; insert URL-link of relevant Wikipedia entry of list of cities in country)	List of Twitter users that are citizens or located in the country and are tweeting about the topic of interest
4	1.4_Download_all_related_tweets_from_local_authors	1.3_Download_all_related_tweets_from_local_authors	· Not required	Corpus with tweets that contain topic-related keywords (semantic link + country-contextual) and are tweeted by authors that are from the country of interest
5	1.5_Filter_by_Topic_and_Cleaning_Text	1.4_Filter_by_Topic_and_Cleaning_Text	· Hashtags that introduce noise · Keywords that introduce noise	Corpus with tweets that contain topic-related keywords (semantic link + country contextual) and are tweeted by authors that are from the country of interest. Corpus contains mainly tweets about the topic of interest. Output is final cleaned corpus.
6	2.1_Recovering_Hashtags_From_Politicians	Not required	· List of right-wing and left-wing politicians · Period of time to analyze the use of general hashtags	Most common hashtags used by right-wing and left-wing politicians during the selected period
7	2.2_Downloading_Left_Right_Hashtags_To_Label 2.3_Labeling_Left_Right_Users	2.1_Downloading_Left_Right_Hashtags_To_Label 2.2_Labeling_New_Users_and_Updating_Data_Sets	· Filtered list of right-wing and left-wing hashtags	Users that talk about the topic of interest labeled according to political affiliation
8	3.1_Download_Retweets_to_Create_Network 3.2_Create_the_RT_network	3.1_Download_Retweets_to_Create_Network 3.2_Update_the_RT_network	· Not required	Retweets network between the list of users

Notes: Notebooks can be found in our GitHub repository under the URL: https://github.com/BSE-DSDM-2022/ChileGov.
Source: Authors' own construction.

3.3 Analysis Tools

In order to ease the exploratory process of analyzing the resulting corpus from our methodology, we have extended the challenge from the CCO and built various analysis tools for quantitative descriptive research. We have built a custom Python library named `TextAnLib` and an interactive dashboard.

3.3.1 `TextAnLib` Python Package

We built a Python library which is tailored to analyzing the data in the resulting corpora from the methodology. The main functions of the library are two-fold:

- Filtering: Using the functions, the data can be easily split by features such as time periods, political affiliation, specific hashtags, verified accounts among others.
- Visualizing: Data visualization functions such as word clouds of top words, bar plots of top hashtags, comparison of use of words between left- and right-leaning users among others.

See Appendix B for an extensive documentation of each of the functions in the `TextAnLib` library.

3.3.2 Interactive Dashboard

The Python packages `Plotly` and `Dash` allow programmers to build user-friendly interactive dashboards. Given the coding complexity of building such dashboards, our product should be viewed as a first prototype and proof-of-concept.

One advantage of interactive dashboards is that it allows quick and easy data visualization, and hereby extends the user base not just to data scientists and statisticians but also to more non-technical staff who might not be familiar with coding.[11] E.g. we imagine communication graduates not to be very familiar with coding but to still be interested in tracking Twitter conversations on a daily basis. An interactive dashboard is by no means a novel invention, but we still believe it improves our final product for the CCO.

Figure A.3 shows a static screenshot for our first prototype of the dashboard.[12] The dashboard makes use of the functions from our custom `TextAnLib`-package from Section 3.3.1. As can be seen from the figure, the dashboard allows to visualize:

- Daily tweet counts
- Top hashtags
- Top authors
- Top words

[11] Hence, providing a custom Python library might not be useful for them as the cost of learning might be too high for non-technical practitioners.

[12] We would have liked to upload our dashboard to a webpage, to ease demonstration purposes. This is quite easy with the `shiny`-package for `R`, but not so straightforward with `Plotly` and `Dash` and is outside of the scope of this thesis. The reason for using `Plotly` and `Dash` is that these are Python-based, i.e. in the same programming language as our `TextAnLib`-package.

When the users clicks and selects the desired metric and time frame, the app executes the code in the background and updates the UI in real-time. With more time on our hands we imagine we could build a significantly more complex app, such that it could show much more information. We believe this app to be a very useful UI for non-technical users. As mentioned, our constructed dashboard is only a proof-of-concept. Section 5.2 addresses thoughts on how it could be improved further and thereby deliver more value for the CCO.

4 Results

This section presents how our developed methodology of corpus construction can be utilized by political actors in practice to gauge public opinion. Because Chile has recently seen an uptick in immigrants (375,388 arrivals in 2010 (Datos Macro, 2010), compared to 1,462,103 in 2020 (Instituto Nacional de Estadísticas, 2020)) and multiple violent anti-immigration protests in recent years, it could be hypothesized that anti-immigration sentiment is on the rise in Chile. Further, it might be the case that right-leaning users hold stronger anti-immigration views while left-leaning users are more embracing of immigration. Our social listening tool can then be used to analyze whether this is the case in Chileans' online conversations about immigration.

As a general political context of the analyzed period, the previous Chilean government was right-wing. The most recent election was held on December 19th, 2021, and elected Gabriel Boric as President of Chile. Boric is affiliated with a leftist party while the other presidential candidate in the second round was José Kast, who is affiliated with a far-right party. Immigration was one of the relevant topics discussed during the campaign, particularly pushed by the candidate Kast. Boric took office on March 11th, 2022.

The corpus in this section is constructed by applying the methodology in Section 3.2 with the following characteristics:

- Country: Chile
- Topic: Immigration
- Time frame: November 1st, 2020 to April 11th, 2022

The time frame is chosen to include periods before and after the last presidential election and three significant local events regarding immigration: One mass deportation of immigrants on February 10th, 2021 and two violent anti-immigration protests in northern border cities on September 26th, 2021 and January 29th, 2022, respectively. A detailed description of how we applied our methodology to immigration in Chile can be found in Appendix A.2.

4.1 General Overview

Descriptive Measures To provide a general overview of the main corpus resulting from our methodology's Step 5, Table 2 presents simple textual characteristics. The corpus consists of 573,999 tweets from 45,525 distinct Twitter users. Appendix Figure A.8 shows four random tweets from the total corpus.

Table 2: Corpus Characteristics, Total Corpus

Measure	Count
Number of Tweets	573,999
Unique Authors	45,525
Unique Words	346,600
Unique Hashtags	24,266

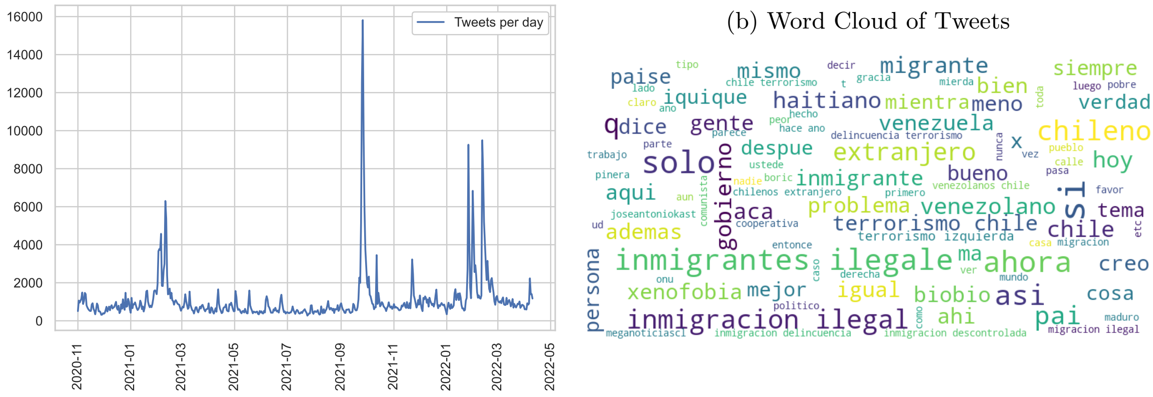
Data Source: Retrieved from Twitter API, spanning the time frame Nov 1st, 2020 – April 11th, 2022. Data is cleaned by our proposed methodology, such that the corpus includes tweets with topic-related keywords (semantic link + country contextual) and are tweeted by Twitter users in Chile or Chilean nationals and contains tweets that mainly regard the topic of immigration.

The tweets' distribution over time is presented in Figure 1a from which three peaks are evident. The first peak is in February 2021 and coincided with a mass deportation of immigrants. The second and third peak occur in late September 2021 and late January 2022, respectively. These two peaks coincided with two violent anti-immigration protests which took place in the northern border city of Iquique.

Out of the 346,600 distinct words in the corpus, Figure 1b presents the most-used ones in a word cloud. As expected, it is seen that the keywords used for construction of the corpus appear prominent, for instance *'inmigracion*, *'inmigrantes*, and *'venezolano'*. There also appears some relevant words that are strongly related with the legal situation of immigrants and the influx of immigrants, for instance *'ilegales'*, *'ilegal'* and *'descontrolada'*. Also, the word cloud shows some common bigrams related to illegal situation like *'inmigrantes, ilegales'* (Eng: Illegals, immigrant) or *'migracion, ilegal'* (Eng: Illegal, immigration). Interestingly, the term *'terrorismo'* is prevalent and also features in multiple of the most-used bigrams in Figure 2b. Terrorism would not normally be considered connected to immigration in Chile. We analyze this finding further on page 12.

Figure 1: Tweet Count and Word Coud, Total Corpus

(a) Tweets per Day

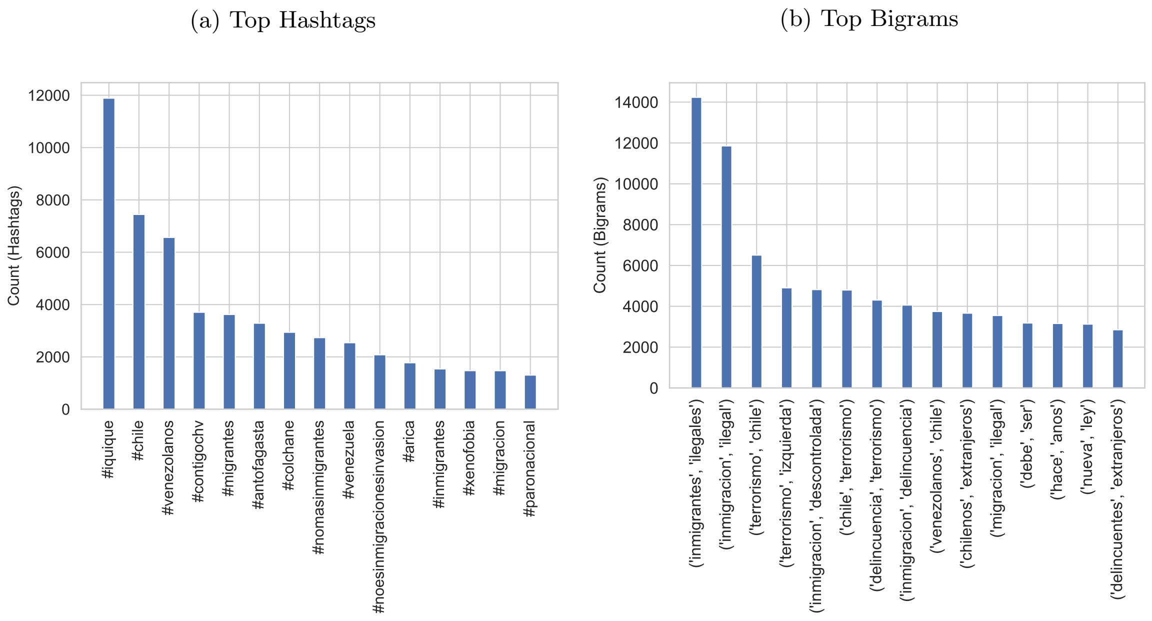

(b) Word Cloud of Tweets

Notes: Textual preprocessing steps in outputs include lowercase enforcement and converting Spanish special characters (e.g. á, ñ, etc.) into corresponding non-accentuated ones.

Data Source: Retrieved from Twitter API, spanning the time frame Nov 1st, 2020 – April 11th, 2022. Corpus includes tweets by Twitter users in Chile or Chilean nationals that mainly pertain to immigration. Corpus cleaned by the steps in our methodology as described in Section 3.2 and Appendix A.2.

Some of the most-used hashtags in the corpus refer to northern Chilean border cities such as `#iquique`, `#antofagasta` and `#arica` as is seen from Figure 2a. This indicates that Chilean Twitter users are concerned about immigrants entering from the norhern border cities. This is further supported by the prominence of the hashtags `#venezolanos` (Eng: Venezuelans) and `#venezuela` as Venezuelan immigrants typically enter Chile from the north. To gauge whether Twitter users talk about immigrants in a negative or positive way, we consider some of the other most-used hashtags. Some hashtags are emotionally neutral such as `#inmigrantes` (Eng: Immigration) or `#migracion` (Eng: Migration). However, some hashtags such as `#nomasinmigrantes` (Eng: No more immgirants) and `#noesimmigracionesinvasion` (Eng: It's not immigration, it's invasion) carry strong negative connotations towards immigrants. We find further descriptive evidence of anti-immigratory agendas by looking at the most-used bigrams in Figure 2b. Here we find talking points of immigration being illegal and related to crime from the bigrams *'inmigrantes, ilegales'* (Eng: Immigrants, illegals), *'inmigracion, descontrolada'* (Eng: Migration, uncontrolled) and *'inmigracion, delincuencia'* (Eng: Migration, crime). However, we also see signs that some Twitter users try to call out the negative speech by highlighting the xenophobic elements of the general talking points with the hashtag `#xenofobia`. Hence, we find a general pattern of the Chilean Twittersphere being against migration, specifically concerned about illegal immigration and the crime it supposedly brings with it. Although some users seem to be against xenophobia.

Figure 2: Top 15 Hashtags and Bigrams, Total Corpus


(a) Top Hashtags

(b) Top Bigrams

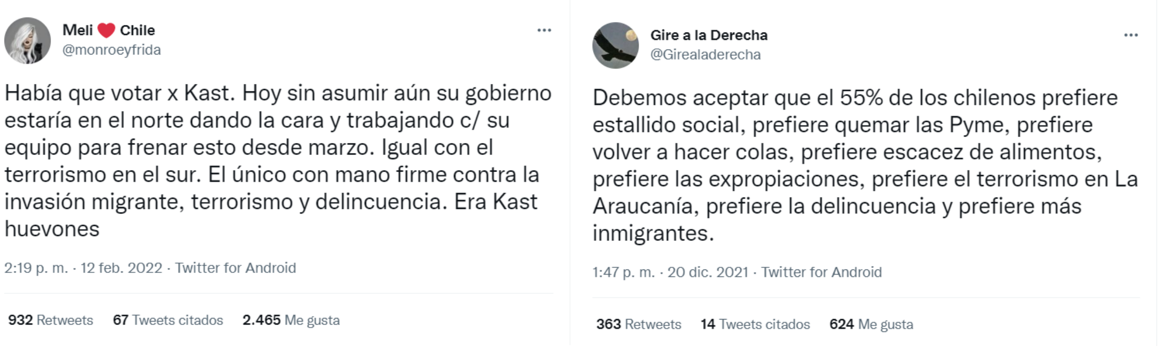
Notes: Textual preprocessing steps in outputs include lowercase enforcement and converting Spanish special characters (e.g. á, ñ, etc.) into corresponding non-accentuated ones.

Data Source: Retrieved from Twitter API, spanning the time frame Nov 1st, 2020 – April 11th, 2022. Corpus includes tweets by Twitter users in Chile or Chilean nationals that mainly pertain to immigration. Corpus cleaned by the steps in our methodology as described in Section 3.2 and Appendix A.2.

Immigration $\overset{?}{=}$ Terrorism Immigration and terrorist attacks are generally considered separate issues in Chile, whereby it is surprising how often the term *'terrorismo'* (Eng: Terrorism) features in the metrics presented.[13] To investigate, Figure 3 shows the two most retweeted tweets mentioning *'terrorismo'* in our corpus.

[13] Contemporary terrorism in Chile is mostly performed by groups supporting rights for indigenous inhabitants in the southern part of the country. Hence these attacks are by no means connected to immigration.

Figure 3: Most Retweeted Tweets that Mention Terrorism



(a) Most Retweeted

Meli ❤ Chile
@monroeyfrida

Había que votar x Kast. Hoy sin asumir aún su gobierno estaría en el norte dando la cara y trabajando c/ su equipo para frenar esto desde marzo. Igual con el terrorismo en el sur. El único con mano firme contra la invasión migrante, terrorismo y delincuencia. Era Kast huevones

2:19 p. m. · 12 feb. 2022 · Twitter for Android

932 Retweets 67 Tweets citados 2.465 Me gusta

(b) Second-Most Retweeted

Gire a la Derecha
@Girealaderecha

Debemos aceptar que el 55% de los chilenos prefiere estallido social, prefiere quemar las Pyme, prefiere volver a hacer colas, prefiere escacez de alimentos, prefiere las expropiaciones, prefiere el terrorismo en La Araucanía, prefiere la delincuencia y prefiere más inmigrantes.

1:47 p. m. · 20 dic. 2021 · Twitter for Android

363 Retweets 14 Tweets citados 624 Me gusta

Notes: Tweet 3a in English (Google Translate): "You had to vote for Kast. Today without assuming his government yet he would be in the north showing his face and working with his team to stop this from March. Same with him terrorism in the south. The only one with a firm hand against the migrant invasion, terrorism and crime. Was Kast."
Tweet 3b in English (Google Translate): "We must accept that 55% of Chileans prefer social outburst, prefers to burn SMEs, prefers queuing again, prefer food scarcity, prefers expropriations, prefers terrorism in La Araucanía, prefers crime and prefers more immigrants."

Data Source: Screenshots from Twitter. Retrieved from Twitter API, spanning the time frame Nov 1st, 2020 – April 11th, 2022. Corpus includes tweets by Twitter users in Chile or Chilean nationals that mainly pertain to immigration. Corpus cleaned by the steps in our methodology as described in Section 3.2 and Appendix A.2.

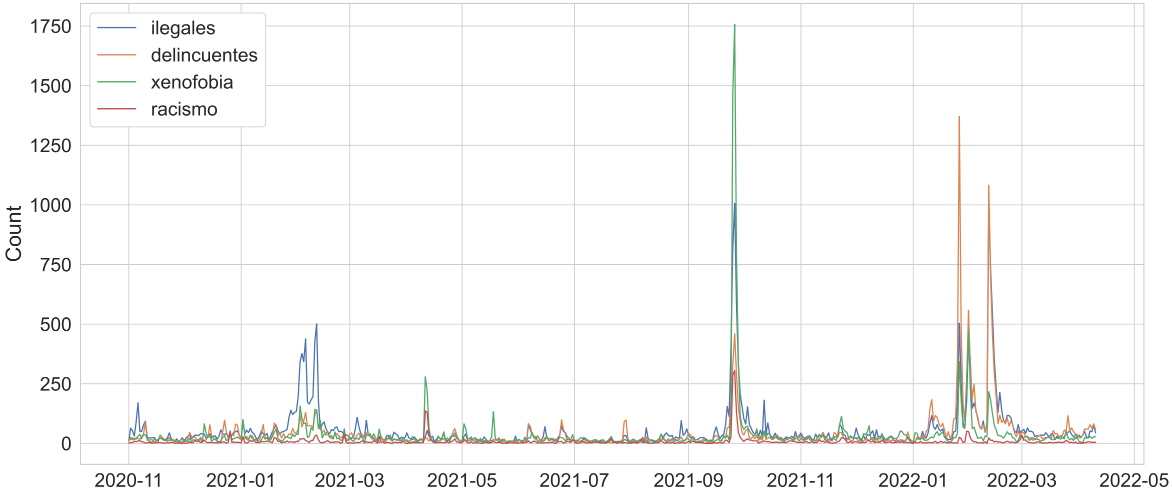
Interestingly, the tweets do not directly relate terrorism with immigration. Rather, they mention the two separate issues together and state that leftist parties and supporters do not address these issues adequately. Hereby we seem to have found a prominent discourse from right-leaning Twitter users: They try to push the agenda that the left is responsible for these separate issues and by equating them paints the left in a negative light. Analyzing the ten most retweeted tweets mentioning terrorism is in line with this finding.

Appendix Tables A.2 and A.3 show the number of users that used the top 15 hashtags and bigrams, respectively. We find that generally the most-used hashtags and bigrams are used by a large number of users. The exception is the bigrams related with terrorism, where only few users use these prominently used hashtags. This can be an expression of certain users aggressively pushing these agendas into the conversation.

Sentiments over Time To analyze how positive and negative talking points regarding immigration have developed over time, we plot four of the most-common words over the entire studied time period. We plot two words with negative connotations towards immigration, *'ilegales'* (Eng: Illegals) and *'delincuentes'* (Eng: Criminals), and two with anti-xenophobia connotations, *'xenofobia'* (Eng: Xenophobia) and *'racismo'* (Eng: Racism). Figure 4 presents the results.

Figure 4 shows that prior to 2022, the term *'ilegales'* (Eng: Illegals) was generally the most common. However, after 2022 the term *'delincuentes'* (Eng: Criminals) begins to become more prominent. In February 2021, where the mass deportations took place, the most prominent term was *'ilegales'*. During the first anti-immigrant protest in September 2021, *'xenofobia'* became the most-used term among the considered ones in Figure 4, while *'ilegales'* was the second most-used. Hence, we find further descriptive evidence on two separate agendas in the Chilean Twittesphere during the protest: (i) Some users highlight the xenophobic nature of the protests, (ii) some users emphasize the undocumented and illegal situation of immigrants. The third peak in February 2022 shows a hardening of the speech as the term *'delincuentes'* begins to overtake *'ilegales'* in popularity. Calling immigrants *'delincuentes'* instead of *'ilegales'* associates immigration directly with crime instead of illegality and hence indicates increasing anti-immigration sentiments among Twitter users. In line with our previous hypothesis, our results indicate that anti-immigration sentiment is on the rise in Chile. To further analyze the observed increasing anti-immigration sentiments from Chilean Twitter users as well as some users' anti-xenophobia agendas the next section discriminates between ideologically left- and right-leaning Twitter users.

Figure 4: Usage of Anti-Immigration and Anti-Xenophobia Terms; Total Corpus



Notes: The anti-immigration terms plotted are 'ilegales' (Eng: Illegals) and 'delincuentes' (Eng: Criminals). The anti-xenophobia terms are 'xenofobia' (Eng: Xenophobia) and 'racismo' (Eng: Racism). See Appendix Figure A.5 for the same figure in logs. Textual preprocessing steps in outputs include lowercase enforcement and converting Spanish special characters (e.g. á, ñ, etc.) into corresponding non-accentuated ones.

Data Source: Retrieved from Twitter API, spanning the time frame Nov 1st, 2020 – April 11th, 2022. Corpus includes tweets by Twitter users in Chile or Chilean nationals that mainly pertain to immigration. Corpus cleaned by the steps in our methodology as described in Section 3.2 and Appendix A.2.

4.2 Political Differences

Labeled Subsample Characteristics Table 3 presents simple textual statistics for the subsample of the main corpus labeled by political affiliation. Appendix Figures A.9 and A.10 show four random tweets from left-leaning and right-leaning Twitter users, respectively. Given that Twitter data generally is biased (as discussed in Section 2), and our labeling strategy is rather simple (see Steps 6-7 in Section 3.2), the subsample of left- and right-leaning Twitter users might be more biased than the main corpus, over-representing more extreme positions. However, the results presented in this section are still consistent with previous studies, as is further discussed in Section 5.1.

Table 3: Corpus Characteristics, Political Affiliation-Labeled Subcorpus

	Count		
	Left-Leaning	Right-Leaning	Unlabeled
Number of Tweets	59,153	157,092	357,754
Unique Authors	4,530	5,076	35,919
Unique Words	83,509	140,283	261,113
Unique Hashtags	4,266	7,821	17,332

Notes: Unique words and hashtags can have duplicates across the subcategories "left-leaning", "right-leaning" or "unlabeled".

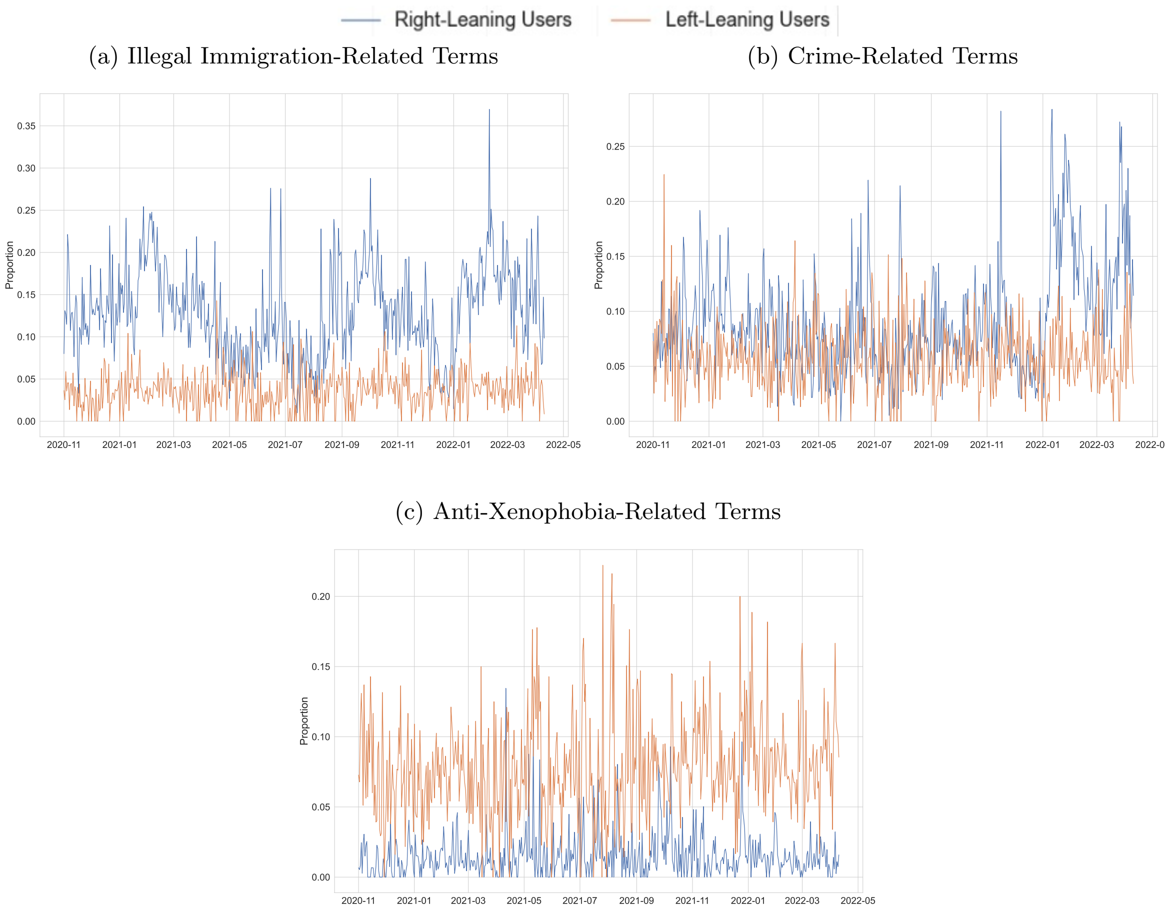
Data Source: Retrieved from Twitter API, spanning the time frame Nov 1st, 2020 – April 11th, 2022. Corpus includes tweets by Twitter users in Chile or Chilean nationals that mainly pertain to immigration. Corpus cleaned by the steps in our methodology as described in Section 3.2 and Appendix A.2.

From Table 3, it is seen that the number of left-leaning and right-leaning labeled Twitter users is approximately balanced. However, right-leaning Twitter users tweet almost thrice as much as left-leaning users in the given time frame. Right-leaning users are responsible for 27.4% of total tweets compared to only 10.3% from left-leaning users. We also find that right-leaning users post on average 31 tweets while left-leaning users post 13 tweets on average. So, it holds that right-leaning Chilean Twitter users are significantly more active regarding immigration than left-leaning users despite the almost equal number of left-leaning and right-leaning Twitter users.[14] We find further support for this using networks metrics in Section 4.3, which also shows that right-leaning Twitter users are more influential and interconnected regarding the topic of immigration than left-leaning users.

Sentiments over Time by Ideology To extend on the findings from Section 4.1, that Chilean Twitter users generally have become more anti-immigration over time, while some highlight anti-xenophobia agendas, we extend Figure 4 by distinguishing Twitter users by political affiliation. Figure 5 presents the percentage of tweets per day that contain terms related to either illegal immigration, crime and xenophobia by political affiliation.

[14] It is difficult to interpret results from the 'Unlabeled' category as this possibly includes center-leaning or ideologically neutral users (such as media outlets), or uncategorized right- and left-leaning users that our classification strategy does not classify. Some of our results, however, seem to indicate that this category mainly consists of right-leaning Twitter users. These findings are presented in Section 4.4. A relevant future improvement for our research is to utilize more accurate methods to label Twitter users by ideology as is discussed further in Section 5.2.

Figure 5: Proportion of Tweets Containing Specific Subtopic-Related Terms by Political Affiliation during the Protest; Sep 21, 2021 – Oct 1, 2021



(a) Illegal Immigration-Related Terms

(b) Crime-Related Terms

(c) Anti-Xenophobia-Related Terms

Notes: Orange line is for left-leaning Twitter users and blue is for right-leaning ones. Tweets in each of the subfigures are filtered by whether they contain at least one term in a specified list of terms. List of terms in Figure 5a: 'ilegal', 'ilegales', 'indocumentado', 'indocumentados'. List of terms in Figure 5b: 'delincuentes', 'delincuencia' , 'crimen', 'criminiales', 'delito', 'delitos', 'robo', 'ladron', 'ladrones'. List of terms in Figure 5c: 'xenofobia', 'racismo', 'discriminacion', 'discriminados'. Textual preprocessing steps in outputs include lowercase enforcement and converting Spanish special characters (e.g. á, ñ, etc.) into corresponding non-accentuated ones.

Data Source: Subsample of main corpus retrieved from Twitter API, spanning the time frame Nov 1st, 2020 – April 11th, 2022. Affiliation labels constructed using Step 6 in our methodology as described in Section 3.2 and Appendix A.2. Corpus includes tweets by Twitter users in Chile or Chilean nationals that mainly pertain to immigration.

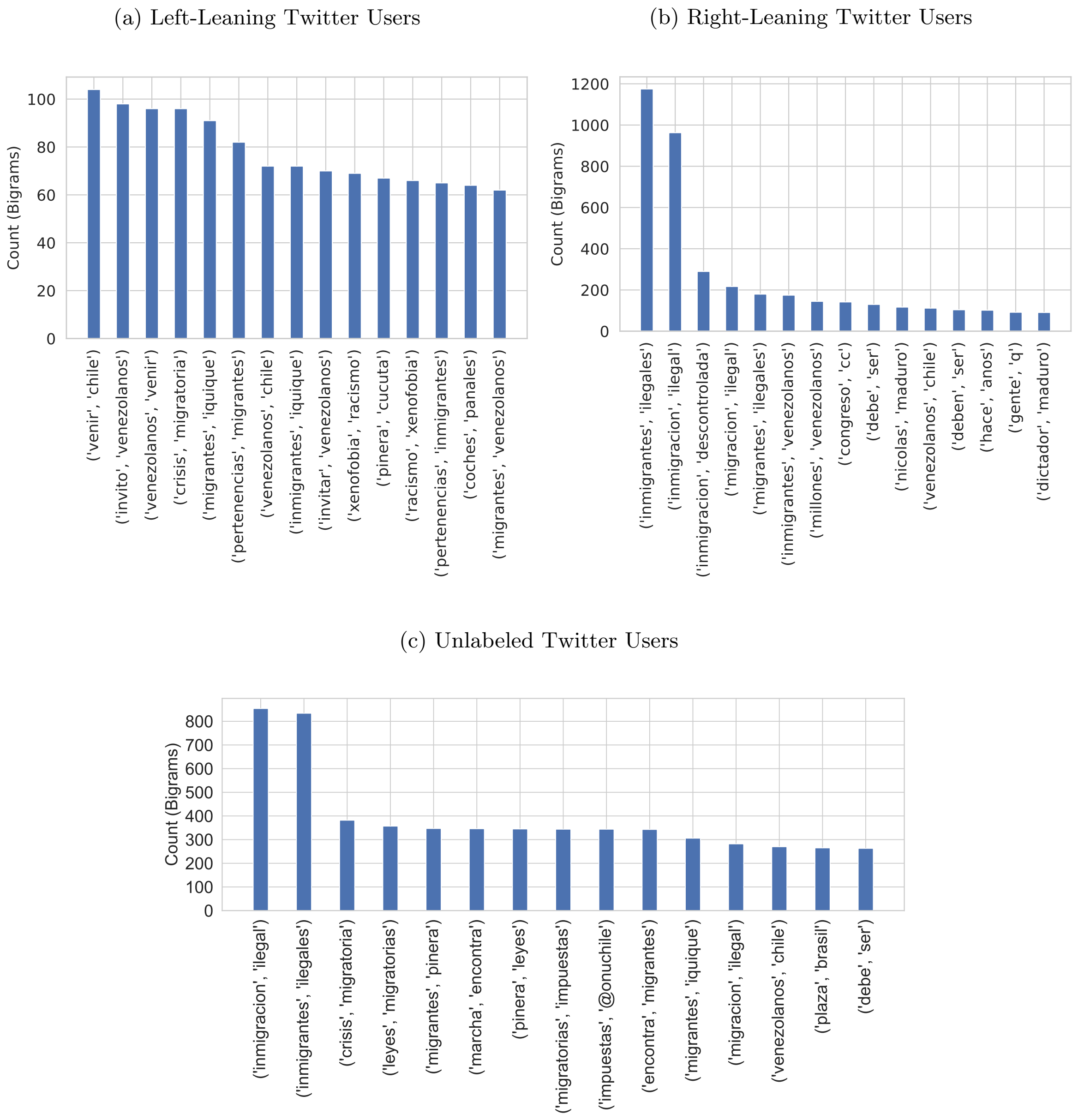
From Figure 5a it is seen that terms regarding illegal immigration are mostly used by right-leaning users while terms related to anti-xenophobic agendas are mostly used by left-leaning users (Figure 5c). With few exceptions these trends are approximately constant across the considered time frame. This finding is consistent with what could be hypothesized for each political affiliation. Surprisingly, the use of crime-related terms gives a more ambiguous picture as shown in Figure 5b. Until the beginning of 2022, crime-related terms are used in approximately equal proportions between right-leaning and left-leaning users. However, since the beginning of 2022 (following the presidential election), right-leaning users begin to increase their usage of crime-related terms. This might explain the findings in Figure 4 where the usage of the term *'delincuentes'* begins to overtake *'ilegales'* in the latter period. The increase in usage of crime-related terms by right-leaning users coincides with the third peak in Twitter activity after the anti-immigration protests in February 2022. Contrary to general Twitter activity following the peak (see Figure 1a), the proportion of crime-related terms in Figure 5b does not diminish after the protest. This might indicate that right-leaning users begin to push agendas equating immigration with crime few weeks after the new leftist government was elected in December 2021.

Analysis of a Specific Event Governments might be interested in analyzing what is happening in the Twittersphere during unusually high peaks of activity. The highest peak of activity in our corpus occurred in September 2021 coinciding with violent anti-immigration protest in the northern border city of Iquique, as mentioned in Section 4.1. For the sake of this analysis, we consider 5 days before and 5 days after the protest, i.e. September 21$^{\text{st}}$, 2022 to October 1$^{\text{st}}$, 2022.

Figure 6 presents the most-used bigrams during the protest period for left- and right-leaning Twitter users, respectively (Figure 6a and 6b, respectively). The figure shows that right-leaning Twitter users put their main emphasis on illegal immigration during the protest. Left-leaning users use a more uniformly distributed collection of bigrams, but a large part of them seem to indicate more embracing and guestfriendly views towards migrants. This is exemplified by bigrams such as *'invito, venezolanos'* (Eng: Invited, Venezuelans) and *'venezolanos, venir'* (Eng: Venezuelans, come). We also find anti-xenophobic sentiments such as *'xenofobia, racismo'* (Eng: Xenophobia, racism) and sympathies towards the immigrants' destroyed belongings by the protesters, e.g. *'pertenencias, migrants'* (Eng: Belongings, immigrants) and *'coches, pañales* (Eng: cars, diapers).[15]

[15] During the most violent episodes of the protests, the protesters burned many of the immigrants belongings such as tents and diapers.

Figure 6: Top 15 Bigrams for Twitter Users during the Protest; Sep 21, 2021 – Oct 1, 2021


(a) Left-Leaning Twitter Users

(b) Right-Leaning Twitter Users

(c) Unlabeled Twitter Users

Notes: Textual preprocessing steps in outputs include lowercase enforcement and converting Spanish special characters (e.g. á, ñ, etc.) into corresponding non-accentuated ones.

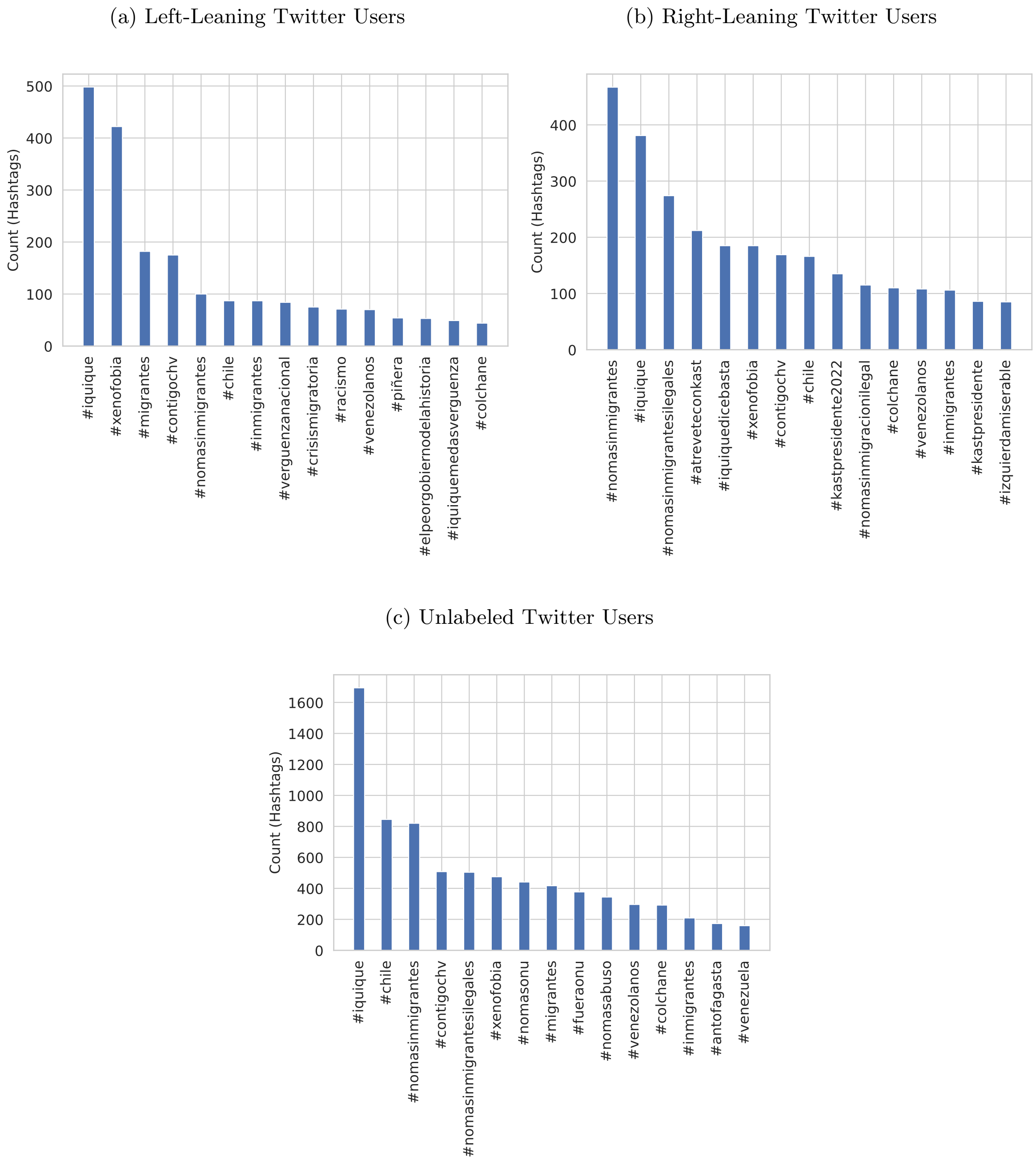
Data Source: Subsample of main corpus retrieved from Twitter API, 5 days before and after the protest, i.e. Sep 21, 2021 to Oct 1, 2021. Affiliation labels constructed using Step 6 in our methodology as described in Section 3.2 and Appendix A.2. Corpus includes tweets by Twitter users in Chile or Chilean nationals that mainly pertain to immigration.

Comparing these findings with the most-used hashtags by left- and right-leaning users in Figure 7 indicates further support for these findings. For left-leaning users' most popular hashtags in Figure 7a, we again find anti-xenophobia agendas exemplified by hashtags such as `#xenofobia` (Eng: Xenophobia), `#racismo` (Eng: Racism) and `#iquiquemedasverguenza` (Eng: Iquique you embarass me). For the right-leaning users we find further support that these hold strong anti-immigration views in Figure 7b, as seen from the use of hashtags such as `#nomasimmigrantes` (Eng: No more immigrants) and `#iquiquedicebasta` (Eng: Iquique says stop). We also find the previous focus on undocumented immigration from hashstags such as `#nomasimmigrantesilegals` (Eng: No more illegal immigrants). From both sides we find opposition to the opposite side of the ideological spectrum: Left-leaning users utilize `#elpeorgobiernodelahistoria` (Eng: Worst government in history[16]), right-leaning users tweet using `#izquierdamiserable` (Eng: Miserable left). An interesting difference between left- and right-leaning users however is that the right-leaning presidential candidate José Kast is endorsed during the protests by his supporters (with hashtags such as `#atraveteconkast` (Eng: Go with Kast) and `#kastpresidente` (Eng: President Kast)) which is not the case for left-leaning users.[17]

[16] Referring to the previous right-wing government.

[17] To distinguish if the popular hashtags and bigrams are prominent because the number of users that include it in their tweets increases or as a result of few users pushing these, Appendix Tables A.4 and A.5 present the count of users using the 15 top hashtags for left- and right-leaning users, respectively. Appendix Tables A.7 and A.8 present the top 15 bigrams by political affiliation. We observe that in general for left- and right-leaning users, the top hashtags and bigrams represent expressions used by a high number of users. The only exception for left-leaning users is the bigram *'coches, pañales'* that makes reference to burned belongings of immigrants. For right-leaning users the exception is *'congreso, cc'* which is related with political institutions (Congress and Constitutional Convention). In the case of unlabeled users appear more terms used only for a few users, mainly related with UN, migration laws and the previous right-wing President.

Figure 7: Top 15 Hashtags for Twitter Users during the Protest; Sep 21, 2021 – Oct 1, 2021



(a) Left-Leaning Twitter Users

(b) Right-Leaning Twitter Users

(c) Unlabeled Twitter Users

Notes: Textual preprocessing steps in outputs include lowercase enforcement and converting Spanish special characters (e.g. á, ñ, etc.) into corresponding non-accentuated ones.

Data Source: Subsample of main corpus retrieved from Twitter API, 5 days before and after the protest, i.e. Sep 21, 2021 to Oct 1, 2021. Affiliation labels constructed using Step 6 in our methodology as described in Section 3.2 and Appendix A.2. Corpus includes tweets by Twitter users in Chile or Chilean nationals that mainly pertain to immigration.

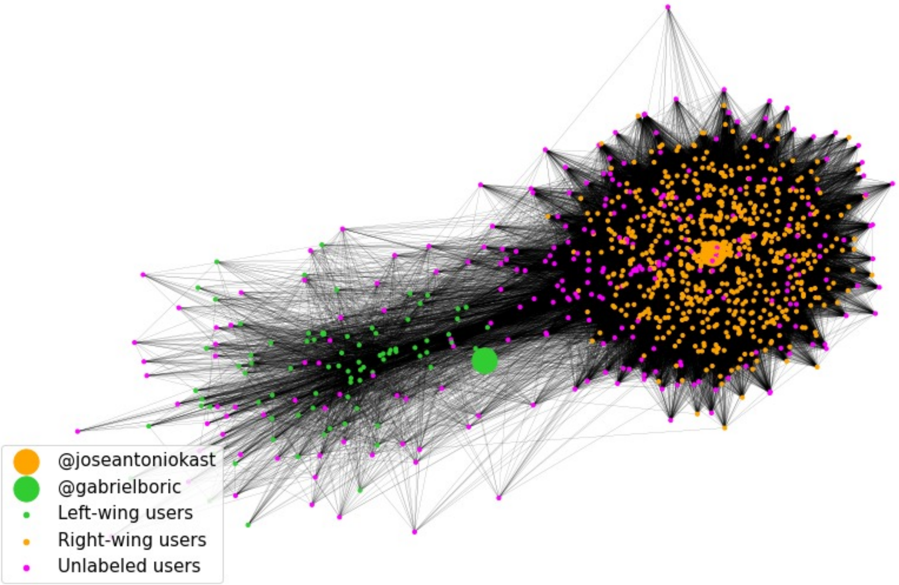
4.3 Network Analysis

Section 4.2 presented support for the hypothesis that right-leaning Twitter users hold strong anti-immigration views and are more active on Twitter. In order to analyze whether this higher activity level also translates to more influence in online conversations, we use network analysis methods to provide descriptive evidence on the most active users, the most influential users and the interconnectedness within left- and right-leaning Twitter users. (We provide a brief mathematical walkthrough of the measures utilized in this section in Appendix A.4. Practitioners can skip this appendix.)

General Network Description Our general retweet network is described by the graph $G = (E, V)$, where V is the set of vertices consisting of the 45,525 Chilean Twitter users, and E is the set of edges consisting of the 578,383 retweets among users. The edges are directed from the person retweeting to the person being retweeted. The links are weighted by total number of times a person A retweeted a person B. For our particular case, the weights are ranging from 1 (representing a unique retweet) to the maximum of 532. We use retweets as links following the literature and the previous practices of the Chilean Communication Office.

Figure 8 presents a simplified visualization of the retweet network between users. We find two clusters: one of right-leaning users and one of left-leaning users. Right-leaning users seem more prominent and more densely connected, while left-leaning users are more spread out. Considering the recent runner-up in the presidential election, the right-leaning politician José Kast, we find that he is more central in his right-wing cluster while the current left-wing president Gabriel Boric is less central in his left-wing cluster. Unlabeled users are more densely connected to the right-leaning users, however some of them are connected to left-leaning users to a higher degree. We investigate this further in Section 4.4.

Figure 8: Simplified Undirected General Graph of Retweet Network



Notes: The network plot is undirected and considers the 1,000 most important users in terms of degree centrality. Data Source: Retweets network retrieved from Twitter API and `twarc2`*'s network plug-in. Result of Step 8 of the applied methodology as described in Section 3.2 and Appendix A.2.*

In order to measure the retweet activity of Twitter users across political affiliation, we consider differences of degree measures. Since the graph is directed, we can differentiate between the volume of retweets received (terminologically called "in-degree") and sent (terminologically called "out-degree"). From Table 4 we find that among the 1,000 Twitter users with the highest degree measures, right-leaning and unlabeled users get significantly more tweets retweeted. This is in line with the findings of higher activity from right-leaning users from Table 3 in Section 4.2. For those users that retweet others, the contrast is even starker. Here right-leaning users retweet significantly more than both left-leaning as well as unlabeled users. Given this information, we find that right-leaning users are the most active about immigration, but might not have the greater influence given the important presence of unlabeled nodes.

Table 4: Degree Measures for Top 1,000 Twitter Users

	Count		
	Left	Right	Unlabeled
In-Degree	115	440	445
Out-Degree	120	829	51

Data Source: Retweets network retrieved from Twitter API and `twarc2`*'s network plug-in. Result of Step 8 of the applied methodology as described in Section 3.2 and Appendix A.2.*

Influence We have found that right-leaning users are the most active ones. In order to measure whether this higher activity translates into influence, we identify the most influential users using two different centrality measures: Degree centrality and eigenvector centrality. Degree centrality measures the influence of users based on their retweets, while eigenvector centrality measures the influence, based on the influence of their nearest neighbors. Table 5 presents the 1,000 most influential Twitter users. Using both measures, the general pattern is the same. We find that more than 60% of the most influential users are right-leaning. The influence of left-leaning and unlabeled users slightly depends on the centrality measure, but generally unlabeled users are more influential than left-leaning users. Hence, the general pattern throughout our findings also holds in terms of influence: Right-leaning Twitter users are more influential than left-leaning ones on the topic of immigration.

Table 5: Centrality Measures for Top 1,000 Twitter Users

	Count		
	Left	Right	Unlabeled
Degree Centrality	83	655	262
Eigenvector Centrality	3	605	392

Data Source: Retweets network retrieved from Twitter API and `twarc2`*'s network plug-in. Result of Step 8 of the applied methodology as described in Section 3.2 and Appendix A.2.*

To identify the most influential users, Table 6 presents the five Twitter users from our corpus with the highest degree centrality measure (by eigenvector centrality in Appendix Table A.10). The account of the right-leaning politician José Kast (runner-up in the recent presidential election) is the most influential Twitter user. This could be explained from the findings from Figure 7b in Section 4.2 that Kast's supporters used endorsing hashtags during the September 2021-protests such as `#atraveteconkast` (Eng: Go with Kast) and `#kastpresidente` (Eng: President Kast).

The unlabeled nodes found in the most influential users are mainly celebrities (`@AldoDuqueSantos`) and media outlets: television program `@T13` and radios `@Biobio` and `@Cooperativa`. The current president and left-wing leader is only found in 146th place as measured by degree centrality (7th among left-leaning users, see Appendix Table A.11) and 799th by eigenvector centrality (3rd among left-leaning users, see Appendix Table A.12). Hence the President is influential within left-leaning users but (given their general low influence level) not influential in the general Chilean Twittersphere.

Generally, we find right-leaning users to be more active than the left-leaning ones in the topic of immigration. However, in terms of influence, many unlabeled users have high scores, not solely right-leaning users.

Interconnectedness In order to analyze the interconnectedness of Chilean Twitter users across political affiliations, we consider the two metrics of density and reciprocity. The measures are conceptually similar. Reciprocity measures the probability that two authors in the network retweet each other. Density measures the proportion of retweets among all possible pairs of users in the network. To measure political affiliation-specific interconnectedness, we built two separate subnetworks: One only containing right-leaning Twitter users, the other only containing left-leaning users. Table 7 shows

Table 6: Five Most Influential Users by Degree Centrality

User	Degree Centrality	Label
`@joseantoniokast`	0.131623	Right
`@AldoDuqueSantos`	0.112336	Unlabeled
`@T13`	0.097004	Unlabeled
`@biobio`	0.088722	Unlabeled
`@Cooperativa`	0.073631	Unlabeled

Data Source: Retweets network retrieved from Twitter API and `twarc2`*'s network plug-in. Result of Step 8 of the applied methodology as described in Section 3.2 and Appendix A.2.*

that right-leaning Twitter users retweet each other more than the left-leaning Twitter users do, and hence are more interconnected. Right-leaning users are also more interconnected than the aggregate network.

Table 7: Interconnectedness Measures

Structure Measures			
	Left	Right	General Network
Density	0.002	0.008	0.0003
Reciprocity	0.014	0.044	0.023

Data Source: Retweets network retrieved from Twitter API and `twarc2`*'s network plug-in. Result of Step 8 of the applied methodology as described in Section 3.2 and Appendix A.2.*

Generally, we find that right-leaning users are more active and well-connected than left-leaning users. Our results also indicate that unlabeled users to a large extent are right-leaning, but not to a all-encompassing extent.

4.4 Unlabeled User Accounts

To further analyze the characteristics of the unlabeled users we can compare the findings from the networks metrics with those from the textual results.

Figure 6c gives insights as to which ideology is most prevalent in the 'Unlabeled' category. The distribution of bigrams is more akin to that of the right-leaning users and so are the connotations of the bigrams. We find unlabeled users to mainly stress the undocumented/illegal situation of the migrants with popular bigrams such as *'inmigrantes, ilegales'* (Eng: Immigrants, illegals), *'crisis, migratoria'* (Eng: Crisis, migratory), *'marcha, encontra'* (Eng: March, against) and *'encontra, migrantes'* (Eng: Against, migrants). So, in the unlabeled category we find that these are more similar to right-leaning users. It is possible that we have some right-leaning users in the 'Unlabeled' category that our labeling strategy classifies wrongly. It could also be the case that unlabeled users are in fact center-leaning and that center voters are more anti-immigration than embracing. This supports the finding from Figure

8 that unlabeled users are mostly centered close to right-leaning users.

From Figure 7c, we also find frequent usage of hashtags such as `#nomasinmigrantes` (Eng: No more immgirants) and `#nomasinmigrantesilegales` (Eng: No more illegal immigrants) which mainly mirror the talking points of the right. However, we find one specific talking point from the left-leaning users which is `#xenofobia` (Eng: Xenophobia). With these findings in mind, we cannot claim that unlabeled users primarily are right-leaning as the use of hashtags is ambiguous across ideological agendas. This is again consistent with findings from Figure 8 that unlabeled users seem most similar to right-leaning users, but that there is a minority more similar to left-leaning users.

While analyzing the unlabeled users, we also discover a new pattern. In Figure 6c the mention of `@onuchile` (UN in Chile's account) is prevalent, while Figure 7c shows that hashtags such as `#fueraonu` (Eng: Out with the UN) and `#nomasonu` (Eng: No more UN) are popular among unlabeled Twitter users. These bigrams and hashtags are neither used by left-leaning nor right-leaning users. From Appendix Tables A.6 and A.9, we find that it is only a few number of users that tweet these hashtags and bigrams while making heavy use of them.

Generally, it seems that unlabeled users are primarily reminiscent of the right-leaning users. More accurately categorizing the unlabeled users is therefore one of the most immediate future improvements to our project, as is further discussed in Section 5.2. Reviewing how many users include these hashtags or bigrams, we can see that there are only a few of them that tweet a lot. Again, we find that something that appears to be a general pattern in truth is a little group of people trying to push some topic in the discussion.

5 Discussion

5.1 Validation of Results

Comparison with Previous Studies In Section 2, we reviewed the most relevant, previous studies about public perception of immigration in Chile. Here we compare these findings with our own from Section 4. Studies that analyze the general perception of immigration in Chile are generally scarce. This is bad for the purpose of comparing our findings but also shows why our tool is relevant – it provides a description of the conversation around immigration that is not easily accessible today.

Centro de Estudios Públicos (2022) included some questions about immigration in their latest poll covering April to May 2022. (The CEP is one of the most prestigious opinion research centers in Chile.) In the study, 13% of respondents mentioned immigration as the main pressing issue for the Government. This contrasts with 6% in August 2021 and 1% in December 2019. This is consistent with our findings from Section 4.1 that Chileans are increasingly concerned about immigration. In a different question, respondents were asked to rank the strictness of their preferred immigration policies on a scale from 1 (most restrictive) to 10 (most lax). 61% of respondents answered between 1 and 4 (i.e. immigration-skeptic), 30% answered 5 or 6 (i.e. center-leaning) and 8% answered 7 to 10 (i.e. immigration-friendly). This is again consistent with our findings, as the most common hashtags and bigrams show opposition towards immigration. This particular question has only been asked in the most recent poll and can hence not be compared over time.

Gálvez et al. (2020) analyzes Twitter data, focusing mainly on classifying discriminatory messages and their authors from January 2018 to August 2020. The study was sponsored by the Jesuit Service for Migrants (henceforth SJM for *Servicio Jesuita a Migrantes*), which is one of the most relevant NGOs working with immigrants in Chile. The study finds that users who use discriminatory language are from the *"political extreme right, nationalist and conservative"* and that they *"declare themselves anti-leftists"*. This is consistent with our findings that right-leaning Twitter users in general hold anti-immigrant positions as seen from hashtags such as `#noesimmigracionesinvasion` which could be considered extreme right. Anti-leftist agendas are seen from the trending hashtag `#izquierdamiserable` by right-leaning users. The study also presents a word cloud of the account description of users that use discriminatory language (attached in Appendix Figure A.6 for the reader's convenience). Comparing with our corresponding Appendix Figure A.4, we find some repeated words like *'Rechazo'* (Eng: Reject) or *'Derecha'* (Eng: Right). Differences do exist due to different sample selection criteria and time frames, but the existence of common patterns is consistent considering that both are analyzing the same topic.

González et al. (2019) is an academic study about general perceptions about immigration in Chile from 2002 to 2017. The study presents polling data showing that 57% of Chileans think that irregular migration is a problem and that the Government should exclude illegal immigrants. The study also finds that right-leaning individuals have slightly higher anti-immigration attitudes compared to individuals with left-leaning or center-aligned political positions. These findings are consistent with what we found for a different period of time.

It generally appears that our findings are consistent with previous studies, giving us confidence in the performance of our developed methodology. We can filter Twitter conversations by users' nationality and topics as well as labeling them by political affiliation, and reproduce previous findings as well as find novel insights. Hence, our social listening tool seems to be accurate and useful in the test subject of immigration in Chile.

Generalizability of Methodology To validate that our corpus construction methodology from Section 3 functions well across other relevant political topics we have run the methodology for the topic of feminism in Chile from March 9 to March 13, 2022.[18] The results are presented in Appendix C. The highest peak of Twitter activity during this period is on March 9, one day after the manifestation of International Women's Day, which seems reasonable. We also find the top hashtag to be `#8m`, which refers to March 8. Boric's presidential inauguration ceremony was held on March 11, where we find the second-highest peak of activity. As a result of this, we find bigrams such as *"feminist government"*. Unlike in the topic of immigration, for feminism we find more left-wing users in the conversation.

5.2 Future Improvements

Advanced Political Affiliation Classification As shown in Section 4, the "unlabeled" users are a quite heterogeneous group. They seem to include both right- and left-leaning users, neutral users (such as media outlets) and presumably center-leaning users. These users did not use enough political hashtags during the election to survive our threshold for classification.

To get a more accurate description of affiliation-specific talking points in the Twittersphere, a future improvement could be to use more advanced classification algorithms to more precisely label users. Possible approaches are machine learning models (considering e.g. users' linguistic patterns, see e.g.

[18] We only consider a small timeframe in order to have quick results.

Conover et al. (2011b)) or label propagation (from network analysis, see e.g. Raghavan et al. (2007)). As presented in Section 2, a lot of research concerning the classification of Twitter users' political affiliation has already been done. Hence, the only obstacle for incorporating this improvement to our methodology is time.

Bot Detection While analyzing Table 4, we became suspicious that some users might be fake accounts. E.g. we found an active retweeter named `@j-pablo-escobar`. This account has no picture, only a fake and famous alias. The bio states *"Patriota, Republicano, voto por kast"* (Eng: "Patriot, Republican, vote for Kast"). Upon manual inspection of this user's activity, we find that the retweets consist of controversial facts or opinion polls which enables his followers to spread these views. Hence, it seems there are some fake accounts in our corpus.

There is a growing academic literature regarding bot detection. Efthimion et al. (2018) provide a supervised machine leanrning model that detects bots with a 2.25% of misclassification rate, using as features the length of the username, sentiment expressions, variability of the activity, among others. This approach's main caveat is that it requires information on the full activity of users, which would be computationally exhausting for long lists of users. Knauth (2019) tries to face this problem and provides different models. Some of these require the full activity, but others only need user metadata like username, profile picture, etc. Unfortunately, the Twitter API does not return e.g. profile pictures, and these models also need a manually labeled training dataset.

A future improvement to our work is to add bot detection algorithms, to allow for distinguishing human conversations by bot-generated ones. Political institutions might want to filter these out to solely analyze their citizens' opinions.

Test Tool Across Countries We have run our methodology for two topics, immigration and feminism (in Appendix C), and validated that the methodology functions well across political topics. It would still have to be tested how well it performs across countries. Testing this might highlight potential issues with the proposed approach in Step 3 of the methodology.

Improved Dashboard If we had substantially more time, the final dashboard could be updated to visualize metrics such as

- Filter by dates, list of authors, verified accounts, accounts with minimum number of follower, tweets that include some word or hashtag and tweets with a minimum number of retweets, likes, quotes or replies.
- Display number of tweets per day, common hashtags, word clouds, common bigrams, use of selected word over time, most popular tweets (measured by retweets, likes, citations) and metrics per tweet. All of these outputs for the entire data set and also for only left-leaning and right-leaning users.
- Interactive network results, allowing to highlight any node in the plot.
- Plotting sentiment scores in a given time frame.
- A button to exclude bots from all outputs.

This allows the user to extract information such as all the outputs shown in Section 4 (and more) that our current app does not allow.

Extended Feature Engineering (Gender and Age) Political Affiliation is a relevant covariate to distinguish users, but other ones can also be informative. Trying to classify users by gender (e.g. using common male and female names for instance) or by age (considering users' linguistic patterns as in Rao et al. (2010)) could provide more detailed information. This would allow practitioners to better understand the different speeches by different groups and better target their communication strategy.

Sentiment Scores Researchers such as Pérez et al. (2021) and Gonzalez et al. (2021) focus their work in training Bidirectional Encoder Representations from Transformers models (also known as `BERT`) in Devlin et al. (2018) on Spanish Twitter corpora. They provide libraries that allow researchers to obtain sentiment score and emotion analysis from Spanish tweets. We tried implementing such models, but they are computationally demanding, leading to a trade-off between how quickly results are needed and the value of recently developed methods. Despite this, we think that a future improvement could be to add these indicators to analyze how they change over time and across political affiliations.

Topic Modeling Algorithms Utilizing topic model algorithms such as the Latent Dirichlet Allocation (Blei et al., 2003) could give further insights into Chilean Twitter users' conversational subtopics within a given main topic.[19] We tried implementing LDA in our project but it proved infeasible as it is computationally demanding and not built for short text corpora, such as Twitter data. It might still prove useful in a future iteration of the project. In this project, as of writing, we have approximated the function of LDA by analyzing bigrams, but specific topic modeling algorithms might give a more complex and accurate description of subtopics.

6 Conclusion

This project has shown that our methodology works as intended to construct a corpus with political topic-specific Twitter conversations. It has been shown to work well across multiple topics and provide relevant descriptive insights regarding users' concerns and agendas and the structure of their interactions. It is important for practitioners to be aware of the project's limitations, primarily that Twitter users do not represent the whole population and that extreme opinions hence might be overrepresented.

We believe this project to be immensely useful for governments as it can provide opinion data at high frequency and at low cost and thereby complement traditional phone- and survey-based opinion polls. Insights from the data can help political institutions realize important topics and agendas within topics across political affiliations. Metrics from network analysis can help identify the most influential Twitter users. Insights can be easily obtained by visualizing the data as an interactive dashboard. Our entire methodology is available in our GitHub repository and is structured with ease-of-usage as a main priority.

This thesis presented the first version of our social listening tool. As future improvements we propose to add more advanced classification algorithms for political affiliation and bot detection.

[19] In our case, such subtopics could be crime, xenophobia, etc.

To provide more precise descriptions of users' concerns and agendas, we further propose to analyze sentiment scores and topic modeling and to distinguish subgroups by constructing features such as users' gender and age.

While applying the social listening tool we obtained novel findings about Chilean Twitter conversations regarding immigration. By analyzing a corpus of Chilean Twitter conversation about immigration from November 2020 to April 2022, we found the Chilean Twittersphere to become increasingly concerned about immigration. This is especially pronounced for right-leaning users. These post up to thrice as often as left-leaning users and retweet more often. We find that right-leaning users are primarily concerned about the undocumented situation of immigrants and crime. They also exploit specific events of higher general Twitter activity, such as a violent protest in September 2021, to campaign for the politicians they support. Right-leaning users also retweet each other more often and are more influential on Twitter – especially their presidential candidate José Kast.

Left-leaning users are primarily concerned about their views of rising xenophobia and racism. They do not have substantial influence on Twitter about immigration and are outdominated by the right-leaning users, especially by José Kast. Our findings are consistent with previous studies.

We also make a particularly surprising finding: Some agendas are prevalent in the general Chilean Twittersphere, but are used only by a few users in high frequency. This indicates that certain users try to aggressively push these agendas into the general conversation. This pertains to two specific agendas: Linking immigration with terrorism and blaming the UN for Chile's migration crisis.

In addition to being useful for practitioners, this thesis also contributes to the academic literature in three ways. First, we provide a general-purpose methodology to accurately construct a corpus of Twitter conversations regarding a specific topic for a given country. Second, we contribute to the field of feature engineering by extrapolating users' geolocation and nationality, which can be useful across virtually all studies using Twitter data. Third, we add to the scarce literature regarding public opinion in Chile towards immigration.

We believe our thesis to have presented new avenues for further research and presented a first prototype for a social listening tool for practitioners.

References

Agarwal, A., B. Xie, I. Vovsha, O. Rambow, and R. J. Passonneau (2011): "Sentiment Analysis of Twitter Data," in *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, 30–38.

Barberá, P. and G. Rivero (2015): "Understanding the Political Representativeness of Twitter Users," *Social Science Computer Review*, 33, 712–729.

Basile, V., C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, and M. Sanguinetti (2019): "SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, Minneapolis, Minnesota, USA: Association for Computational Linguistics, 54–63.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003): "Latent dirichlet allocation," *Journal of machine Learning research*, 3, 993–1022.

Cataldi, M., L. Di Caro, and C. Schifanella (2010): "Emerging Topic Detection on Twitter Based on Temporal and Social Terms Evaluation," in *Proceedings of the tenth international workshop on multimedia data mining*, 1–10.

Centro de Estudios Públicos (2022): "Estudio Nacional de Opinión Pública – Encuesta CEP 86," Report, last accessed June 21, 2022, https://www.cepchile.cl/cep/site/docs/20220608/20220608124401/encuestacep_abril_mayo2022.pdf.

Conover, M., J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini (2011a): "Political Polarization On Twitter," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 5, 89–96.

Conover, M. D., B. Goncalves, J. Ratkiewicz, A. Flammini, and F. Menczer (2011b): "Predicting the Political Alignment of Twitter Users," in *2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing*, IEEE, 192–199.

Datos Macro (2010): "Aumenta el número de inmigrantes en Chile," *Datos Macro - Expansión, https://datosmacro.expansion.com/demografia/migracion/inmigracion/chile?anio=2010*.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2018): "Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805*.

Efthimion, P. G., S. Payne, and N. Proferes (2018): "Supervised Machine Learning Bot Detection Techniques to Identify Social Twitter Bots," *SMU Data Science Review*, 1, 5.

Gonzalez, J. A., L.-F. Hurtado, and F. Pla (2021): "TWilBert: Pre-trained Deep Bidirectional Transformers for Spanish Twitter," *Neurocomputing*, 426, 58–69.

González, R., E. Muñoz, and B. Mackenna (2019): "Cómo quieren en Chile al amigo cuando es forastero: Actitudes de los chilenos hacia la inmigración," *Isabel Aninat and Rodrigo Vergara (des), Inmigración en Chile. Una mirada multidimensional*, 321–346.

Gálvez, D., P. Durán, T. Lawrence, and N. R. Pedemonte (2020): "Barómetro de Percepción de la Migración 2018-2020," .

Himelboim, I., M. A. Smith, L. Rainie, B. Shneiderman, and C. Espina (2017): "Classifying Twitter Topic-Networks Using Social Network Analysis," *Social Media + Society*, 3, 205630511769154.

Hong, L. and B. D. Davison (2010): "Empirical Study of Topic Modeling In twitter," in *Proceedings of the first workshop on social media analytics*, 80–88.

Instituto Nacional de Estadísticas (2020): "Población extranjera residente en Chile llegó a 1.462.103 personas en 2020, un 0,8% más que en 2019," *Instituto Nacional de Estadísticas*, *https://www.ine.cl/prensa/2021/07/29/poblaci%C3%B3n-extranjera-residente-en-chile-lleg%C3%B3-a-1.462.103-personas-en-2020-un-0-8-m%C3%A1s-que-en-2019*.

Jungherr, A. (2016): "Twitter Use in Election Campaigns: A Systematic Literature Review," *Journal of information technology & politics*, 13, 72–91.

Knauth, J. (2019): "Language-Agnostic Twitter-Bot Detection," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, 550–558.

Kruspe, A., M. Häberle, E. J. Hoffmann, S. Rode-Hasinger, K. Abdulahhad, and X. X. Zhu (2021): "Changes in Twitter Geolocations: Insights And Suggestions for Future Usage," in *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, Online: Association for Computational Linguistics, 212–221.

Lin, Y.-R., B. Keegan, D. Margolin, and D. Lazer (2014): "Rising Tides or Rising Stars?: Dynamics of Shared Attention on Twitter during Media Events," *PloS one*, 9, e94093.

Maharani, W., A. A. Gozali, et al. (2014): "Degree Centrality and Eigenvector Centrality in Twitter," in *2014 8th international conference on telecommunication systems services and applications (TSSA)*, IEEE, 1–5.

Pennacchiotti, M. and A.-M. Popescu (2011): "A Machine Learning Approach to Twitter User Classification," in *Proceedings of the international AAAI conference on web and social media*, vol. 5.

Pereira-Kohatsu, J. C., L. Quijano-Sánchez, F. Liberatore, and M. Camacho-Collados (2019): "Detecting and Monitoring Hate Speech in Twitter," *Sensors*, 19, 4654.

Plaza-del Arco, F. M., M. D. Molina-González, L. A. Ureña-López, and M. T. Martín-Valdivia (2021): "Comparing Pre-trained Language Models for Spanish Hate Speech Detection," *Expert Systems with Applications*, 166, 114120.

Pérez, J. M., D. A. Furman, L. A. Alemany, and F. Luque (2021): "RoBERTuito: A Pre-trained Language Model for Social Media Text in Spanish," .

Raghavan, U. N., R. Albert, and S. Kumara (2007): "Near linear time algorithm to detect community structures in large-scale networks," *Physical review E*, 76, 036106.

Rao, D., D. Yarowsky, A. Shreevats, and M. Gupta (2010): "Classifying Latent User Attributes in Twitter," in *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, 37–44.

Saif, H., Y. He, and H. Alani (2012): "Semantic Sentiment Analysis of Twitter," in *International semantic web conference*, Springer, 508–524.

Small, T. A. (2011): "What the Hashtag? A Content Analysis of Canadian Politics On Twitter," *Information, communication & society*, 14, 872–895.

Tufekci, Z. (2014): "Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls," in *Eighth international AAAI conference on weblogs and social media*.

Zhao, W. X., J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li (2011): "Comparing Twitter and Traditional Media Using Topic Models," in *European conference on information retrieval*, Springer, 338–349.

A Appendices

A.1 Original Thesis Challenge

> *"Desarrollar una metodología para analizar la conversación chilena en twitter sobre un tópico específico, considerando afiliación política de los usuarios. Describir las narrativas asociadas que aparecen a lo largo del tiempo y la estructura de red de los grupos involucrados. Aplicar la metodología desarrollada utilizando como tema de prueba inmigración."*
>
> — *Claudio Villegas Oliva, Chilean Communication Office*

A.2 Applying the Methodology

This section presents how we have applied the general methodology presented in Section 3.2.

All the notebooks can be found in the `methodology_immigration`-folder of the GitHub repository (particularly in the subfolder `new_topic`).

Step 1 We input the term "immigration" into *Semantic Link*. Utilizing our domain knowledge, we exclude words returned by *Semantic Link* that (*i*) are unrelated with immigration in Chile such as "Aliyah" or (*ii*) highly related to other topics such as "unskilled". We translate the relevant words and then we run the notebook[20] `1.1_First_Query_to_Twarc.ipynb` with our specified list of semantically linked keywords to return the first corpus with geolocated tweets from Chile. See Appendix A.3.1 for our full query for Step 1.

Output: Corpus of geolocated tweets from Chile that contain semantically linked keywords to the term "immigration". The resulting corpus from Step 1 had 15,339

Validation: Defining the appropriate set of words is of great importance. Including a too narrow list in the tweet retrieve search, would potentially fail to capture relevant tweets. Specifying a too broad list, would potentially capture too much noise and risks exceeding the maximum download limits set by the API. In this step we try different sets of words and we explored the outputs.

Step 2 We extend the first list of semantically linked keywords to "immigration" by exploring the output retrieved from Step 1. We investigate the top 200 words and hashtags, and from this select the terms "venezolanos", "haitianos","xenofobia" and "extranjeros" to add to our list of keywords. Further, we again utilize our domain knowledge and also we created a list of Chilean related words with the last name of the president and the two candidates in last election, names of cities, and all the regional capitals, most common slangs and the account of Chilean Government to obtain tweets related with Chile. The first query require that the tweets have to contain one immigration related word and one Chilean related word. The second one require one immigration related word and a tweet geo-located in Chile. We run the notebook `1.2_Adding_words_looking_the_Chilean_context` to download the corresponding tweets. See Appendix A.3.2 for our full query for Step 2.

[20] The words that we used were: inmigración migración, migrante,migrantes, inmigrante, inmigrantes, emigrantes, deportación, deportado, deportados, refugiado and refugiados

Output: Corpus of geolocated tweets from Chile or tweets that used Chilean related words that contain semantically or contextual linked keywords to the term "immigration". The resulting corpus from Step 2 had 502,510 tweets.

Step 3 To filter the corpus from Step 2 by citizens that are either located in Chile or Chileans living abroad apply the criteria outlined in Step 3 in Section 3.2 as follows. We only include Twitter users that have one or more of the following in their self-written author descriptions and/or locations: (*i*) A Chilean flag as an emoji ("`flag:chile`"); (*ii*) *'chile'* as a regular expression, including derivations thereof, such as *'chileno'*, *'chilena'*; (*iii*) An unambiguous Chilean city either as an n-gram or unigram. We run the notebook `1.3_Filter_by_Chileans` to filter the Twitter users. See Appendix Table A.13 for the list of unigram excluded (ambiguous) list of cities in Chile and the included ones; Appendix Table A.14 presents the list of n-gram excluded (ambiguous) list of Chilean cities and the included ones.

Output: List of authors located in Chile or Chilean citizens tweeting about immigration as specified by the semantically linked keywords to the term "immigration" and immigration-Chile-contextual keywords. The output from this step was a list of 45,550 unique authors.

Validation: Reviewing the list of Twitter users obtained after Step 3, we are confident that the vast majority of tweets in the resulting corpus are written by Twitter users located in Chile or Chilean citizens by reviewing a random subsample of authors' location and description, the top 20 author locations, and wordcloud of author description. See Appendix Table A.7, Appendix Figure A.2 and Appendix Figure A.4 for the previously mentioned investigations, respectively.[21]

Step 4 This step retrieves all the tweets about immigration from Chilean authors identified in Step 3. We add the the list of semantic linked and Chile-contextual keywords from Step 2 into the `twarc2` query. We also again specify the time frame. We run the notebook `1.4_Download_all_related_tweets_from_local_authors` to obtain the full list of tweets from the specified authors.[22] See Appendix A.3.3 for our full query for Step 4.[23]

Output: Corpus with tweets that contain-immigration related words (by semantic link and Chilean Twitter context), only included Twitter users that are either located in Chile or Chilean citizens, and are active in the conversations on immigration. The resulting corpus from Step 4 had 574,219 tweets.

Validation To ensure that our methodology works as intended up to Step 4, we analyze a random selection of Twitter users in the corpus and investigated their author description and location. We have also looked at the most common author location and word clouds of the author descriptions. This is done for our specific application on the topic of immigration in Chile and is presented in Appendix Figures A.2 and A.4. This exploratory analysis gives us confidence that the methodology works as intended and that the vast majority of tweets in the resulting corpus are written by Twitter users from

[21] Appendix Figure A.2 shows that in the 20 most popular author locations in our sample are all from Chile. Appendix Figure A.4 shows the wordcloud of author biographies where domain knowledge confirms is that the words to a large degree are regarding Chilean users. The wordcloud includes Chilean words and phrases related to the topic of immigration in the country.

[22] This is a computationally intensive task. The list of authors was split into three and ran on separate computers. Each query took between 5-7 hours to complete in each of the machines.

[23] Considering that twarc has a maximum number of characters for a single query, we split the authors list and we create a list of queries that we run from a txt file using the option searches.

the country studied. However, the corpus after Step 4 can still contain noise and tweets that do not solely pertain to the topic of interest.

Step 5 The corpus returned from Step 4 still contains some irrelevant tweets. We check this by looking at the top words, hashtags, etc. E.g. we find a multitude of tweets regarding football matches between the Chilean and Venezuelan national football teams. Using the word- and hashtag-based filtering approach as described in Step 5 in Section 3.2, we clean our corpus from irrelevant tweets.[24] The total number of tweets that we remove in Step 5 is 218.

Output: Corpus from Step 4 with noise filtered out. This output is our final data set to analyze and had 573,999 tweets from 45,525 unique authors.

Validation We reviewed a random sample of 200 tweets after this step to ensure that mainly of our corpus contain tweets related to immigration in Chile. 187 of these tweets are related to our topic of interest.

Step 6 Here aim to characterize the authors by their political affiliation. To do so, we label the users according to the candidate that they supported in the last election. (this can be done also by ML models, but it can increase the uncertainty of measurements)

Our first step was to take a list of left and right politicians and download their tweets during the electoral period. From this data set, we extract the 15 most popular hashtags for each affiliation, previous a manually filtering according with hashtags that specifically support one of the candidates. Table A.1 shows the specific hashtags chosen to label Twitter users in the corpus into left- and right-leaning political affiliation, respectively.

Output: List of right leaning hashtags and left leaning hashtags.

Step 7 Using the previous list of hashtags we download all the tweets from the users that have tweets in the output of the step 5 in the electoral period that used one of the political hashtags. We labeled as left leaning all the authors that used more than 40 left hashtags during the campaign and that more than 80% of the political hashtags that they used are from the list of left leaning hashtags. We do the same for the right leaning.

Output: A list of users that tweet about immigration with their political affiliation. The dimensions of the data frame with authors labeled left or right was 9,606 (21% of all our authors). We merge this information in our previous data set, to obtain labeled tweets and we add the label "Unlabeled" to all the users that didn't have Right or Left label.

Validation: Reviewing of a random sample of 40 tweets for each affiliation during the electoral period (the tweets that we used to label users), we found that 40/40 tweets from right wing users supports one of the right wing candidates, and 39/40 tweets from left wing users support one of the left wing candidates, only 1/40 was neutral.

Step 8 With the same list of keywords that we have from step 2, we download from twarc all the retweets from our list of users that contains one of the keywords. After with the plugin networks, we

[24] We exclude the tweets that contain the hashtags: 'apostilla','25deJulio','venezolanosenelmundo','vamoscolocolo','vamoslau' and the tweets that contain the words:'gol','futbol','foul','apostilla','futbolista'

Table A.1: Hashtags to Label Twitter Users from Corpus into Left- And Right-Leaning Political Affiliation

Left		Right	
Hashtag	Relative Frequency	Hashtag	Relative Frequency
`#boricpresidente`	1,095	`#atreveteconkast`	189
`#seguimos`	606	`#kastpresidente2022`	166
`#boricpresidentedechile`	312	`#kastpresidente`	139
`#apruebodignidad`	195	`#vota2votakast`	90
`#rutaesperanzaxboric`	176	`#atreveteporchile`	76
`#boricnosune`	152	`#sepuede`	68
`#boricpresidente2022`	152	`#todochileconkast`	68
`#unmillondepuertasxboric`	142	`#kastledaesperanzaachile`	57
`#1millondepuertasxboric`	137	`#chilevotakast`	48
`#meunoconboric`	136	`#atrevidos`	47
`#ahorayasna`	117	`#consichelsepuede`	44
`#boricpresidentedechile2022`	108	`#sichelpresidente`	41
`#boricenprimeravuelta`	106	`#atrevidosporkast`	34
`#vota1`	87	`#votakast`	31
`#paravivirmejor`	81	`#mujeresporkast`	30

Notes: Textual preprocessing steps include lowercase enforcement and converting Spanish special characters (e.g. á, ñ, etc.) into corresponding non-accentuated ones.
Data Source: Retrieved from Twitter API utilizing the `twarc2` *command line tool, spanning the time frame Oct 19st 2021 – Dec 20th, 2021 (electoral period).*

obtain a `.csv`-file with the necessary information to create the network (who retweeted, who received the retweet and the amount of retweets that the second users gives to the first one). Adding the previous information about labels, we create the network considering only the retweets between two users that we have in our previous list. Each node has an attribute with their political affiliation and each edge has a weight indicating the number of retweets.

Output: Network of retweets with Chilean users that talk about immigration during the period of interest. The network has 45,525 nodes (users) and 578,383 edges (retweets between users).

A.3 Full `twarc2` Queries for Application of Methodology

A.3.1 Full `twarc2` Query for Step 1

```
twarc2 search --archive "(inmigración OR inmigracion OR migración OR migracion OR migrante
OR migrantes OR inmigrante OR inmigrantes OR emigrantes OR deportación OR deportacion OR deportado
OR deportados OR refugiado OR refugiados) place_country:CL -is:retweet" --start-time "2020-11-01"
--end-time "2022-04-11"
```

A.3.2 Full `twarc2` Queries for Step 2

```
twarc2 search --archive "(inmigracion OR  migracion OR migrante OR migrantes OR inmigrante
OR inmigrantes OR emigrantes OR deportacion OR deportado OR deportados OR refugiado OR refugiados
OR venezolanos OR extranjeros OR xenofobia OR haitianos) place_country:CL -is:retweet"
--start-time "2020-11-01" --end-time "2022-04-11"

twarc2 search --archive "(inmigracion OR  migracion OR migrante OR migrantes OR inmigrante
OR inmigrantes OR emigrantes OR deportacion OR deportado OR deportados OR refugiado OR refugiados
OR venezolanos OR extranjeros OR xenofobia OR haitianos) (chile OR santiago OR iquique OR arica
OR piñera OR pinera OR boric OR kast OR valparaiso OR antofagasta OR colchane OR copiapo OR
coquimbo OR rancagua OR talca OR conce OR temuco OR puerto montt OR valdivia OR coyhaique OR
punta arenas OR weon OR weona OR sebastianpinera OR gabrielboric OR joseantoniokast OR
GobiernodeChile) -is:retweet" --start-time "2020-11-01" --end-time "2022-04-11"
```

A.3.3 Full `twarc2` Query for Step 4

```
twarc2 search --archive "(inmigracion OR  migracion OR migrante OR migrantes OR inmigrante OR
inmigrantes OR emigrantes OR deportacion OR deportado OR deportados OR refugiado OR refugiados
OR venezolanos OR extranjeros OR xenofobia OR haitianos)(from:user1 OR from:user2 OR ...)
-is:retweet" --start-time "2020-11-01" --end-time "2022-04-11"
```

A.3.4 Full `twarc2` Query for Step 6

```
twarc2 timelines --use-search --start-time "2021-10-19" --end-time "2021-12-20" left_accounts.txt

twarc2 timelines --use-search --start-time "2021-10-19" --end-time "2021-12-20" right_accounts.txt
```

A.3.5 Full `twarc2` Query for Step 7

```
twarc2 search "(#boricpresidente OR #seguimos OR #boricpresidentedechile OR #apruebodignidad OR
#rutaesperanzaxboric OR #boricnosune OR #boricpresidente2022 OR #unmillondepuertasxboric OR
#1millondepuertasxboric OR #meunoconboric OR #ahorayasna OR #boricpresidentedechile2022 OR
#boricenprimeravuelta OR #vota1 OR #paravivirmejor OR #atreveteconkast OR #kastpresidente2022 OR
#kastpresidente OR #vota2votakast OR #atreveteporchile OR #sepuede OR #todochileconkast OR
#atreveteconkast OR #kastledaesperanzaachile OR #chilevotakast OR #atrevidos OR #consichelsepuede
OR #sichelpresidente OR #atrevidosporkast OR #votakast OR #mujeresporkast)
(from:user1 OR from:user2 OR from:user3..."--start-time "2021-10-19" --end-time "2021-12-20"
```

A.3.6 Full `twarc2` Queries for Step 8

```
twarc2 search --archive "(inmigracion OR  migracion OR migrante OR migrantes OR inmigrante
OR inmigrantes OR emigrantes OR deportacion OR deportado OR deportados OR refugiado OR refugiados
OR venezolanos OR extranjeros OR xenofobia OR haitianos)(from:user1 OR from:user2 OR ...)
is:retweet" --start-time "2020-11-01" --end-time "2022-04-11" output_file.json

twarc2 network output_file.json --format csv network_final.csv --edges retweet
```

A.4 Theory for Network Analysis

Our project mainly presents simple descriptive measures such as counts, percentages, etc. which we assume readers to already be familiar with. However, some more advanced measures are used, especially in our networks analysis part, and this section provides the mathematical definitions of these measures.

Graph Theory A network can be defined by the graph $G : G = (E, V)$. Here V is the set of vertices (observable units, such as individuals), while E is the set of edges (the measurement of connection between units, such social interactions). Considering a social network, the total number of individuals in the network n is the vertices set cardinality, $n = |V|$. Similarly, the number of social interactions m, such as retweets, is $m = |E|$. The vertices can be weighted and the edges can be weighted and/or directed. The adjacency matrix A_{ij} is a square matrix indicating the weight and direction of connections between each (i, j)-pairs of vertices in the graph. (A is symmetric if the connections are undirected, and the weigh of a_{ij} is 0 if no connection is occurring.)

Degree When the edges are not directed, the degree k_i of a vertex i represents the number of nearest neighbors the vertex has. By definition, for a node i in a undirected network

$$k_i \equiv \sum_j a_{ij} \equiv \sum_j a_{ji} \tag{1}$$

When edges are directed, we distinguish between in-degrees (the number of incoming links) and out-degrees measures of the vertex (the number of outgoing links). For a node i in a directed network:

$$k_i^{out} \equiv \sum_j a_{ij} \tag{2}$$

$$k_i^{in} \equiv \sum_j a_{ji} \tag{3}$$

In other words, the out-degree of a node i is the sum of the row i of the adjacency matrix A. On the other hand, the in-degree of a node is the sum of the column i of the adjacency matrix.

Centrality Centrality is the measure of the influence a node has. For our project, we consider two centrality measures: *Degree centrality* and *eigenvector centrality.* For both measures, the higher the centrality score is for a node, the more influential it is.

Degree centrality depends on the number of nearest neighbors. It consists of measuring for each nodes, the fraction of nodes it is connected to. Formally, for the node i:

$$C_i \equiv \frac{k_i}{(n-1)} \tag{4}$$

where C_i is centrality of node i, k_i is the degree of the node i and n the number of nodes in the network.

Eigenvector centrality depends on the importance of its neighbors. For node i, the centrality is given by the i^{th} element of the vector x in the equation:

$$Ax \equiv \lambda x \tag{5}$$

where A is the adjacency matrix that represents the network and λ is the eigenvalue of the matrix. The vector x represents the eigenvector centrality for each node.

Density The density of a network d is based on the ratio of number of edges m to number of nodes n in the graph:

$$d \equiv \frac{m}{n(n-1)} \tag{6}$$

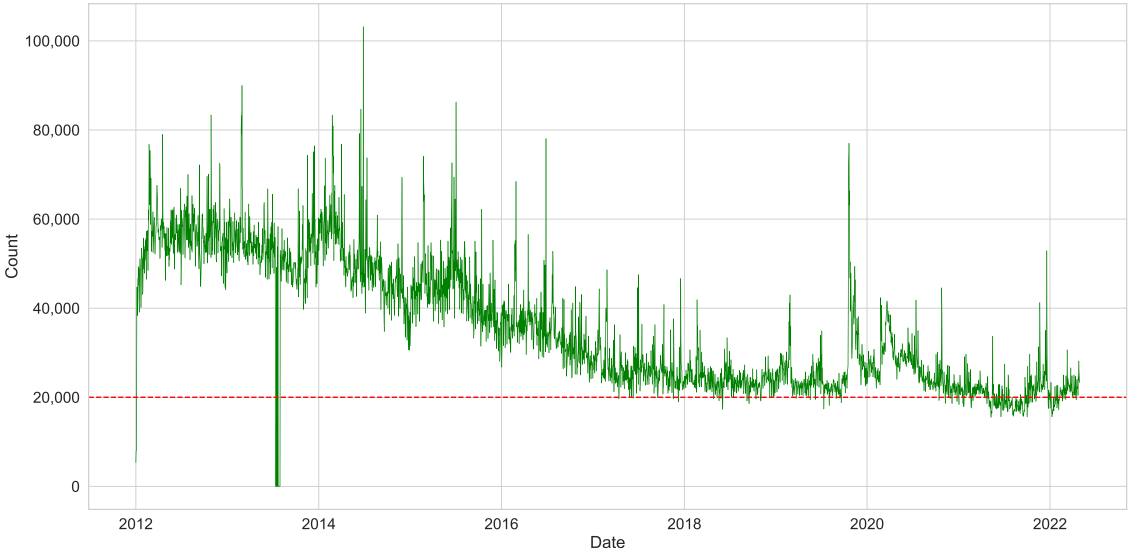
The measure is bounded $d \in [0,1]$. The closer to unity, the more dense the network is. Hence, a network with a density measure of 0 represents a network without any links.

Reciprocity In directed graphs, reciprocity r is the ratio of the number of edges pointing in both directions to the total number of edges in the graph:

$$r \equiv \frac{|(i,j) \in E \cap (j,i) \in E|}{|(i,j) \in E|} \tag{7}$$

where E is the set of all the edges in the network and (i,j) are all the possible pairs of nodes.

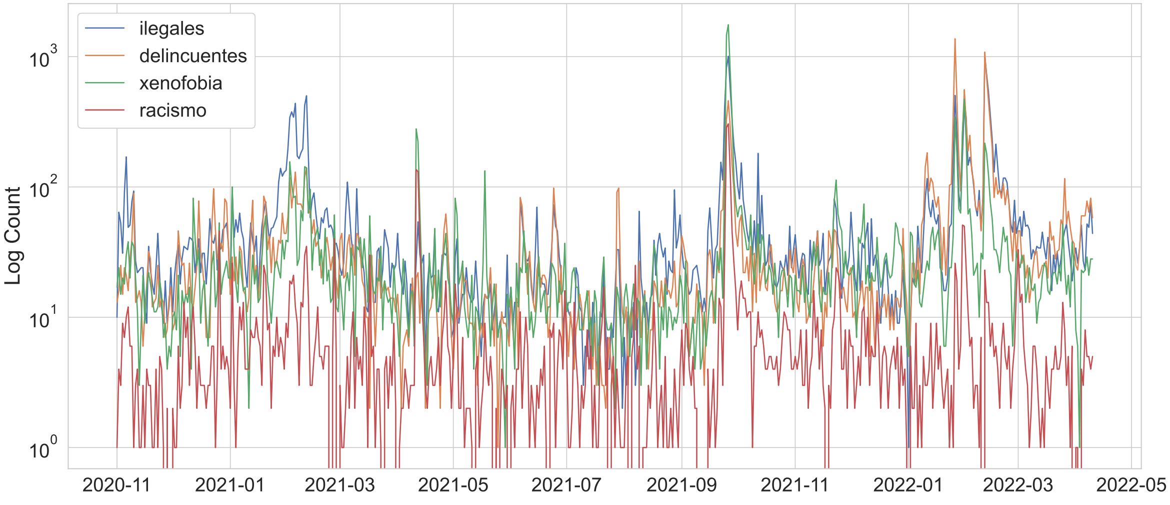
Figure A.1: Geotagged Tweets From Chile, 2012–2022



Notes: Red dashed line shows lower bound of 20,000 tweets.
Source: Twitter data downloaded using `twarc2` *API.*

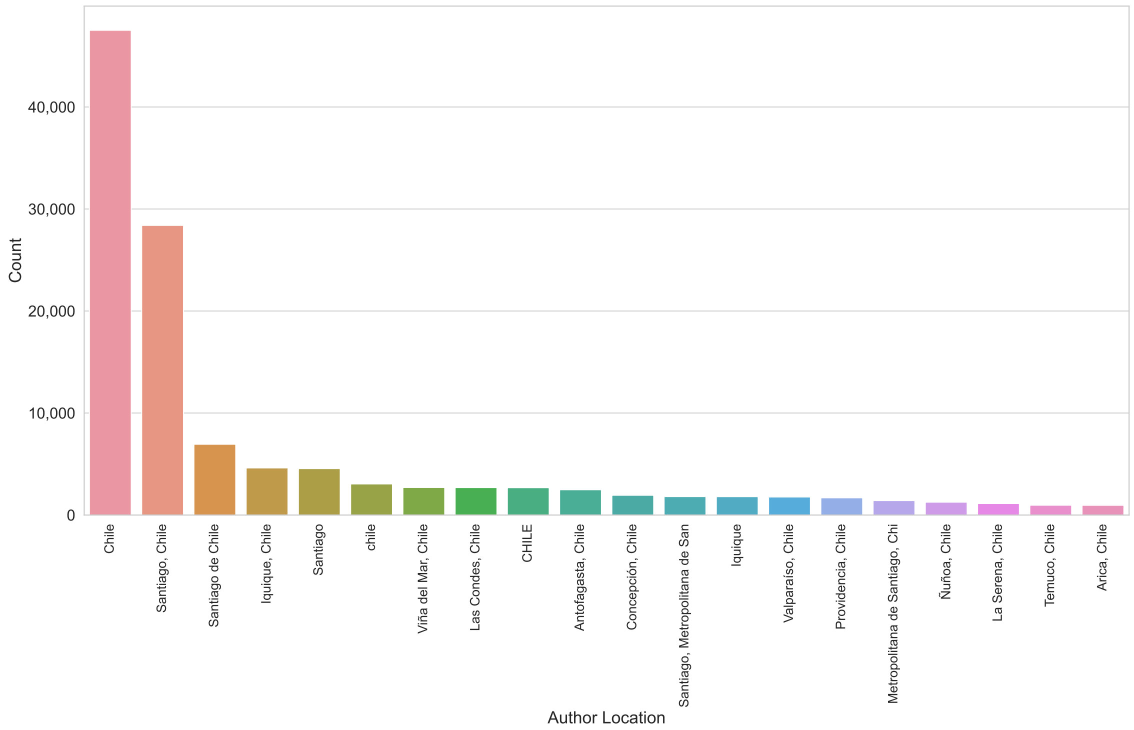
A.5 Figures

Figure A.5: Log Usage of Anti-Immigration and Anti-Xenophobia Terms; Total Corpus



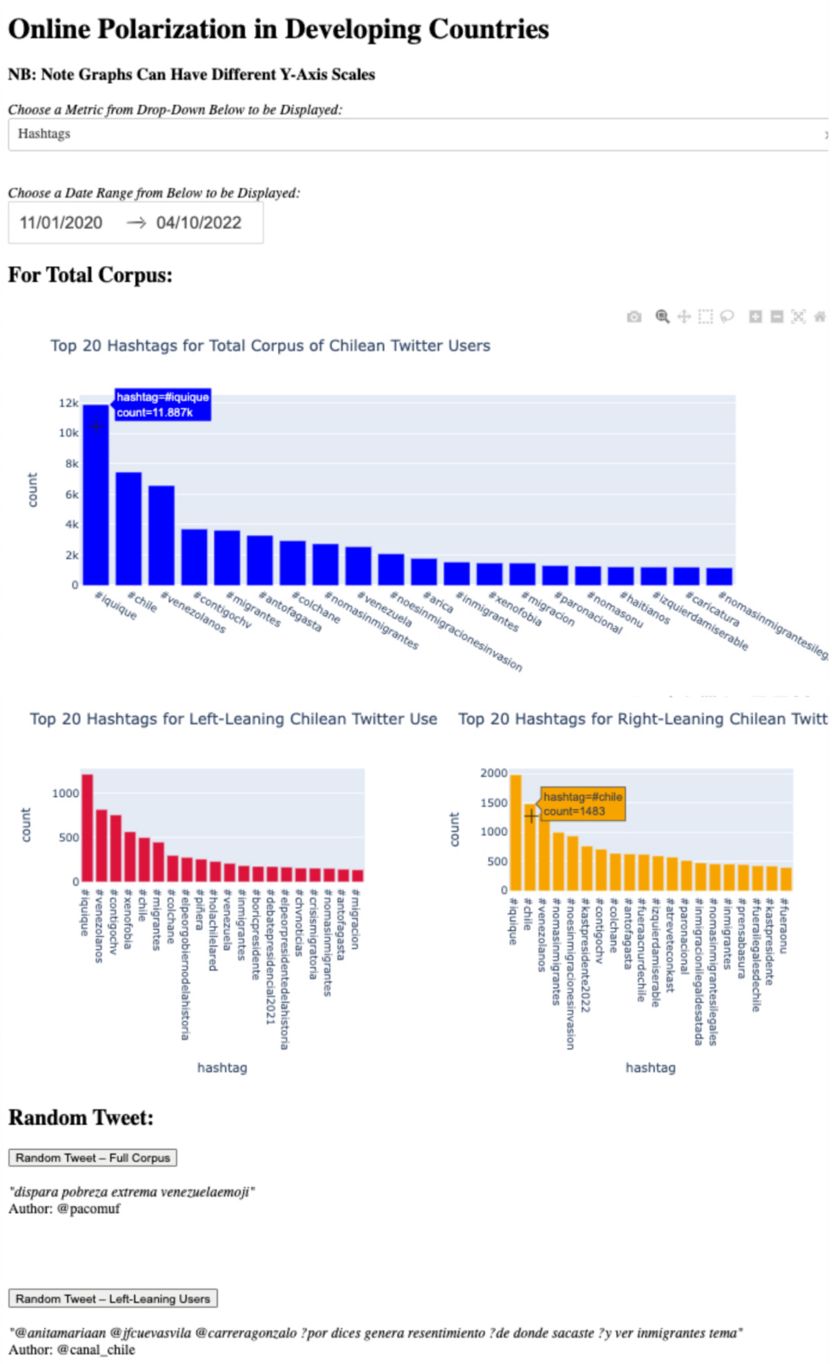
Notes: The anti-immigration terms plotted are 'ilegales' (Eng: Illegals) and 'delincuentes' (Eng: Criminals). The anti-xenophobia terms are 'xenofobia' (Eng: Xenophobia) and 'racismo' (Eng: Racism). Textual preprocessing steps include lowercase enforcement and converting Spanish special characters (e.g. á, ñ, etc.) into corresponding non-accentuated ones.
Data Source: Same as Figure 4.

Figure A.2: Top 20 Author Locations in Corpus After Filtering by Citizens



Data Source: Retrieved from Twitter API, spanning the time frame Nov 1st, 2020 – April 11th, 2022. Corpus includes tweets by Twitter users in Chile or Chilean nationals that mainly pertain to immigration.

Figure A.3: Screenshot of First Prototype of Interactive Dashboard



Online Polarization in Developing Countries

NB: Note Graphs Can Have Different Y-Axis Scales

Choose a Metric from Drop-Down Below to be Displayed:

Hashtags

Choose a Date Range from Below to be Displayed:

11/01/2020 → 04/10/2022

For Total Corpus:

Random Tweet:

Random Tweet – Full Corpus

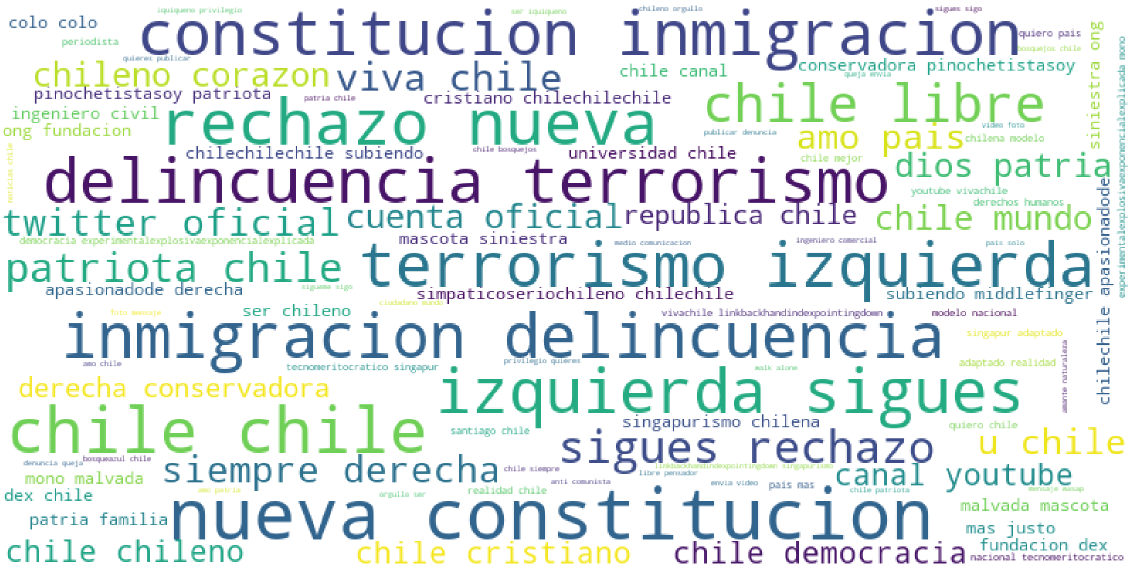
"dispara pobreza extrema venezuelaemoji"
Author: @pacomuf

Random Tweet – Left-Leaning Users

"@anitamariaan @jfcuevasvila @carreragonzalo ?por dices genera resentimiento ?de donde sacaste ?y ver inmigrantes tema"
Author: @canal_chile

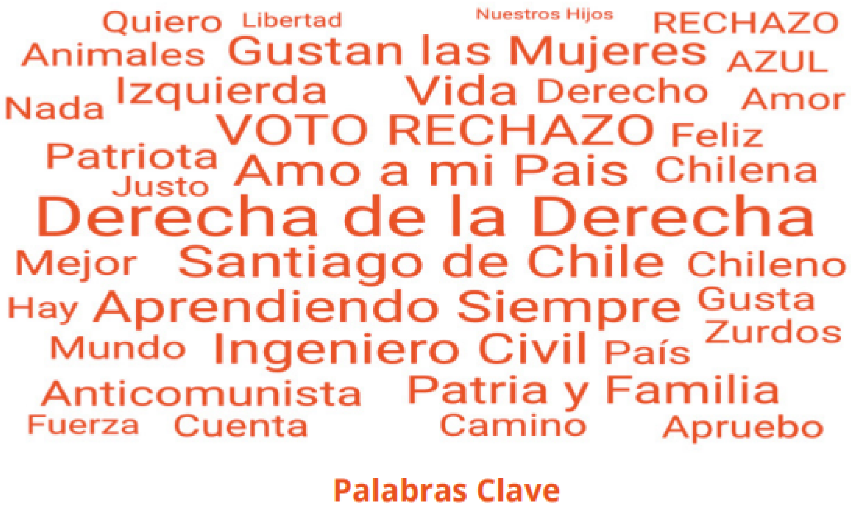
Notes: Dashboard coded using the Python packages `Plotly` *and* `Dash` *and using the corpus from Step 5 as input. Textual preprocessing steps in outputs include lowercase enforcement and converting Spanish special characters (e.g. á, ñ, etc.) into corresponding non-accentuated ones.*
Data Source: Retrieved from Twitter API, spanning the time frame Nov 1st, 2020 – April 11th, 2022. Corpus includes tweets by Twitter users in Chile or Chilean nationals that mainly pertain to immigration. Corpus cleaned by the steps in our methodology as described in Section 3.2 and Appendix A.2.

Figure A.4: Wordcloud of Author Biographies in Corpus After Filtering by Citizens



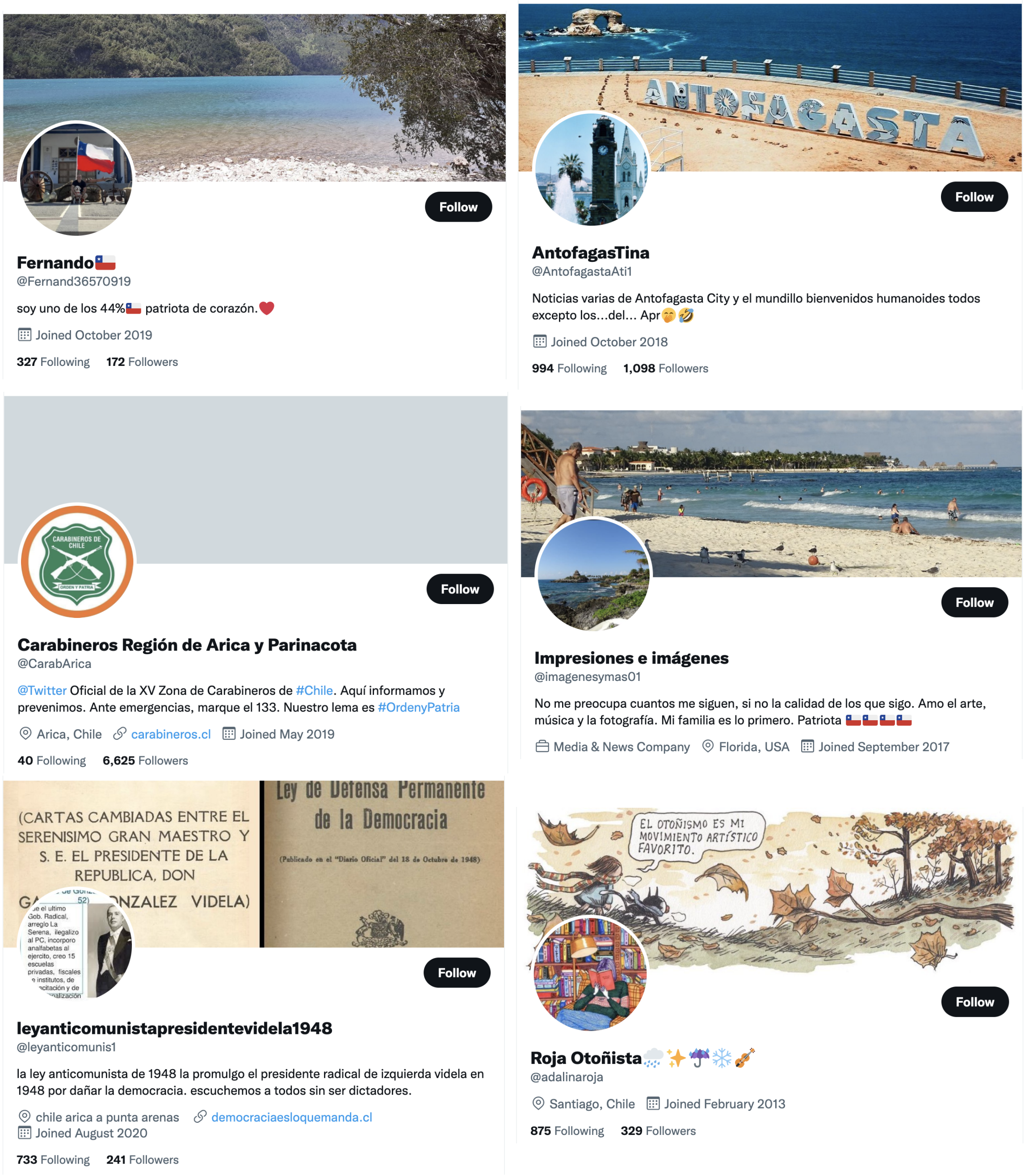
Data Source: Retrieved from Twitter API, spanning the time frame Nov 1st*, 2020 – April 11*th*, 2022. Corpus includes tweets by Twitter users in Chile or Chilean nationals that mainly pertain to immigration.*

Figure A.6: Wordcloud for Author Descriptions of Twitter Users That Use Discriminatory Language Towards Immigrants from the SJM Study



Source: Gálvez et al. (2020)

Figure A.7: Random Author Location and Descriptions after Step 3


Follow
ANTOFAGASTA
Follow
Follow
Follow
(CARTAS CAMBIADAS ENTRE EL SERENISIMO GRAN MAESTRO Y S. E. EL PRESIDENTE DE LA REPUBLICA, DON GA NZALEZ VIDELA)
Ley de Defensa Permanente de la Democracia
(Publicado en el "Diario Oficial" del 18 de Octubre de 1948)
Follow
EL OTOÑISMO ES MI MOVIMIENTO ARTÍSTICO FAVORITO.
Follow

Fernando🇨🇱
@Fernand36570919

soy uno de los 44%🇨🇱 patriota de corazón.❤️

Joined October 2019

327 Following **172** Followers

AntofagasTina
@AntofagastaAti1

Noticias varias de Antofagasta City y el mundillo bienvenidos humanoides todos excepto los...del... Apr🤭🤣

Joined October 2018

994 Following **1,098** Followers

Carabineros Región de Arica y Parinacota
@CarabArica

@Twitter Oficial de la XV Zona de Carabineros de #Chile. Aquí informamos y prevenimos. Ante emergencias, marque el 133. Nuestro lema es #OrdenyPatria

Arica, Chile carabineros.cl Joined May 2019

40 Following **6,625** Followers

Impresiones e imágenes
@imagenesymas01

No me preocupa cuantos me siguen, si no la calidad de los que sigo. Amo el arte, música y la fotografía. Mi familia es lo primero. Patriota 🇨🇱🇨🇱🇨🇱🇨🇱

Media & News Company Florida, USA Joined September 2017

leyanticomunistapresidentevidela1948
@leyanticomunis1

la ley anticomunista de 1948 la promulgo el presidente radical de izquierda videla en 1948 por dañar la democracia. escuchemos a todos sin ser dictadores.

chile arica a punta arenas democraciaesloquemanda.cl
Joined August 2020

733 Following **241** Followers

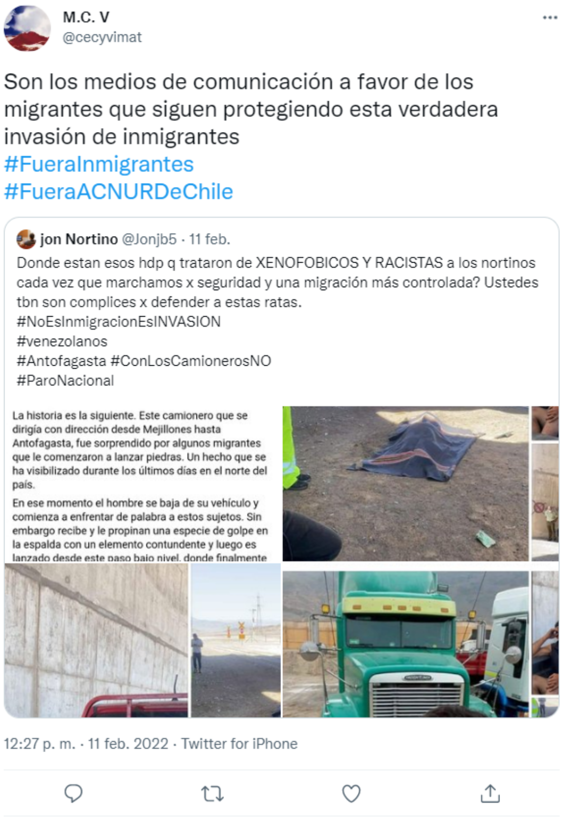
Roja Otoñista🌨️✨☔❄️🎻
@adalinaroja

Santiago, Chile Joined February 2013

875 Following **329** Followers

Data Source: Screenshots from Twitter. Retrieved from Twitter API, spanning the time frame Nov 1st*, 2020 – April 11*th*, 2022. Corpus includes tweets by Twitter users in Chile or Chilean nationals that mainly pertain to immigration. Corpus cleaned by the steps in our methodology as described in Section 3.2 and Appendix A.2.*

Figure A.8: Random Selection of Tweets, Total Corpus



M.C. V
@cecyvimat

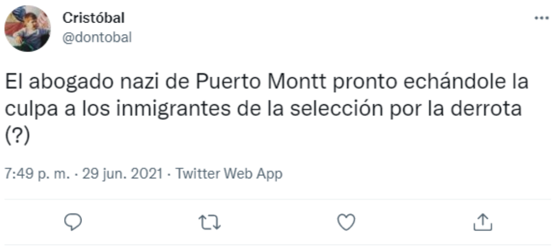
Son los medios de comunicación a favor de los migrantes que siguen protegiendo esta verdadera invasión de inmigrantes
#FueraInmigrantes
#FueraACNURDeChile

jon Nortino @Jonjb5 · 11 feb.
Donde estan esos hdp q trataron de XENOFOBICOS Y RACISTAS a los nortinos cada vez que marchamos x seguridad y una migración más controlada? Ustedes tbn son complices x defender a estas ratas.
#NoEsInmigracionEsINVASION
#venezolanos
#Antofagasta #ConLosCamionerosNO
#ParoNacional

La historia es la siguiente. Este camionero que se dirigía con dirección desde Mejillones hasta Antofagasta, fue sorprendido por algunos migrantes que le comenzaron a lanzar piedras. Un hecho que se ha visibilizado durante los últimos días en el norte del país.

En ese momento el hombre se baja de su vehículo y comienza a enfrentar de palabra a estos sujetos. Sin embargo recibe y le propinan una especie de golpe en la espalda con un elemento contundente y luego es lanzado desde este paso bajo nivel, donde finalmente

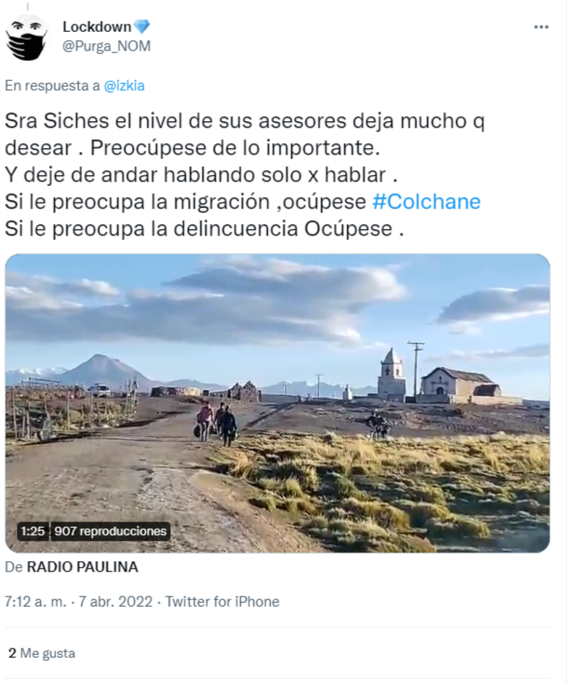
12:27 p. m. · 11 feb. 2022 · Twitter for iPhone



Cristóbal
@dontobal

El abogado nazi de Puerto Montt pronto echándole la culpa a los inmigrantes de la selección por la derrota (?)

7:49 p. m. · 29 jun. 2021 · Twitter Web App



Lockdown💎
@Purga_NOM

En respuesta a @izkia

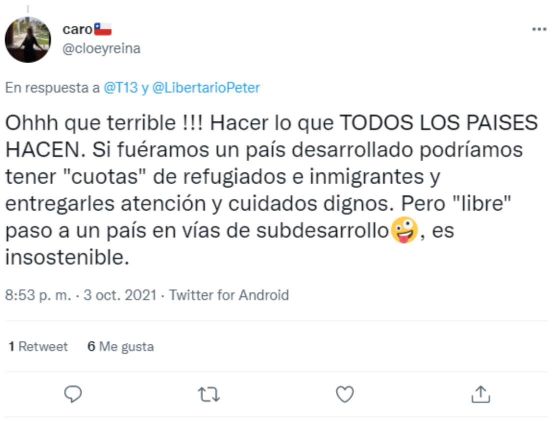
Sra Siches el nivel de sus asesores deja mucho q desear . Preocúpese de lo importante.
Y deje de andar hablando solo x hablar .
Si le preocupa la migración ,ocúpese #Colchane
Si le preocupa la delincuencia Ocúpese .

1:25 907 reproducciones

De RADIO PAULINA

7:12 a. m. · 7 abr. 2022 · Twitter for iPhone

2 Me gusta



caro🇨🇱
@cloeyreina

En respuesta a @T13 y @LibertarioPeter

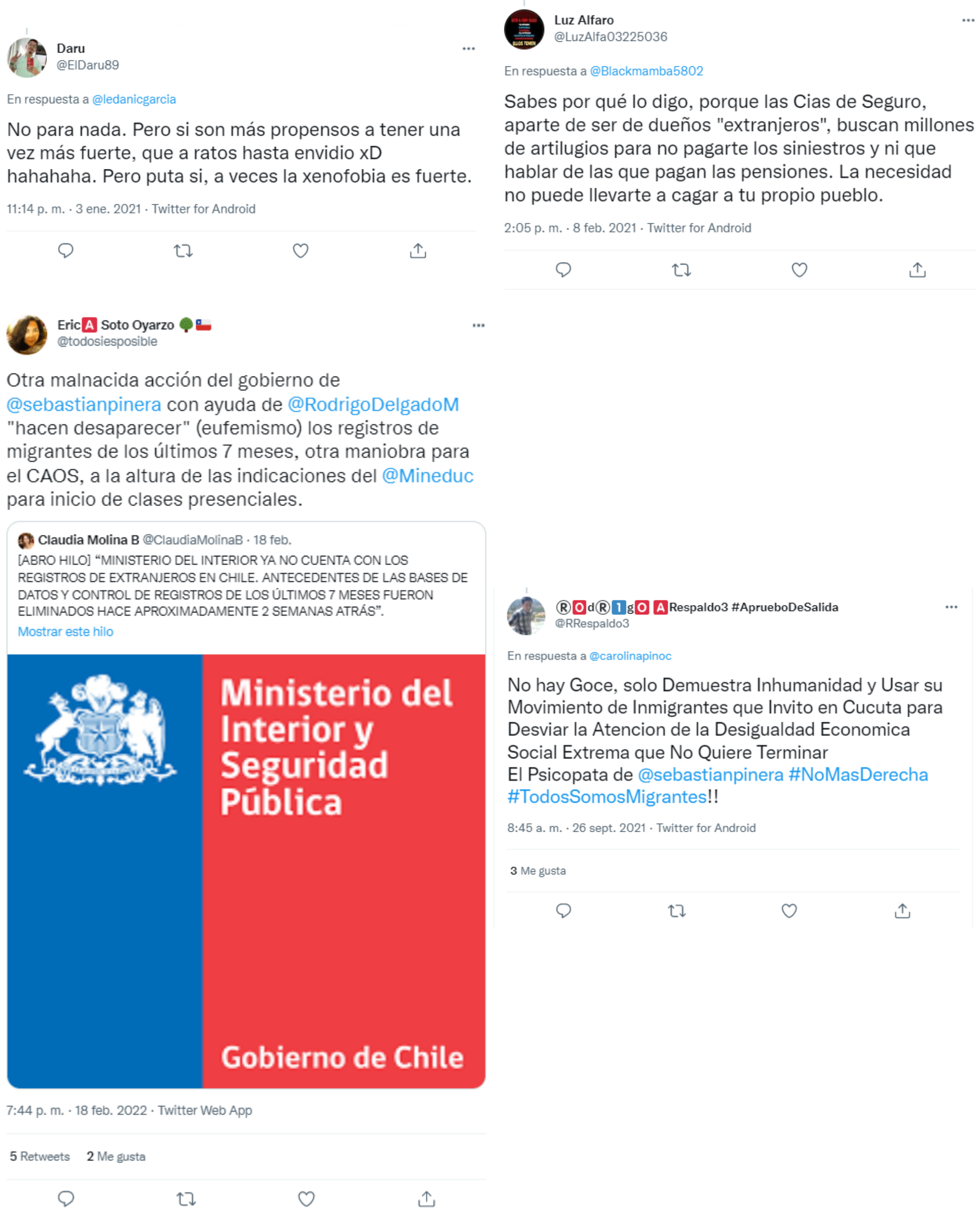
Ohhh que terrible !!! Hacer lo que TODOS LOS PAISES HACEN. Si fuéramos un país desarrollado podríamos tener "cuotas" de refugiados e inmigrantes y entregarles atención y cuidados dignos. Pero "libre" paso a un país en vías de subdesarrollo🤪, es insostenible.

8:53 p. m. · 3 oct. 2021 · Twitter for Android

1 Retweet 6 Me gusta

Data Source: Screenshots from Twitter. Corpus resulting after Step 3 in application of methodology to immigration in Chile as described in Appendix A.2. Corpus spans the time frame Nov 1st, 2020 – April 11th, 2022 and includes tweets by Twitter users in Chile or Chilean nationals that mainly pertain to immigration.

Figure A.9: Random Selection of Tweets, Left-Leaning Users



Daru
@ElDaru89

En respuesta a @ledanicgarcia

No para nada. Pero si son más propensos a tener una vez más fuerte, que a ratos hasta envidio xD hahahaha. Pero puta si, a veces la xenofobia es fuerte.

11:14 p. m. · 3 ene. 2021 · Twitter for Android

Eric🅰 Soto Oyarzo 🌳🇨🇱
@todosiesposible

Otra malnacida acción del gobierno de @sebastianpinera con ayuda de @RodrigoDelgadoM "hacen desaparecer" (eufemismo) los registros de migrantes de los últimos 7 meses, otra maniobra para el CAOS, a la altura de las indicaciones del @Mineduc para inicio de clases presenciales.

Claudia Molina B @ClaudiaMolinaB · 18 feb.
[ABRO HILO] "MINISTERIO DEL INTERIOR YA NO CUENTA CON LOS REGISTROS DE EXTRANJEROS EN CHILE. ANTECEDENTES DE LAS BASES DE DATOS Y CONTROL DE REGISTROS DE LOS ÚLTIMOS 7 MESES FUERON ELIMINADOS HACE APROXIMADAMENTE 2 SEMANAS ATRÁS".
Mostrar este hilo

7:44 p. m. · 18 feb. 2022 · Twitter Web App

5 Retweets 2 Me gusta

Luz Alfaro
@LuzAlfa03225036

En respuesta a @Blackmamba5802

Sabes por qué lo digo, porque las Cias de Seguro, aparte de ser de dueños "extranjeros", buscan millones de artilugios para no pagarte los siniestros y ni que hablar de las que pagan las pensiones. La necesidad no puede llevarte a cagar a tu propio pueblo.

2:05 p. m. · 8 feb. 2021 · Twitter for Android

ⓇOdⓇ1gO 🅰Respaldo3 #AprueboDeSalida
@RRespaldo3

En respuesta a @carolinapinoc

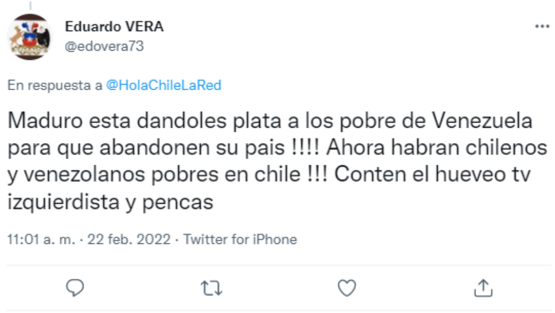
No hay Goce, solo Demuestra Inhumanidad y Usar su Movimiento de Inmigrantes que Invito en Cucuta para Desviar la Atencion de la Desigualdad Economica Social Extrema que No Quiere Terminar
El Psicopata de @sebastianpinera #NoMasDerecha #TodosSomosMigrantes!!

8:45 a. m. · 26 sept. 2021 · Twitter for Android

3 Me gusta

Data Source: Screenshots from Twitter. Corpus resulting after Step 3 in application of methodology to immigration in Chile as described in Appendix A.2. Corpus spans the time frame Nov 1st, 2020 – April 11th, 2022 and includes tweets by Twitter users in Chile or Chilean nationals that mainly pertain to immigration.

Figure A.10: Random Selection of Tweets, Right-Leaning Users

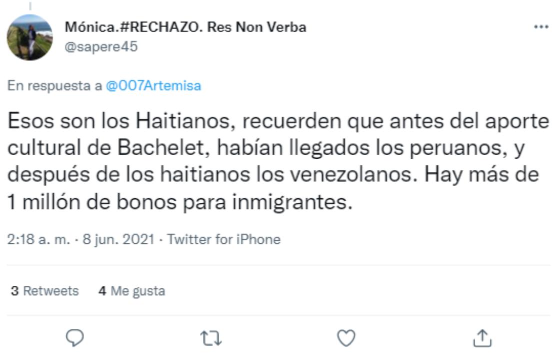


Eduardo VERA
@edovera73

En respuesta a @HolaChileLaRed

Maduro esta dandoles plata a los pobre de Venezuela para que abandonen su pais !!!! Ahora habran chilenos y venezolanos pobres en chile !!! Conten el hueveo tv izquierdista y pencas

11:01 a. m. · 22 feb. 2022 · Twitter for iPhone



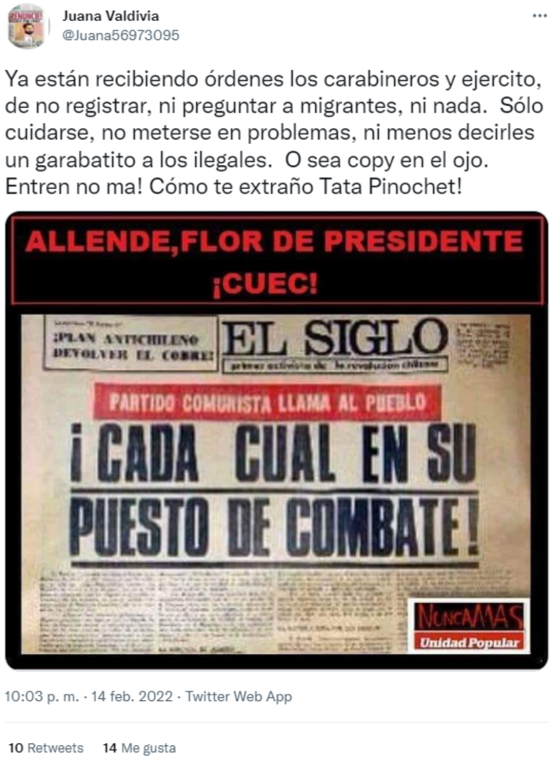
Mónica.#RECHAZO. Res Non Verba
@sapere45

En respuesta a @007Artemisa

Esos son los Haitianos, recuerden que antes del aporte cultural de Bachelet, habían llegados los peruanos, y después de los haitianos los venezolanos. Hay más de 1 millón de bonos para inmigrantes.

2:18 a. m. · 8 jun. 2021 · Twitter for iPhone

3 Retweets 4 Me gusta



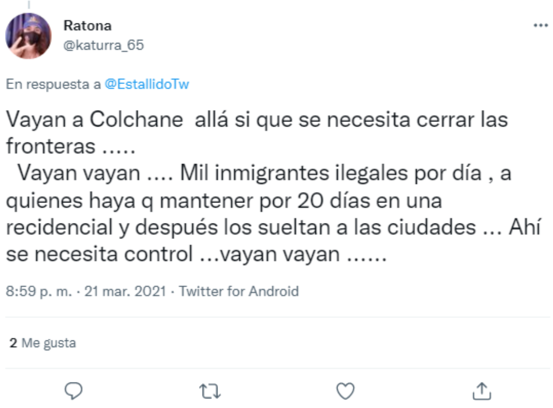
Juana Valdivia
@Juana56973095

Ya están recibiendo órdenes los carabineros y ejercito, de no registrar, ni preguntar a migrantes, ni nada. Sólo cuidarse, no meterse en problemas, ni menos decirles un garabatito a los ilegales. O sea copy en el ojo. Entren no ma! Cómo te extraño Tata Pinochet!

ALLENDE,FLOR DE PRESIDENTE
¡CUEC!
¡PLAN ANTICHILENO DEVOLVER EL COBRE!
EL SIGLO
PARTIDO COMUNISTA LLAMA AL PUEBLO
¡CADA CUAL EN SU PUESTO DE COMBATE!
NUNCA MAS
Unidad Popular

10:03 p. m. · 14 feb. 2022 · Twitter Web App

10 Retweets 14 Me gusta



Ratona
@katurra_65

En respuesta a @EstallidoTw

Vayan a Colchane allá si que se necesita cerrar las fronteras .....
Vayan vayan .... Mil inmigrantes ilegales por día , a quienes haya q mantener por 20 días en una recidencial y después los sueltan a las ciudades ... Ahí se necesita control ...vayan vayan ......

8:59 p. m. · 21 mar. 2021 · Twitter for Android

2 Me gusta

Data Source: Screenshots from Twitter. Corpus resulting after Step 3 in application of methodology to immigration in Chile as described in Appendix A.2. Corpus spans the time frame Nov 1st, 2020 – April 11th, 2022 and includes tweets by Twitter users in Chile or Chilean nationals that mainly pertain to immigration.

A.6 Tables

Table A.2: Top 15 Hashtags Count and Number of Users Using Each Hashtag

Hashtag	Count	No. users
#iquique	11,887	2,988
#chile	7,444	1,879
#venezolanos	6,566	2,001
#contigochv	3,704	1,473
#migrantes	3,620	1,061
#antofagasta	3,286	730
#colchane	2,939	976
#nomasinmigrantes	2,735	1,021
#venezuela	2,540	629
#noesinmigracionesinvasion	2,079	650
#arica	1,774	430
#inmigrantes	1,538	711
#xenofobia	1,473	746
#migracion	1,471	689
#paronacional	1,303	604

Data Source: Retrieved from Twitter API, spanning the time frame Nov 1st, 2020 – April 11th, 2022. Data is cleaned by our proposed methodology, such that the corpus includes tweets with topic-related keywords (semantic link + country contextual) and are tweeted by Twitter users in Chile or Chilean nationals and contains tweets that mainly regard the topic of immigration.

Table A.3: Top 15 Bigram Count and Number of Users Using Each Bigram

Bigram	Count	No. users
('inmigrantes', 'ilegales')	14,234	4,990
('inmigracion', 'ilegal')	11,852	4,433
('terrorismo', 'chile')	6,506	33
('terrorismo', 'izquierda')	4,904	11
('inmigracion', 'descontrolada')	4,816	2,468
('chile', 'terrorismo')	4,796	57
('delincuencia', 'terrorismo')	4,305	322
('inmigracion', 'delincuencia')	4,058	336
('venezolanos', 'chile')	3,745	2,683
('chilenos', 'extranjeros')	3,662	2,418
('migracion', 'ilegal')	3,548	1,848
('debe', 'ser')	3,179	2,401
('hace', 'anos')	3,158	2,235
('nueva', 'ley')	3,125	1,110
('delincuentes', 'extranjeros')	2,849	1,714

Data Source: Retrieved from Twitter API, spanning the time frame Nov 1st, 2020 – April 11th, 2022. Data is cleaned by our proposed methodology, such that the corpus includes tweets with topic-related keywords (semantic link + country contextual) and are tweeted by Twitter users in Chile or Chilean nationals and contains tweets that mainly regard the topic of immigration.

Table A.4: Top 15 Bigram Count and Number of Users Using Each Bigram; Left-Leaning Users during the Protest

Bigram	Count	No. users
('venir', 'chile')	104	89
('invito', 'venezolanos')	98	82
('venezolanos', 'venir')	96	81
('crisis', 'migratoria')	96	84
('migrantes', 'iquique')	91	72
('pertenencias', 'migrantes')	82	68
('inmigrantes', 'iquique')	73	65
('venezolanos', 'chile')	72	67
('invitar', 'venezolanos')	70	65
('xenofobia', 'racismo')	69	61
('pinera', 'cucuta')	67	58
('racismo', 'xenofobia')	66	52
('pertenencias', 'inmigrantes')	65	57
('coches', 'panales')	64	8
('migrantes', 'venezolanos')	62	53

Data Source: Retrieved from Twitter API, spanning the time frame Nov 1st, 2020 – April 11th, 2022, filtered to consider only left-wing users during the protest.

Table A.5: Top 15 Bigram Count and Number of Users Using Each Bigram - Right-Leaning Users during the Protest

Bigram	Count	No. users
('inmigrantes', 'ilegales')	1,176	541
('inmigracion', 'ilegal')	963	490
('inmigracion', 'descontrolada')	290	210
('migracion', 'ilegal'))	217	141
('migrantes', 'ilegales')	182	132
('inmigrantes', 'venezolanos')	175	94
('congreso', 'cc')	142	3
('millones', 'venezolanos')	141	118
('debe', 'ser')	130	111
('nicolas', 'maduro')	117	67
('venezolanos', 'chile')	112	84
('deben', 'ser')	104	78
('dictador', 'maduro')	92	53
('gente', 'q')	92	25
('chilenos', 'extranjeros')	90	76

Data Source: Retrieved from Twitter API, spanning the time frame Nov 1st, 2020 – April 11th, 2022, filtered to consider only right-wing users during the protest.

Table A.6: Top 15 Bigram Count and Number of Users Using Each Bigram - Unlabeled Users during the Protest

Bigram	Count	No. users
('inmigracion', 'ilegal')	854	540
('inmigrantes', 'ilegales')	834	546
('crisis', 'migratoria')	382	270
('leyes', 'migratorias')	357	12
('migrantes', 'pinera')	347	3
('marcha', 'encontra')	346	3
('pinera', 'leyes')	345	1
('migratorias', 'impuestas')	344	1
('impuestas', '@onuchile')	344	1
('encontra', 'migrantes')	343	1
('migrantes', 'iquique')	306	219
('migracion', 'ilegal')	282	223
('venezolanos', 'chile')	270	229
('plaza', 'brasil')	265	172
('debe', 'ser')	263	236

Data Source: Retrieved from Twitter API, spanning the time frame Nov 1st, 2020 – April 11th, 2022, filtered to consider only unlabeled users during the protest.

Table A.7: Top 15 Hashtags Count and Number of Users Using Each Hashtag - Left-Leaning Users during the Protest

Hashtag	Count	No. users
#iquique	498	276
#xenofobia	422	189
#migrantes	182	89
#contigochv	175	104
#nomasinmigrantes	100	72
#chile	87	59
#inmigrantes	87	56
#verguenzanacional	84	66
#crisismigratoria	75	28
#racismo	71	19
#venezolanos	70	38
#piñera	54	29
#elpeorgobiernodelahistoria	53	32
#iquiquemedasverguenza	49	33
#colchane	44	25

Data Source: Retrieved from Twitter API, spanning the time frame Nov 1st, 2020 – April 11th, 2022, filtered to consider only left-wing users during the protest.

Table A.8: Top 15 Hashtags Count and Number of Users Using Each Hashtags - Right-Leaning Users during the Protest

Hashtag	Count	No. users
#nomasinmigrantes	467	208
#iquique	381	162
#nomasinmigrantesilegales	274	141
#atreveteconkast	212	137
#iquiquedicebasta	185	92
#xenofobia	185	77
#contigochv	169	95
#chile	166	67
#kastpresidente2022	135	81
#nomasinmigracionilegal	115	71
#colchane	110	67
#venezolanos	108	60
#inmigrantes	106	59
#kastpresidente	86	57
#izquierdamiserable	85	53

Data Source: Retrieved from Twitter API, spanning the time frame Nov 1st, 2020 – April 11th, 2022, filtered to consider only right-wing users during the protest.

Table A.9: Top 15 Hashtags Count and Number of Users Using Each Hashtags - Unlabeled Users during the Protest

Hashtag	Count	No. users
#iquique	1,694	809
#chile	845	292
#nomasinmigrantes	820	293
#contigochv	508	309
#nomasinmigrantesilegales	504	115
#xenofobia	475	325
#nomasonu	441	7
#migrantes	417	254
#fueraonu	377	15
#nomasabuso	344	1
#venezolanos	296	139
#colchane	292	143
#inmigrantes	209	154
#antofagast	173	52
#venezuela	159	87

Data Source: Retrieved from Twitter API, spanning the time frame Nov 1st, 2020 – April 11th, 2022, filtered to consider only unlabeled users during the protest.

Table A.10: Five Most Influential Users by Eigenvector Centrality

User	Eigenvector Centrality	Label
@joseantoniokast	0.200379	Right
@AldoDuqueSantos	0.175374	Unlabeled
@T13	0.141765	Unlabeled
@biobio	0.135003	Unlabeled
@Florencia_Pink	0.132579	Right

Data Source: Retweets network retrieved from Twitter API and twarc2's network plug-in. Result of the step 8 of the applied methodology as described in Section 3.2 and Appendix A.2.

Table A.11: Degree Centrality by Political Affiliations

User	Degree Centrality	Label
Panel A. Left-Leaning Users		
@NachoOrtega	0.041077	Left
@RodriguezManuel	0.025305	Left
@baradit	0.024866	Left
@danieljadue	0.024449	Left
@LuisErrazuriz	0.022098	Left
@MattyLL	0.0181	Left
@gabrielboric	0.017925	Left
@PaulinaAstrozaS	0.017485	Left
@jgalemparte	0.017441	Left
@El_Ciudadano	0.017178	Left
Panel B. Right-Leaning Users		
@joseantoniokast	0.131623	Right
@Florencia_Pink	0.059617	Right
@AlejandroMery1	0.057047	Right
@NatyDerecha	0.056476	Right
@camilaemiliasv	0.054323	Right
@Alberto85366967	0.052302	Right
@Francis25830521	0.052236	Right
@cherrAL62	0.051885	Right
@carreragonzalo	0.046876	Right
@lamonsalveg	0.046503	Right

Data Source: Retweets network retrieved from Twitter API and twarc2's network plug-in. Result of Step 8 of the applied methodology as described in Section 3.2 and Appendix A.2.

Table A.13: Unigram Chilean Cities, Excluded and Included in Step 3 of Applied Methodology

Excluded	Included
aguila	ancud

Continued On Next Page...

Table A.13 – Continued From Previous Page

Excluded	Included
aguirre	andacollo
agustin	angol
alamos	antofagasta
alegre	araucania
alemana	arauco
algarrobo	arica
almagro	atacama
almonte	aysen
alto	barnechea
amarilla	biobio
andes	buin
angeles	calama
antonio	calbuco
arenas	calera
bajos	canete
barbara	carahue
bernardo	catemu
bosque	cauquenes
bueno	cerrillos
bulnes	chanaral
cabildo	chiguayante
cabras	chillan
cabrero	chimbarongo
caldera	chuquicamata
calle	codegua
carlos	coelemu
cartagena	coihueco
casablanca	collipulli
casas	combarbala
castro	conchali
central	concon
cerda	copiapo
cerro	coquimbo
cisterna	coyhaique
clemente	cunco
colina	curacautin
concepcion	curacavi
condes	curanilahue
constitucion	curico
coronel	donihue

Continued On Next Page...

Table A.13 – Continued From Previous Page

Excluded	Included
cruz	frutillar
diego	futrono
domingo	graneros
elena	gultro
espejo	hijuelas
estacion	hualane
esteban	hualpen
felipe	hualqui
fernando	huasco
florida	huechuraba
freire	huepil
fresia	illapel
gorbea	iquique
granja	lanco
hospicio	lebu
hospital	ligua
hurtado	limache
imperial	llanquihue
independencia	llay-llay
isla	loncoche
islita	longavi
jahuel	machali
javier	macul
joaquin	maipo
jose	maipu
juana	mariquina
labranza	maule
lagos	melipilla
laja	muermos
lampa	mulchen
larga	nancagua
lautaro	nunoa
linares	o'higgins
lota	olmue
magallanes	ovalle
mar	paillaco
maria	paine
mejillones	panguipulli
melon	parinacota
metropolitan	penalolen

Continued On Next Page...

Table A.13 – Continued From Previous Page

Excluded	Included
miguel	pichilemu
miranda	pintana
molina	pirque
monte	pitrufquen
montt	placilla
mostazal	pucon
nacimiento	pudahuel
natales	puren
navia	purranque
negro	putaendo
nogales	quellon
normal	quilicura
nueva	quillon
osorno	quillota
padre	quilpue
palmilla	quirihue
palqui	quisco
parral	rancagua
patria	renaico
paz	renca
pedro	requinoa
penaflor	santiago
penco	talagante
penuelas	talca
peumo	talcahuano
pozo	taltal
prado	tarapaca
providencia	temuco
puente	tilcoco
puerto	tiltil
punta	tocopilla
quinta	traiguen
quintero	valdivia
ramon	vallenar
recoleta	valparaiso
reina	villarrica
rengo	vilos
rinconada	vitacura
rio	yumbel
rios	

Continued On Next Page...

Table A.13 – Continued From Previous Page

Excluded	Included
rosendo	
salamanca	
salvador	
santa	
santo	
serena	
tabo	
tagua	
teno	
tierra	
tome	
union	
varas	
ventanas	
vicente	
victoria	
vicuna	
viejo	
villa	
vina	
yungay	
zaldivar	

Source: Wikipedia entry on "List of cities in Chile". Retrieved on June 15, 2022, 19.41.

Table A.14: n-Gram Chilean Cities, Excluded and Included in Step 3 of Applied Methodology

Excluded	Included
algarrobo	alto hospicio
bulnes	alto jahuel
cabildo	ancud
caldera	andacollo
calle larga	angol
cartagena	antofagasta
casablanca	araucania
castro	arauco
colina	arica
concepcion	arica and parinacota region
constitucion	atacama
coronel	aysen

Continued On Next Page...

Table A.14 – Continued From Previous Page

Excluded	Included
diego de almagro	bajos de san agustin
el bosque	biobio
el melon	buin
el monte	cabrero
el salvador	calama
estacion central	calbuco
freire	canete
fresia	carahue
gorbea	catemu
graneros	cauquenes
hospital	cerrillos
independencia	cerro navia
la calera	chanaral
la cruz	chiguayante
la florida	chillan
la granja	chillan viejo
la islita	chimbarongo
la laja	chuquicamata
la reina	codegua
la serena	coelemu
la union	coihueco
labranza	collipulli
lampa	combarbala
las cabras	conchali
las ventanas	concon
lautaro	copiapo
linares	coquimbo
lo espejo	coyhaique
los alamos	cunco
los andes	curacautin
los angeles	curacavi
los lagos	curanilahue
los rios	curico
lota	donihue
magallanes	el palqui
maria elena	el quisco
mejillones	el tabo
molina	estacion zaldivar
mostazal	frutillar
nacimiento	futrono

Continued On Next Page...

Table A.14 – Continued From Previous Page

Excluded	Included
nogales	gultro
osorno	hijuelas
padre hurtado	hualane
palmilla	hualpen
pedro aguirre cerda	hualqui
penaflor	huasco
penco	huechuraba
peumo	huepil
providencia	illapel
quintero	iquique
recoleta	isla de maipo
rengo	la cisterna
rio bueno	la ligua
rio negro	la pintana
salamanca	lanco
san antonio	las condes
san bernardo	lebu
san carlos	limache
san clemente	llanquihue
san esteban	llay-llay
san felipe	lo barnechea
san fernando	lo miranda
san javier	lo prado
san joaquin	loncoche
san miguel	longavi
san ramon	los muermos
san rosendo	los vilos
santa barbara	machali
santa cruz	macul
santa juana	maipu
santa maria	mariquina
santo domingo	maule
teno	melipilla
tome	monte aguila
victoria	monte patria
vicuna	mulchen
yungay	nancagua
	nueva imperial
	nunoa
	o'higgins

Continued On Next Page...

Table A.14 – Continued From Previous Page

Excluded	Included
	olmue
	ovalle
	padre las casas
	paillaco
	paine
	panguipulli
	parral
	penalolen
	pichilemu
	pirque
	pitrufquen
	placilla de penuelas
	pozo almonte
	pucon
	pudahuel
	puente alto
	puerto aysen
	puerto montt
	puerto natales
	puerto varas
	punta arenas
	puren
	purranque
	putaendo
	quellon
	quilicura
	quillon
	quillota
	quilpue
	quinta de tilcoco
	quinta normal
	quirihue
	rancagua
	renaico
	renca
	requinoa
	rinconada
	san jose de maipo
	san pedro de la paz
	san vicente de tagua tagua

Continued On Next Page...

Table A.14 – Continued From Previous Page

Excluded	Included
	santiago
	santiago metropolitan
	santiago metropolitan region
	talagante
	talca
	talcahuano
	taltal
	tarapaca
	temuco
	tierra amarilla
	tiltil
	tocopilla
	traiguen
	valdivia
	vallenar
	valparaiso
	villa alegre
	villa alemana
	villarrica
	vina del mar
	vitacura
	yumbel

Source: Wikipedia entry on "List of cities in Chile". Retrieved on June 15, 2022, 19.41.

Table A.12: Eigenvector Centrality by Political Affiliations

User	Eigenvector Centrality	Label
Panel A. Left-Leaning Users		
@LuisErrazuriz	0.012359	Left
@Lucianoabrahamm	0.010078	Left
@gabrielboric	0.008336	Left
@renenaranjo	0.006337	Left
@eveoca	0.006158	Left
@danieljadue	0.004129	Left
@RodriguezManuel	0.003813	Left
@PaulinaAstrozaS	0.003275	Left
@NachoOrtega	0.002976	Left
@Vitalicio7020	0.002808	Left
Panel B. Right-Leaning Users		
@joseantoniokast	0.200379	Right
@Florencia_Pink	0.132579	Right
@AlejandroMery1	0.11526	Right
@Francis25830521	0.113777	Right
@lamonsalveg	0.110292	Right
@AlboradaDeChile	0.10901	Right
@cherrAL62	0.107874	Right
@aprachile	0.107562	Right
@NatyDerecha	0.106167	Right
@camilaemiliasv	0.106151	Right

Data Source: Retweets network retrieved from Twitter API and twarc2*'s network plug-in. Result of Step 8 of the applied methodology as described in Section 3.2 and Appendix A.2.*

B Documentation of TextAnLib Library

NB: See also https://github.com/BSE-DSDM-2022/ChileGov/blob/master/methodology_blank/z_explore_data/TxtAnLib/README.md for nicer formatting.

B.1 Functions to Filter Corpus

f_authors (DF: pd.DataFrame, authors_list: list): Receive as input a list of authors and returns a data frame with the tweets from these authors.
Parameters:
DF: Data frame to filter; Type: Pandas Dataframe
author_list: list of authors usernames to select the tweets from these users; Type: List of strings.
Return: DataFrame with the same columns that the input DF, but considering only the tweets written by authors in the input list.

f_dates(DF: pd.DataFrame, start_time =”2020-10-31”, end_time = ”2022-04-12”): Receive as input a start time and end time and return the data frame only with the tweets tweeted

between these two dates.
Parameters:
DF: Data frame to filter; Type: Pandas Dataframe
start_time: Initial date of the period to filter; Type: String in format YYYY-MM-DD; Default: ”2020-10-31”
end_time: Final date of the period to filter; Type: String in format YYYY-MM-DD; Default: ”2022-04-12”
Return:DataFrame with the same columns that the input DF, but considering only the tweets written in the period between start_time and end_time.

f_words(DF: pd.DataFrame, list_of_words: list): Receive a list of words and return a data frame only with the tweets that contain one of these words
Parameters:
DF: Data frame to filter; Type: Pandas Dataframe
list_of_words: list of words to select the tweets that contains one of these words; Type: List of strings
Return: DataFrame with the same columns that the input DF, but considering only the tweets that contain at least one of the words in list_of_words.

f_hashtags(DF: pd.DataFrame, list_of_hashtags: list) Receive a list of hashtags and return a data frame only with the tweets that contain one of these hashtags.
Parameters:
DF: Data frame to filter; Type: Pandas Dataframe
list_of_hashtags list of authors words to select the tweets that contains one of these hashtags; Type: List of strings
Return: DataFrame with the same columns that the input DF, but considering only the tweets that contain at least one of the hashtags in list_of_hashtags.

f_metrics(DF: pd.DataFrame, metric: str , threshold: int): Filter tweets with a minimum number of RT, likes, quotes or resplies.
Parameters:
DF: Data frame to filter; Type: Pandas Dataframe
metric: Name of the metric that the user wants to use to filter. Options: ”Likes”, ”Retweets”, ”Quotes” or ”Replies”; Type: String
threshold: Minimum number of the “metric” to filter tweets.
Return: DataFrame with the same columns that the input DF, but considering only the tweets that have more than the threshold number of the selected metric.

f_location(DF: pd.DataFrame, list_of_locations: list) Receive a list of words and return a data frame only with tweets from authors that have one of these words in the location.
Parameters:
DF: Data frame to filter; Type: Pandas Dataframe
list_of_location: list of words to select the tweets from authors that contain at least one of these words in their location; Type: List of strings

Return: DataFrame with the same columns that the input DF, but considering only the tweets written by authors that contain in their location at least one of the words of the list of locations.

f_affiliation(DF,Affiliation="Unlabeled"): Receive an affiliation (Left, Right or Unlabeled) and return a DataFrame only with tweets from authors of this affiliation.
Parameters:
DF: Data frame to filter; Type: Pandas Dataframe
Affiliation: Affiliation to filter. Options: "Left", "Right" or "Unlabeled". Default:"Unlabeled";Type: String.
Return: DataFrame with the same columns that the input DF, but considering only the tweets written by authors of the selected affiliation.

f_verified(DF) Return a Dataframe with tweets written by verified accounts.
Parameters:
DF: Data frame to filter; Type: Pandas Dataframe
Return: DataFrame with the same columns that the input DF, but considering only the tweets written by verified accounts.

f_minfollowers(DF,min_followers=200) Return a Dataframe with tweets written by accounts with a minimum number of followers.
Parameters:
DF: Data frame to filter; Type: Pandas Dataframe
min_followers: Minimum number of followers to filter the data frame. Default:200; Type:Int
Return: DataFrame with the same columns that the input DF, but considering only the tweets written by accounts with more than min_followers followers.

B.2 Functions for Visualization

word_cloud(DF: pd.DataFrame, number_of_words = 100): Plot a Word Cloud of the input data set cleaned text.
Parameters:
DF: Data frame with input data; Type: Pandas Dataframe
number_of_words: Maximum number of words to display in the Word Cloud. Default: 100;Type: Int
Return: None
Action: Plot a Word Cloud of cleaned text from the DF.

top_words(DF: pd.DataFrame, number_of_words = 100): Print the n most used words in the Data Frame and the number of occurrences.
Parameters:
DF: Data frame with input data; Type: Pandas Dataframe
number_of_words: Number of most used words to display. Default: 100;Type: Int
Return: List of n tuples with the most used words and the number of occurrences.
Action: Print the n most used words and the number of occurrences.

top_authors(DF: pd.DataFrame, number_of_authors = 100): Print n authors username that tweets more often in the DF.
Parameters:
DF: Data frame with input data; Type: Pandas Dataframe
number_of_authors: Number of most common authors to display. Default: 100;Type: Int
Return: List of n tuples with the n authors that tweet most often in DF and the number of tweets that each one has.
Action: Print the n authors that tweeted more and the number of tweets for each one.

top_hashtags(DF: pd.DataFrame,number_print=100,number_plot=20, title = ' '): Print the n_print most used hashtags in the Data Frame and the number of occurrences. Also plot a bar plot with the n_plot top hashtags.
Parameters:
DF: Data frame with input data; Type: Pandas Dataframe
number_print: Number of most used hashtags to print. Default:100;Type:int. number_plot: Number of most used hashtags to include in the bar plot. Default:20;Type:int. title: Title for the bar plot. Default:" "; Type: String.
Return: List of n_print tuples with the n most used hashtags and the number of occurrences.
Action: Print the n_print most used hashtags in the Data Frame and the number of occurrences. Also display a bar plot with the n_plot number of most used hashtags. Each bar represent the number of occurrences.

top_bigrams(DF:pd.DataFrame,number_of_bigrams=100,number_plot=20, title = "", split_bi = False): Print the n_print most used bigrams in the Data Frame and the number of occurrences. Also plot a bar plot with the n_plot top bigrams.
Parameters:
DF: Data frame with input data; Type: Pandas Dataframe
number_of_bigrams: Number of most used bigrams to print. Default:100;Type:int. number_plot: Number of most used bigrams to include in the bar plot. Default:20;Type:int. title: Title for the bar plot. Default:" "; Type: String.
Split_bi: Boolean to indicate if in the labels of the bar plot the bigrams are presented in two lines (True) or in one line (False). Default: False; Type: Boolean
Return: List of n_print tuples with the n most used bigrams and the number of occurrences.
Action: Print the n_print most used bigrams in the Data Frame and the number of occurrences. Also display a bar plot with the n_plot number of most used bigrams. Each bar represent the number of occurrences.

pday_tweets(DF: pd.DataFrame): Print a time series indicating the number of tweets per day.
Parameters:
DF: Data frame with input data; Type: Pandas Dataframe

Return: None
Action: Display time series plot indicating the number of tweets per day.

pday_metrics(DF: pd.DataFrame, RT = True, Likes = True, Quotes = True, Reply = True): Print time series for each indicated metric with the number of these metrics per tweet for each day.
Parameters:
DF: Data frame with input data; Type: Pandas Dataframe
RT: If True, display the time series of retweets per tweet. Default: True; Type: Boolean
Likes: If True, display the time series of likes per tweet. Default: True; Type: Boolean
Quotes: If True, display the time series of quotes per tweet. Default: True; Type: Boolean
Reply: If True, display the time series of replies per tweet. Default: True; Type: Boolean
Return: None
Action: Display time series plots indicating the number of the selected metrics per tweet for each day.

pday_hashtags(DF: pd.DataFrame, list_of_hashtags: list): Plot one time series for each hashtag in the list, showing the number of tweets per day that contain the hashtags. All the time series are displayed together in the same plot.
Parameters:
DF: Data frame with input data; Type: Pandas Dataframe
list_of_hashtags: List of hashtags that the user wants to see the time series; Type: List of strings. It is not necessary to include the # before the hashtag.
Return: None
Action: Display time series plot showing the number of tweets that contains each of the hashtags in the list.

pday_word(DF: pd.DataFrame, list_of_words: list): Plot one time series for each word in the list, showing the number of tweets per day that contain the word. All the time series are displayed together in the same plot.
Parameters:
DF: Data frame with input data; Type: Pandas Dataframe
list_of_words: List of words that the user wants to see the time series; Type: List of strings.
Return: None
Action: Display time series plot showing the number of tweets that contains each of the words in the list.

popular_tweets(DF: pd.DataFrame, metric = "Retweets", number = 10): Print the n most popular tweets. Popularity is measured as the tweets with the highest number of reactions of the selected metric.
Parameters:
DF: Data frame with input data; Type: Pandas Dataframe
metric: Name of the metric that the user wants to use to filter. Options: "Likes", "Retweets", "Quotes" or "Replies". Default: "Retweets"; Type: String

Number: Number of tweets to print. Default: 10; Type: Int
Return: None
Action: Print the n tweets most popular according to the selected metric.

compare_word_pday (DF,list_of_words,standarized=True): Plot time series with the number (or proportion) of tweets per day that contain at least one of the words of the list for right leaning and left leaning people. Both time series are displayed in the same plot to compare.
Parameters:
DF: Data frame with input data; Type: Pandas Dataframe list_of_words: List of words that the user wants to see their use on time for right and left leaning people. Type: List of strings standardized: If it is True, display the proportion of tweets that contain the words per day, if it is False, display the total number of tweets that contain these words. Default: True; Type: Boolean.
Return: None
Action: Time series with the use of the selected words over time for left- and right-wing users.

C Test of Methodology for Different Topic

To test our methodology for corpus construction, we select a different topic and a new period of time. We select the topic feminism and the period of time between March 9th and 11th, 2022 (we selected a short period of time to have quick results). After, we also tested the scripts to update the data considering the period from March 11th to 13th, 2022. So our final data set have the tweets that talk about feminism between March 9th to 13th, 2022. Here we present the final characteristics of this corpus and the same plots that we used to describe the immigration corpus.

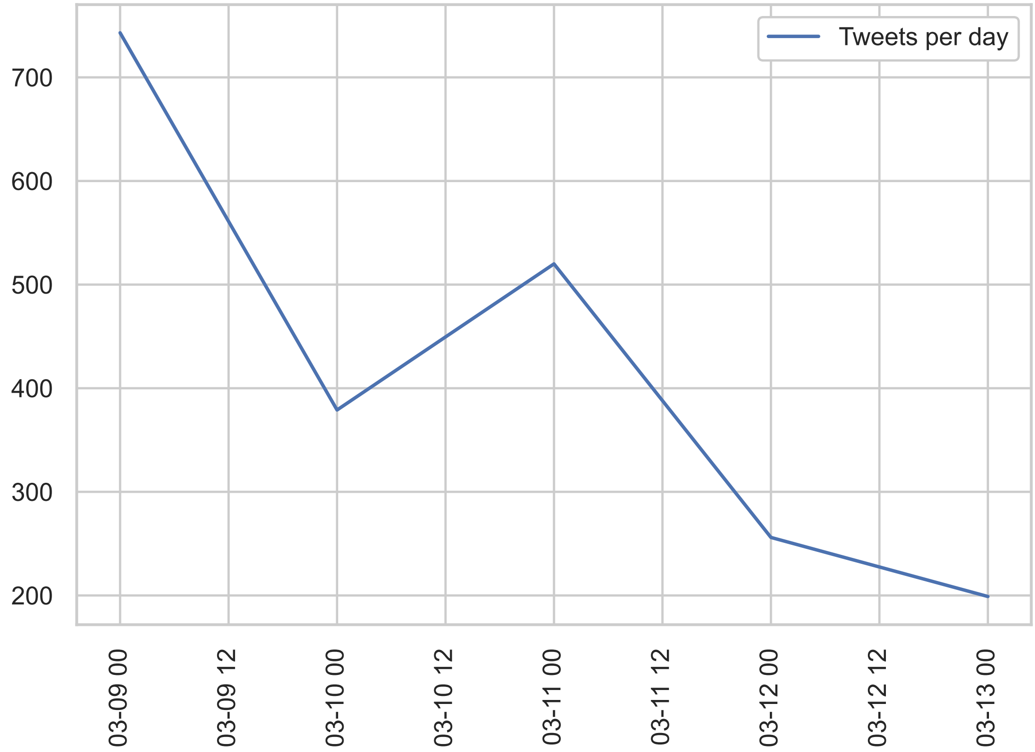
Table C.1: Corpus Characteristics, Feminism Corpus and Labeled Subsample

Panel A. Main Corpus for Feminism			
Measure	Count, Total		
Number of Tweets	2,097		
Unique Authors	1,012		
Unique Words	10,010		
Unique Hashtags	511		
Panel B. Labeled Subsample from Step 6			
Measure	Count, Left-Leaning	Count, Right-Leaning	Count, Unlabeled
Number of Tweets	302	245	1,550
Unique Authors	158	119	736
Unique Words	2,538	2,076	7,918
Unique Hashtags	97	92	399

Notes: Unique words and hashtags can have duplicates across the subcategories "left-leaning", "right-leaning" or "unlabeled".

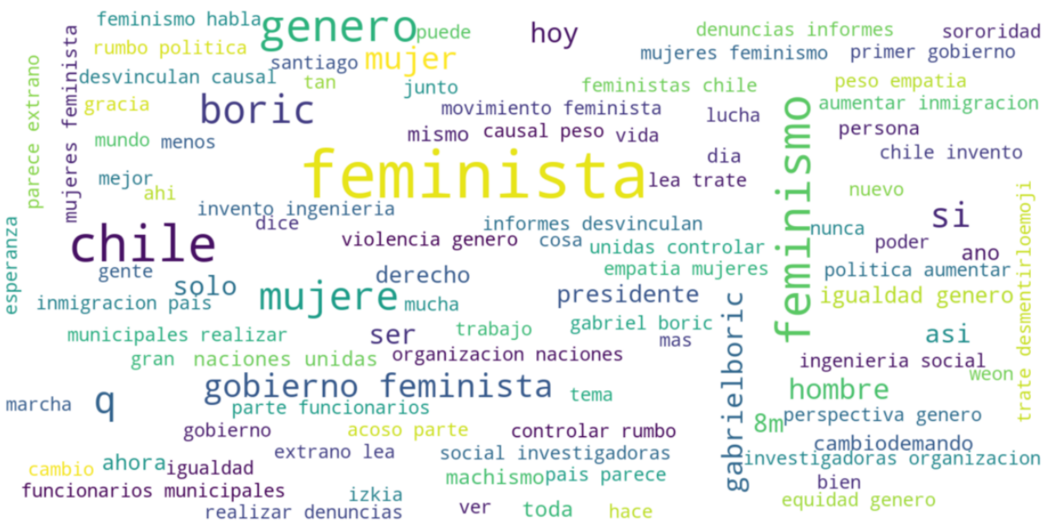
Data Source: Retrieved from Twitter API, spanning the time frame Mar 9th, 2022 – Mar 13th, 2022. Corpus includes tweets by Twitter users in Chile or Chilean nationals that mainly pertain to feminism.

Figure C.1: Tweets per Day, Feminism Corpus



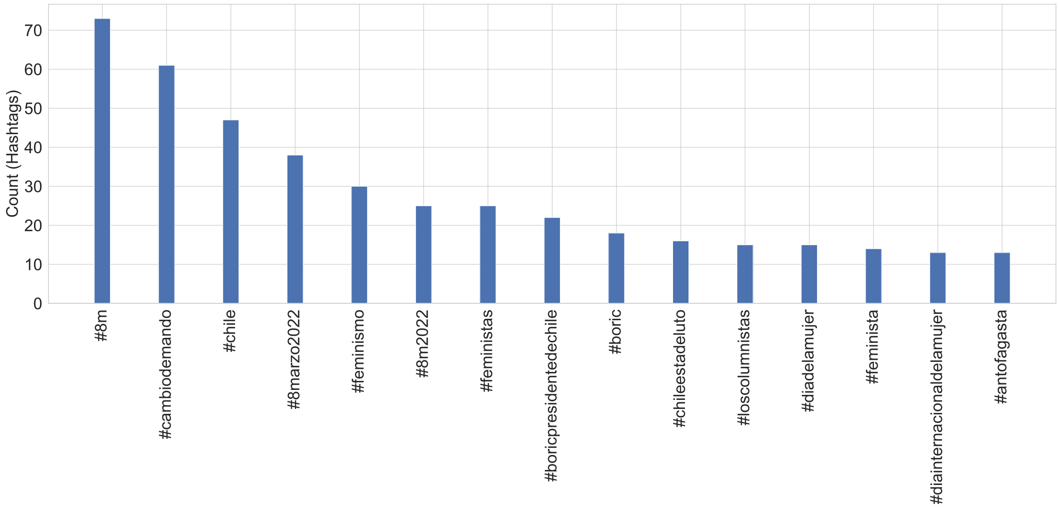
Data Source: Retrieved from Twitter API, spanning the time frame Mar 9th, 2022 – Mar 13th, 2022. Corpus includes tweets by Twitter users in Chile or Chilean nationals that mainly pertain to feminism.

Figure C.2: Word Cloud of Tweets, Feminism Corpus



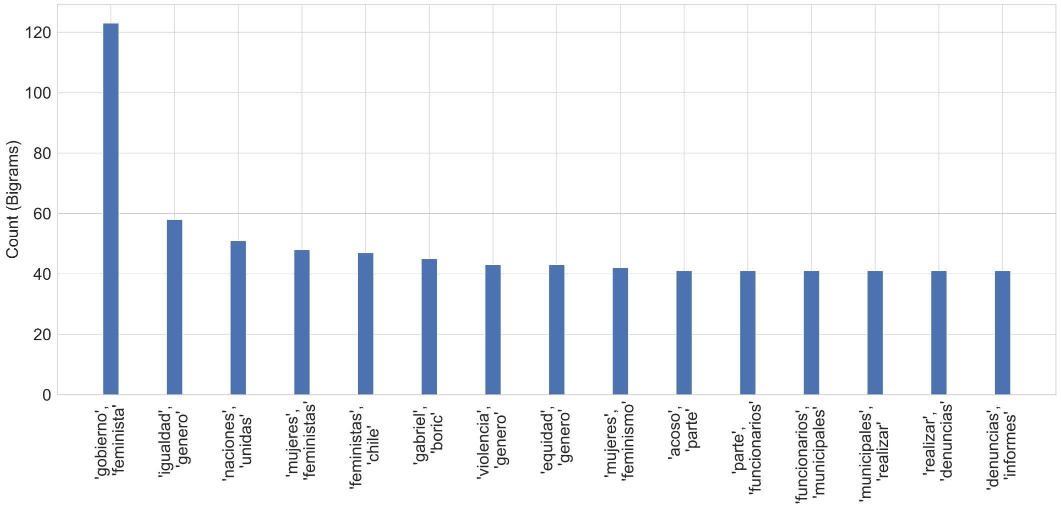
Notes: Textual preprocessing steps include lowercase enforcement and converting Spanish special characters (e.g. á, ñ, etc.) into corresponding non-accentuated ones.
Data Source: Retrieved from Twitter API, spanning the time frame Mar 9th, 2022 – Mar 13th, 2022. Corpus includes tweets by Twitter users in Chile or Chilean nationals that mainly pertain to feminism.

Figure C.3: Top Hashtags, Feminism Corpus



Notes: Textual preprocessing steps include lowercase enforcement and converting Spanish special characters (e.g. á, ñ, etc.) into corresponding non-accentuated ones.
Data Source: Retrieved from Twitter API, spanning the time frame Mar 9th, 2022 – Mar 13th, 2022. Corpus includes tweets by Twitter users in Chile or Chilean nationals that mainly pertain to feminism.

Figure C.4: Top Bigrams, Feminism Corpus



Notes: Textual preprocessing steps include lowercase enforcement and converting Spanish special characters (e.g. á, ñ, etc.) into corresponding non-accentuated ones.
Data Source: Retrieved from Twitter API, spanning the time frame Mar 9th, 2022 – Mar 13th, 2022. Corpus includes tweets by Twitter users in Chile or Chilean nationals that mainly pertain to feminism.