Understanding Opinions towards Migrants in Transit An Analysis of Tweets on Migrant Caravans in the US and Mexico

Tun-Mendicuti, Abigail; Kim, Jisu; Mulder, Clara H.

Published in:
Proceedings of the 16th ACM Web Science Conference, WebSci 2024

DOI:
10.1145/3614419.3644008

Publication date:
2024

Understanding Opinions towards Migrants in Transit

An Analysis of Tweets on Migrant Caravans in the US and Mexico

Abigail, Tun-Mendicuti
Population Research Centre, Faculty of Spatial Sciences, University of Groningen, Groningen, The Netherlands; Department of Digital and Computational Demography, Max Planck Institute for Demographic Research, Rostock, Germany
a.tun.mendicuti@rug.nl

Jisu, Kim
Department of Digital and Computational Demography, Max Planck Institute for Demographic Research, Rostock, Germany
kim@demogr.mpg.de

Clara H., Mulder
Population Research Centre, Faculty of Spatial Sciences, University of Groningen, Groningen, The Netherlands
c.h.mulder@rug.nl

ABSTRACT

The study of opinions towards migrants is profoundly important to understanding migration as well as to politics. Previous research has contributed to understanding anti-immigrant attitudes using social media data. However, there is still a need for a better understanding of opinions towards migrants in transit. We study the case of Central American migrant caravans from 2018 to 2021 by looking at the opinions in both the US, the destination country, and Mexico, the transit country. Media highly covered these events, and an online debate about them started on social media. Our research aims to understand how migrant caravans are discussed online. We are particularly interested in how media salience and geographical variables are associated with the sentiment intensity of the opinions. We combine geolocated data from Twitter, GDELT (Global Database of Events, Language, and Tone), and Survey and Census data for the US and Mexico. We use topic modeling to find the latent topics within the online Twitter discussion, and VADER sentiment analysis to quantify tweets' sentiments to calculate the sentiment intensity score that is used as the dependent variable of our OLS regression models. For both countries, we found that similar topics were discussed with a more political discussion in the US. Our analysis of the sentiment score revealed that sentiment does not reflect stance adequately, which led us to analyze the sentiment intensity score (absolute value of sentiment). We found that, for Mexico, when the media generated a higher number of news articles about migrant caravans, the sentiment intensity was higher. For the geographical variables, we found no significant association in the US; however, for Mexico, tweets in bordering states had a lower sentiment intensity. These results shed light on the differences in the determinants of sentiment intensity in opinions between the two countries.

Websci '24, May 21–24, 2024, Stuttgart, Germany
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0334-8/24/05
https://doi.org/10.1145/3614419.3644008

CCS CONCEPTS

• **Human-centered computing**; • **Collaborative and social computing**; • **Collaborative and social computing theory, concepts and paradigms**; • **Social media**;

KEYWORDS

Opinions, Migration, Online discussion, Twitter, Sentiment, Topics

ACM Reference Format:
Abigail, Tun-Mendicuti, Jisu, Kim, and Clara H., Mulder. 2024. Understanding Opinions towards Migrants in Transit: An Analysis of Tweets on Migrant Caravans in the US and Mexico. In *ACM Web Science Conference (Websci '24), May 21–24, 2024, Stuttgart, Germany*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3614419.3644008

1 INTRODUCTION

Migration emerges as a crucial and highly debated topic in the public and political spheres. For example, political stances and migration policies often influence electoral outcomes and party preferences [1] in many countries. Understanding and gauging opinions towards migrants and migration policy has become a significant factor in shaping political strategies [2, 3] and decisions regarding social cohesion, economic integration, and state sovereignty. As migration continues to play a prominent role in contemporary society, it is likely to remain a central point of discussion and contention in politics for the foreseeable future. In this paper, our specific interest is in opinions towards migrants in transit by looking at the particular case of opinions towards migrant caravans in the US and Mexico.

Migrant caravans as we know them today started in 2011 with the first "Migrant Viacrucis" (the Way of the Cross)[1] as a form of protest to denounce the violations of the Human Rights of migrants[2] [4]. This movement caught the media's attention in 2018 due to its large size, and its explicit aim to reach and to cross the US border, which in turn fueled the political discussion in the US and in Mexico.

[1] Their origin is from the catholic tradition of reliving Christ's suffering walk to the cross. The journey of these Viacrucis resembles the journey of suffering migrants have to go through in their journey through Mexico. In addition to these Viacrucis, other mass mobility protests are organized like the Caravan of Mothers of Missing Migrants, who are relatives of missing migrants. These caravans go through Mexico but with no intention of crossing to the US.

[2] During these processions, activists and undocumented migrants walk the routes that migrants usually transited during their journey through Mexico to the United States with no specific intention of crossing the border.

Migrant caravans emerged in a context of intense violence and danger for migrants in transit through Mexico and a tense political environment between the US and Mexico in terms of migration policy. In this context, migrant caravans were an alternative form of mass mobility considered by migrants as less expensive, more visible, and therefore safer [5]. These forms of mass mobility have been highly discussed online in the news and on social media, causing different reactions in public opinion.

The media depicted stories of migrants in transit in various ways, portraying the risks related to the journey itself, including exhaustion and transportation accidents. Since 2011, the conversation has focused on the escalation of violence against migrants, which included kidnapping, robbery, extortion, human trafficking, organized crime, and drug-related crime in their journey through Mexico [6]. Media coverage of these dangers not only raised the need for better migration policies but also promoted the creation of solidarity networks for undocumented migrants in transit, showing the relevance of media in the understanding of irregular transit migration.

This research aims to study the opinions that people express towards migrant caravans, using social media data. Particularly, we are interested in answering the following research questions for the US (as a destination country) and Mexico (as a transit country):

RQ1: How are migrant caravans discussed online?

RQ2: How do media salience and geographic variables influence opinions towards migrant caravans?

Previous research has focused on the study of opinions towards immigrants, building on inter-group contact theory [7, 8] and media effects theories [9, 10]. In our research, we build on these theories to study opinions towards migrants in transit. We contribute to the growing literature of opinions on migrants in transit using social media data by looking at opinions in the transit country and the destination country.

To address our research questions, we integrate data from several sources. These range from traditional sources, such as the census, to innovative sources, such as Twitter and the Global Database of Events, Language, and Tone (GDELT). Using these data sources, we study different aspects of opinions expressed in tweets. We first employ topic modeling to find the topics within the tweets. Secondly, we use sentiment analysis to calculate the sentiment intensity score of tweets. Lastly, we use OLS regression to understand the different associations of sentiment intensity to media salience and geographical variables

2 RELATED WORKS

Customarily, surveys are the data sources to study opinions and perceptions about migration. The surveys often include questions like whether there are too many foreigners in the country [11] or whether immigration is the most important issue to solve [12]. These questions help identify anti-immigrant attitudes. However, surveys are subject to time and space constraints since they are only conducted at specific moments, and not all countries have the same resources to do surveys [13]. Today, as society evolves in the digital era, new data sources have become available to study opinions, attitudes, and sentiments. Among these are social media data, which provide real-time observation of individuals' discussions online.

Social media data has been used for the study of different aspects of migration [14]. Some examples are migratory patterns [15], identification of migrants and analysis of their cultural integration [16], language acquisition of migrants [17], and attitudes and perceptions towards migrants. Some of the advantages of the data are the access to personal points of view about particular topics of interest, access to data in a chosen period of time for a particular event, and access to geolocations. These advantages of social media data opened up new avenues for researchers to gain valuable insights into public sentiments and opinions regarding migrants and refugees. By tapping into the vast sea of user-generated content on platforms like Twitter, we have access to a diverse range of opinions and experiences related to this important topic. Researchers have harnessed techniques such as topic modeling and sentiment analysis as powerful tools in this endeavor. These methods help enhance the understanding of opinions and emotions. For example, they have been employed to detect hate speech directed towards migrants [18] and to discern anti-immigrant sentiment, particularly during the challenging times of the COVID-19 pandemic [19]. However, there is still room to study opinions towards migrants in transit, especially in countries in the global south.

A recent effort to study transit migration with social media data was the analysis of the digital life of the Central-American migrant caravans [20]. This study looked into the interaction between different participants in the online conversation using social network analysis. Then, sentiment and content analysis of the tweets were performed for the different groups the authors identified through social network analysis. The public discourse around migrant caravans has also been studied by analyzing tweets from January to mid-February 2019, looking at the differences and similarities between countries of different migratory profiles [21]. The study found that the media widely promoted the public discourse and that 85% of the analyzed tweets had a neutral sentiment, with a higher number of retweets of positive tweets than of negative tweets.

In today's connected world, media is relevant to how people form their opinions about a specific topic. Considering the relevance of media in forming opinions, we build our research on the second-level agenda-setting theory [22]. This theory emphasizes two levels at which the media can influence people's opinions. The first level is the object of interest. It can be summarized as: "Although the media cannot tell us what to think; it can lead us into what to think" [23]. This means that how often a matter is discussed by the media shapes what people discuss. The second level is the object's attributes. At this level, attributes such as the tone or the framing of the matter in the media play a role in shaping people's opinions [24]. This theory has been used to study, for example, whether variation in news coverage affects anti-immigration attitudes [12] and whether positive or negative-toned TV news influences attitudes towards North African Americans in Belgium [9].

Media connects us and influences people's opinions. Our daily context, such as the people we interact with or encounter in our daily lives, also influences our opinions. Group threat and intergroup contact theories are two contrasting theories used to understand the role of interaction with immigrants in forming opinions towards them. Group threat theory [25] assumes that the local group would feel threatened by the outside group (immigrants). In contrast, inter-group contact theory [26] assumes that contact between majority and minority groups will alleviate tensions between the groups. Contact theory has been used to study anti-immigrant sentiment; for example, Mirwaldt [27] applied contact theory to border regions of Chequia with Germany, finding that cross-border interactions influence the opinions about neighboring countries.

Recent studies have combined inter-group contact theory and media effects theories to study opinions toward immigrants. Cyzmara and Stephan [10] looked into the effect of media issue salience on such opinions in Germany using longitudinal survey data. They found that greater visibility of immigration issues in the newspapers raises more individual concerns about immigration. They also found that the effect of the media salience is more substantial depending on whether individuals live in places with a lower or higher share of foreigners. A recent study uses these theories with social media data. Menshikova and Van Tubergen [28] proposed an approach to understanding the anti-immigrant sentiment by creating a panel data with Twitter data to test group threat theory and investigate the influence of media salience in anti-immigrant sentiments. They used multilevel models to identify effects at the regional, user and daily levels.

In the current study, we focus on opinions about transit migration, employing a case study of Migrant Caravans. Transit migration is a complex migratory dynamic that includes an origin country, a transit country, and a destination country. We look at opinions in the US, the final destination country, and Mexico, the transit country, providing a more comprehensive understanding of how migrants are perceived throughout their journey. We study different aspects of opinion: the topics discussed, the sentiment, and the sentiment intensity. With our study, we contribute to the growing literature that combines media effects theory with inter-group contact theory using social media data. Furthermore, to study the influence of media on opinions, we focus on an online setting with the use of GDELT as a data source for online news salience.

3 DATA

3.1 Twitter Data

Twitter is a micro-blogging platform where users can post short messages and interact with other users through replies, retweets, and mentions. We accessed these data through Twitter's application programming interface (API)[3] with an academic developer account that was available until March 2023. We collected geolocated tweets that were tweeted from the United States and Mexico from the 1st of January 2018 to the 31st of December 2021. We selected this time frame to include the Viacrucis events of March 2018 and to analyze the evolution of opinions regarding migrant caravans since their initial media prominence. The search query for the tweets was "migrant caravan" in English and Spanish and also in capital letters. We kept the query of search "migrant caravan" as a single term because if our query of search included either "caravan" or "migrant" separately, we obtained tweets that were not particularly related to the specific event of the migrant caravans.

The total number of tweets we obtained was 14,932, of which 54.34% were in English, and 37.75% were in Spanish. The rest were in other or undefined languages. Tweets of which the language was undefined mostly only contained URLs. We were not interested in such tweets, as these tweets do not contain explicit opinions of the user. For the analysis, we only kept those tweets written in either English or Spanish, relying on Twitter's language identification. After downloading the tweets, only tweets written by users with unverified accounts were kept for further analysis. In this way, we focus on the tweets of individual users rather than those from news and media accounts. Additionally, we considered retweets as individual tweets as they contribute to the discussion of migrant caravans as well. The final number of tweets was 6,740 for the US and 3,802 for Mexico.

The geolocations that Twitter identified did not always include a state of either the US or Mexico. To gather geolocation information at the state level, we therefore used the Twitter API[4] to acquire the longitude and latitude of a location and subsequently determined the corresponding state for this location.

3.2 GDELT data

In addition to Twitter data, we collected data from the Global Database of Events, Language and Tone (GDELT) project[5] to obtain information on the number of news articles and the average tone of online news[6] for every day. GDELT monitors broadcast, print, and online news from all over the world. GDELT data provide a large number of variables that are useful to predict social unrest events [29, 30]. One important variable is the average tone, expressed in degrees of positive and negative. The tone of the news articles has proven to be a valuable proxy for examining the country-specific ”media response” [31].

We use GDELT 2.0 through the GDELT 2.0 DOC API, from which it is possible to download data from the full archive and other general information like timeline, volume of news, and daily average tone of news in a given period and within specific countries. The GDELT 2.0 database started in 2015, and it updates every 15 minutes. Immediately after collecting a news article, it undergoes machine translation. The initial steps in this process involve language detection, word segmentation, morphological analysis, and sentiment analysis to determine the tone of the news. GDELT has developed emotional dictionaries for identifying the tone in the text. These emotional dictionaries are available in 15 languages, including English and Spanish. For articles in another language than English, the sentiment analysis is done after translation. GDELT measures the tone of a news article with these emotional dictionaries. In this process, words in the article are classified as having a positive, negative or neutral connotation. The positive/negative score is the percentage of positive/negative words in the article, where each score varies from 0 to 100. The overall tone of the article is defined by the difference of scores: positive score minus negative score. The overall score can range from -100 (negative tone) to 100 (positive tone), with 0 being neutral. The tone scores of the news items on

[3] https://developer.twitter.com/en

[4] https://developer.twitter.com/en/docs/twitter-api/v1/geo/place-information/api-reference/get-geo-id-place_id

[5] https://www.gdeltproject.org/

[6] http://data.gdeltproject.org/documentation/ISA.2013.GDELT.pdf

each day are then used to calculate the daily average tone of the news.

From this database, we downloaded data from the 1st of January 2018 to the 31st of December 2021 to obtain the daily volume of the news articles and the average tone of the daily news articles that contain the query "migrant caravan". We restricted our search to news articles coming from news outlets in Mexico and the United States, relying on GDELT's algorithm[7] to identify the country source of each news outlet. Their algorithm relies on the journalistic geographic bias that news outlets cover more events that occur in close spatial proximity to their location than in other areas of the world. It uses the top country mentioned in the news outlet to determine the country of origin of the news outlet and, consequently, the news article. We divided the daily tone by 100 to scale it from -1 to 1.

3.3 Survey and Census data

We utilized data from the American Community Survey of 2018, 2019, 2020, and 2021 accessed through the Integrated Public Use Microdata Series (IPUMS) USA[8] to obtain the number of Guatemalan, Honduran and Salvadoran immigrants in the US. In the case of Mexico, no yearly survey data is available to obtain these estimates. We therefore approximated the count of immigrants from Guatemala, Honduras, and El Salvador by determining the population born in any of these nations. For the years 2018 and 2019, we used data from the 2015 Intercensal Survey[9]. For the years 2020 and 2021, we relied on data from the 2020 sample of the National Census[10].

4 METHODS

4.1 Topic Modeling

Topic Modelling is an unsupervised machine learning technique that counts words and looks for patterns to find the underlying topics within a document. For this research and analysis, each tweet is a document. There are several models for doing this analysis. This research uses Latent Dirichlet Allocation (LDA), a standard model for topic modeling that has been used before for the classification of tweets [18, 19]. It is a well-established method that has been widely used in the literature [32, 33]. The model assumes that the topics are hidden in the documents; that is, they are latent. Another critical assumption is that the distribution of topics in a document and words in topics each follow a Dirichlet distribution. The number of topics must be previously selected. Initially, LDA assigns a certain probability for the words to belong to a topic and for a topic to belong to a document. Afterwards, through a series of iterations, the best assignment of words to topics and topics to documents is found. With this procedure, the probability for each word to belong to a topic and the likelihood for a topic to belong to a tweet are estimated. We characterized each topic by looking at the ten words with the highest probability of belonging to a topic. To determine the general theme of the topic, we looked at the most common words in a topic. Additionally, we read random tweets that were most likely associated with a topic to better understand what the topics were about.

By employing the metric proposed by Cao et al. [34], we determined the number of topics for the tweets from the US and Mexico, as further explained in Appendix A. Through this process, we selected four as the number of topics for both countries.

4.2 Sentiment Analysis

Sentiment analysis is a natural language processing technique to identify sentiments in text. It has been used to understand public opinion about specific topics, including anti-immigrant sentiments [19, 28]. The two main approaches to sentiment analysis are machine learning models (e.g., Naive Bayes [35], Support Vector Machine [36], Maximum Entropy [37]) and lexicon and rule-based tools based on emotional dictionaries of words (e.g., Linguistic Inquiry and Word Count [38] and SentiWordNet [39]). This research uses the second approach by using VADER (Valence Aware Dictionary and sEntiment Reasoner) [40], which is tailored to analyze text in a micro blog-like context. VADER considers grammar and syntactic rules used by humans when using social media to express their feelings or opinions and classifies sentiment into positive, negative, and neutral. The significant advantages of VADER are that it gathers lexical features of established sentiment lexicons, it is computationally economical, and it is transparent about the rules and lexicon it uses. Furthermore, VADER was initially tested using Twitter data. Other advantages are that it can handle typical negations, emoticons, emojis, commonly used acronyms, and the use of punctuation and uppercase to signal increased sentiment and emphasis. The analysis returns a compound score as a measure of the overall sentiment of the tweet. This compound score is computed with the valence scores of each word in the tweet and is normalized to be a measure between -1 (most negative) and 1 (most positive). An advantage of having a continuous score is that it allows a measurement of how positive or negative the feeling expressed in the opinion is. The standard thresholds for classification are: positive if the score is higher than 0.05, negative if the score is lower than -0.05, and neutral if it is between -0.05 and 0.05 [40].

4.3 OLS regression analysis

We employ Ordinary Least Squares (OLS) regression to explore how various factors, such as media salience, the size of the immigrant population, and the geographical origin of tweets, contribute to the sentiment intensity score of individual tweets. The sentiment intensity score is defined as the absolute value of the VADER sentiment score assigned to each tweet. The reason for taking the absolute value of the sentiment score is explained in Section 5.2.2. Our model takes the following form:

$$\begin{gathered} \textit{sentiment intensity score} = \textit{media salience variables} \\ +\textit{geographical variables} + \textit{control variables} \\ +\textit{constant} + \textit{error} \end{gathered}$$

The media salience variables were derived from GDELT. They include the volume of news articles about "migrant caravans" and the absolute value of the average daily tone of news articles on national online news, which we call the daily tone intensity. The volume of news items is the proportion of news items that contain

[7] https://blog.gdeltproject.org/mapping-the-media-a-geographic-lookup-of-gdelts-sources-2015-2021/

[8] https://usa.ipums.org/usa/

[9] https://www.inegi.org.mx/programas/intercensal/2015/

[10] https://www.inegi.org.mx/programas/ccpv/2020/

the query “migrant caravans” from the total number of news items of that day worldwide. We scaled this variable by a factor of 100 to obtain readable results (the volume of news was always less than 0.0004). Our geographical variables were derived from the census and the geolocation of tweets. We include the percentage of Northern Central American immigrants in each US and Mexican state and a categorical indicator of whether the state belongs to a bordering region. The percentage of Northern Central American immigrants takes into account the proportion of individuals born in El Salvador, Honduras, and Guatemala living in each state in the United States and Mexico. For the US, we use a dummy variable for the states in the US that share a border with Mexico. For Mexico, the indicator for border states has three categories: the state borders with the US, with Guatemala or Belize, or with none of them.

We control for the year when the tweet was written to account for temporal factors such as the occurrence of significant events or changes in public sentiment. For the US, we additionally control for the language of the tweet. Language is particularly relevant in the US, because of the sizable Latin immigrant population. Finally, taking into account that tweets are clustered within states and that the percentage of immigrants is a variable that varies per state and year, we estimated our coefficients with clustered standard errors for state and year.

5 RESULTS

5.1 Topics of the online discussion

The results of the topic modeling (Table 1) show the relevance of the borders, support for migrant caravans, and the government response towards them for both the US and Mexico. For the US, the first topic includes keywords that show intention (will, can, and want) and action (get and can). The second topic's keywords are associated with the situation at the border, such as asylum, children, and Tijuana. The topic includes opinions about the aggression occurring at the border, the asylum situation, and the matter of children migrating by themselves or with parents. The third topic contains words related to the right to transit (come, stop, immigration, and troop), which are associated with the measurements governments took to contain migrant caravans. The topic includes opinions about either welcoming or stopping immigrants with the sending of troops to the border. The fourth topic's keywords include political figures like former president Trump, democrats, and the news media, as well as concepts related to information (report, claim, and support) and transactions (use and fund). This topic includes opinions about the political figures' statements toward migrant caravans and whether they financially benefited from the migrant caravans.

For Mexico, the first topic contains keywords like border, cross, and Chiapas (a Mexican state on the south border) associated with the situation at the border. This topic includes opinions about the two borders of Mexico: north and south. The second topic's keywords include words associated with governance (govern, country, and nation) and words associated with the right to migrate (right, enter, human, pass). This topic includes opinions about the response of the Mexican government since the tweets also expressed opinions on whether free migration is a human right or not. The third topic contains keywords associated with assistance to migrant caravans in border cities (Tijuana, shelter, arrive, and support). This topic includes opinions about the consequences of the arrival of migrants to cities. The fourth topic's keywords include words related to intention (can, want, help), communication (say and know), and former president Trump. This topic includes opinions about the intentions of migrants and Trump's declarations.

Table 1: Topics and the 10 most common words of tweets with geolocation in the US and Mexico

US		Mexico	
Topic	Keywords	Topic	Keywords
1. Intentions and actions of migrants	people, will, american, countri, get, america, want, can, central, make	1.Matters of the borders	border, mexican, central, american, state, honduran, cross, unit, member, chiapa
2. Situations at the border	border, mexico, asylum, tijuana, mexican, children, member, group, hondura, head	2. Government and humanitarian response	mexico, countri, govern, right, enter, human, migrat, ask, nation, pass
3. Welcome or stop migrant caravans	like, immigr, just, say, stop, now, come, one, troop, state	3. Arrival of migrants	will, tijuana, arriv, support, today, hondura, alread, shelter, leav, citi
4. US government actions and politics	trump, news, presid, use, new, report, fund, claim, support, democrat	4. Reactions towards US actions	peopl, want, help, one, say, trump, like, know, can, mani

5.2 Sentiment of the online discussion

5.2.1 Sentiment of tweets. Figure 1 shows the daily average sentiment in the US and Mexico. In the second quarter of 2018, we observe more days with a negative average in the US and more days with a positive average in Mexico. In the last quarter of 2018, we observe similar patterns; however, we also observed more values close to zero. At other times, we observe fluctuation in the sentiment of tweets coming from both countries with no clear pattern. Sentiment goes up and down within a larger range of values for the US than for Mexico.

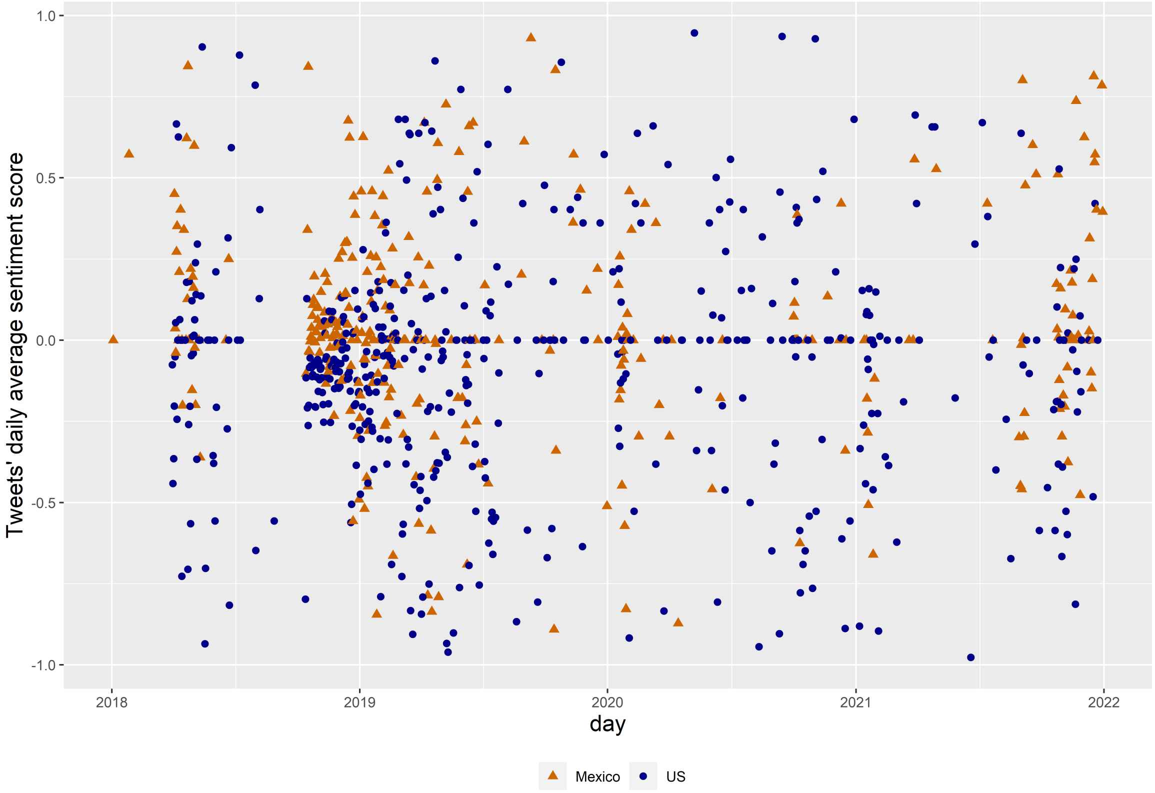


Figure 1: Daily average sentiment score for the US and Mexico

Figure 2 shows the distributions of the sentiment for the US and Mexico. We observe that for both countries, there is a large number of tweets with neutral sentiment. For the US, there are more negative than positive sentiment tweets. For tweets from Mexico, there seems to be a slightly higher concentration of tweets with positive sentiment than for tweets with negative sentiment. The mean value for the US tweets with a positive sentiment is 0.470, and for the negatives it is -0.533. For Mexico, the mean value of positive tweets is 0.486, and for negative tweets it is -0.490.

5.2.2 Sentiment vs Stance. Sentiment analysis is a way to look at opinions, but it may not fully capture an individual's stance on a topic. To better understand the differences between stance and sentiment in our sample, we took a 5% random sample to manually code the stance of the tweet in three categories: in favor of migrant caravans, against, or neutral. We found that there were tweets that we could not directly categorize as in favor, against, or neutral as they were not directly related to migrant caravans but rather to other political matters. Therefore, we classified them as unknown. An example of this category is:

> *"Are we seriously going to act like Obama wouldn't have just sent this migrant caravan back to where they came from?"*

Table 2 shows the comparison between our manual stance coding and the VADER sentiment results. For the US, we do not find a clear association between the sentiment classification and our stance coding. For Mexico, there are more observations in the diagonal cells than in other cells, but the association is far from perfect. Some

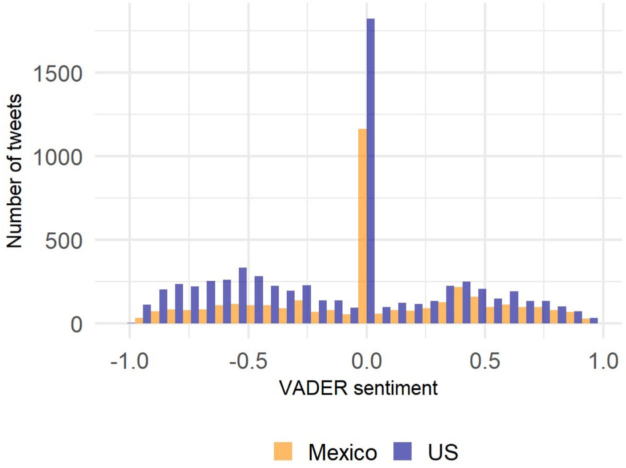


Figure 2: Distributions of sentiment score for tweets from the US and Mexico

Table 2: Cross-tabulation of sentiment and stance based on a 5% sample of tweets

	Sentiment					
	US			Mexico		
Opinion	Negative	Neutral	Positive	Negative	Neutral	Positive
Against	30	16	19	16	5	10
Neutral	45	53	38	14	34	14
In favor	57	10	36	20	5	47
Unknown	15	11	10	6	13	7

examples of tweets[11] that were classified by VADER as exhibiting a positive sentiment while we classified them as opposing migrant caravans are:

> *"Seems like Trump was on point (yet again). Those MS-13 gang folks are apparently making their way to the US in the caravan."* (Tweet from the US)
>
> *"#MigrantCaravan, don't come here. There are no jobs, and there are drug dealers and dangerous people. The United States doesn't even want you. It's better to look for opportunities in your own country. If you need help from other places, do it in an organized way, but stay in your country."* (Tweet from Mexico)

At the same time, there are also tweets classified by VADER as expressing negative sentiment but with a favorable opinion towards migrant caravans, for example:

> *"Let's cut the politics out of the #MigrantCaravan situation. We're losing our #humanity here. These brothers and sisters are just looking for #asylum from #violence."* (Tweet from the US)
>
> *"This is a serious crisis, they are FLEEING from Honduras. We need to be more humane and put ourselves in their shoes, or at least try to understand their reality a little. #CaravanaMigrante #CaravanaDeMigrantesHondureños"* (Tweet from Mexico)

The sentiment score indicates whether an opinion is expressed in a more positive or negative sentiment. However, from our analysis, we observed that supporting and opposing stances can be expressed with both positive and negative sentiments. This raises the challenge that we lack a direct interpretation of negative and positive sentiments as opinions towards migrant caravans. In order to address this issue, we adopt a different metric: the absolute value of the sentiment scores, which we denote as sentiment intensity. This measurement allows us to evaluate the emotional strength of opinions regardless of the specific stance of the tweet. We use this measurement as the dependent variable in our regression analysis (Section 5.3).

5.3 Sentiment intensity of the online discussion

Tables 3 and 4 show the results of the OLS regression of the sentiment intensity score (absolute values of VADER sentiment). Table 3 presents the results for the tweets written in the US. We observe that none of our variables of interest are statistically significant. However, we observe a negative relationship with sentiment intensity in the years 2020 and 2021. We also observe that tweets written in Spanish have lower sentiment intensity than those written in English.

Our results for Mexico are different, as shown in Table 4. The results for Mexico indicate that as the volume of news increases, sentiment intensity also tends to rise. When including the variable for border states (Model 1), we found that tweets from the Mexico-US border and the southern border had, on average, lower sentiment intensity compared to those from non-border states. When we introduced the variable measuring the percentage of Northern Central American immigrants (Model 2), we observed that, on average, tweets that come from states with a higher percentage of Northern Central American immigrants have a less intense sentiment. However, when we include both geographical variables (Model 3), we find that the results are not statistically significant, possibly due to confounding between these two variables.

6 DISCUSSION AND CONCLUSIONS

6.1 Discussion

In this research, we explored the online discourse surrounding migrant caravans. We examined the topics of discussion and the variation in sentiment intensity of Twitter users' opinions in relation to media salience and geographical variables. We focused on two countries that played distinct roles for migrant caravans: the US as the destination country and Mexico as a transit country. Through topic modeling, we found that for both the US and Mexico, discussions around migrant caravans include topics of the arrival conditions of migrants, their asylum-seeking intentions, the assistance and support given to them, the legality of border crossings, the deployment of the military to stop the migrant caravans, and the responses to the actions of both governments. We also found differences in topics between the US and Mexico, aligning with their functions as destination versus transit countries. For instance, in the US, the discussion exhibits a more political direction, emphasizing militarization and the right to enter the country. In Mexico, the discussion revolves around the journey of migrant caravans, the diverse challenges they encounter, the humanitarian response, and the situation in regions near the borders of Mexico.

To further understand how migrant caravans are discussed online, we analyzed the sentiment of the tweets. In the literature, sentiment scores have often been used as a proxy for opinions. Nevertheless, our analysis based on manual coding of the tweets revealed that sentiment analysis inadequately captures the specific opinions we aim to explain, such as supporting or opposing migrant caravans. This additional analysis also revealed that individuals associate migration issues with other political events, such as the US elections. When other political concerns are expressed and linked to migration, the emotional overlay could extend to how migration is perceived. This is particularly relevant when opposition to political matters disseminates into opposition to migration, since previous research has found that opinions with stronger sentiments are more likely to be shared [41, 42].

[11] All tweets presented in this article were rephrased using ChatGPT.

Table 3: Regression of the sentiment intensity score of tweets from the US

Dependent variable: sentiment intensity score			
	Model 1	Model 2	Model 3
Daily average tone intensity	-0.226	-0.212	-0.221
Daily volume of news items	0.574	0.602	0.586
Mex-US border (Ref: Non border)	-0.013		-0.017
% of NCA immigrants		-0.002	0.004
Year 2019 (Ref: Year2018)	-0.008	-0.008	-0.008
Year 2020	-0.050*	-0.049*	-0.050*
Year 2021	-0.054**	-0.052*	-0.054**
Tweet in Spanish	-0.040***	-0.042***	-0.041***
Constant	0.379***	0.375***	0.375***
Observations	6,740	6,740	6,740
Adjusted R2	0.003	0.003	0.003
F Statistic	3.830***	3.458***	3.419***

Note: Model 1 includes media salience variables, border indicator, and controls. Model 2 includes media salience variables, percentage of immigrants from Guatemala, Honduras and El Salvador, and controls. Model 3 includes media salience variables, all geographic variables, and controls.

*p<0.1; **p<0.05; ***p<0.01

Table 4: Regression of the sentiment intensity score of tweets from Mexico

Dependent variable: sentiment intensity score			
	Model 1	Model 2	Model 3
Daily average tone intensity	0.933	0.902	0.92
Daily volume of news items	1.884***	2.003***	1.916***
Mex-US border (Ref: Non border)	-0.018		-0.021
Southern border	-0.052**		-0.085
% of NCA immigrants		-0.067**	0.06
Year 2019 (Ref: Year 2018)	-0.006	-0.006	-0.005
Year 2020	-0.046	-0.038	-0.05
Year 2021	-0.008	-0.001	-0.011
Constant	0.294***	0.289***	0.292***
Observations	3,802	3,802	3,802
Adjusted R2	0.008	0.007	0.008
F Statistic	5.517***	5.660***	4.923***

Note: Model 1 includes media salience variables, border indicator, and controls. Model 2 includes media salience variables, percentage of immigrants from Guatemala, Honduras and El Salvador, and controls. Model 3 includes media salience variables, all geographic variables, and controls.

*p<0.1; **p<0.05; ***p<0.01

Furthermore, previous studies have shown that media salience has an influence on anti-immigrant opinions. Through our analysis, we found that the media effects differ between Mexico and the US, emphasizing the difference in how opinions are formed in a transit country and in a destination country. For instance, we found that in Mexico, the increase in media salience heightens the intensity of expressed opinions, which suggests a sense of crisis. For the US, we did not find an association between media salience and sentiment intensity. However, through our topic modeling, we found keywords associated with the media in one of our topics, showing that the media plays an important role in the discussion of the politics around migrant caravans. This accentuates the role of mass media in our current world, particularly in sensitive topics such as migration, since the continuous generation of news can affect how people express their opinions. When examining geographical variables, significant results were only observed in the Mexican context. The findings suggest that opinions in border states are less polarized compared to non-border states. In the case of the US, we only found a significantly lower intensity of the sentiment from 2020 and for tweets in Spanish.

6.2 Conclusion

Transit migration is a multifaceted phenomenon that engages various stakeholders, and often becomes entangled in debates surrounding legality and human rights. The complexities surrounding transit migration make it a challenging subject to understand fully. Despite the complexity, this study shed light on different aspects of the issue, specifically focusing on the sentiments expressed and their intensity. By delving into these topics, we offer insights that contribute to a better understanding of the factors influencing public opinions toward migrant caravans. This understanding, in turn, might contribute to the development of more informed and nuanced policies and discussions surrounding a specific flow of transit migration.

6.3 Limitations

Despite unveiling interesting discussions and sentiment intensity around the topic of the migrant caravan, our project suffers from a few limitations. One of the limitations is the bias present in our data. The Twitter data do not represent the whole population as it captures the perspectives of users active on the platform who allowed geolocation, excluding those without online presence and with different communication preferences [13]. Additionally, our data is limited to tweets that have geo-tags, and our media variables are restricted to the online articles within the collection of GDELT. Therefore, our results cannot be generalized for the whole population of the US or Mexico.

6.4 Ethical Statement

In this study, no individual-specific details have been released at any stage during the analysis. When possible, the results are presented at an aggregated level to protect privacy. We also ensured that all quoted tweets have been paraphrased while maintaining the original tone.

ACKNOWLEDGMENTS

We thank Emilio Zagheni, MPIDR's round table, the PRC struggle seminar, and Claudia Masferrer for their feedback and commentaries. We also thank the International Max Planck Research School for Population Health and Data Science (IMPRS-PHDS) for the resources provided.

REFERENCES

[1] C. Dustmann, K. Vasiljeva, and A. Piil Damm, "Refugee Migration and Electoral Outcomes," Rev. Econ. Stud., vol. 86, no. 5, pp. 2035–2091, Oct. 2019, doi: 10.1093/restud/rdy047.

[2] C. Boswell, A. Geddes, and P. Scholten, "The Role of Narratives in Migration Policy-Making: A Research Framework," Br. J. Polit. Int. Relat., vol. 13, no. 1, pp. 1–11, Feb. 2011, doi: 10.1111/j.1467-856X.2010.00435.x.

[3] N. Lauwers, J. Orbie, and S. Delputte, "The Politicization of the Migration–Development Nexus: Parliamentary Discourse on the European Union Trust Fund on Migration," JCMS J. Common Mark. Stud., vol. 59, no. 1, pp. 72–90, 2021, doi: 10.1111/jcms.13140.

[4] F. de J. Vargas Carrasco, "El vía crucis del migrante: demandas y membresía," Trace México DF, no. 73, pp. 117–133, 2018.

[5] A. Islas Colín, "Caravanas de migrantes y refugiados en México," Rev. Castell.-Manchega Cienc. Soc., 2019.

[6] G. Martínez, S. D. Cobo, and J. C. Narváez, "Trazando rutas de la migración de tránsito irregular o no documentada por México," Perfiles Latinoam., vol. 23, no. 45, pp. 127–155, 2015.

[7] L. M. McLaren, "Anti-Immigrant Prejudice in Europe: Contact, Threat Perception, and Preferences for the Exclusion of Migrants," Soc. Forces, vol. 81, no. 3, pp. 909–936, Mar. 2003, doi: 10.1353/SOF.2003.0038.

[8] S. L. Schneider, "Anti-Immigrant Attitudes in Europe: Outgroup Size and Perceived Ethnic Threat," Eur. Sociol. Rev., vol. 24, no. 1, pp. 53–67, Sep. 2007, doi: 10.1093/esr/jcm034.

[9] L. Jacobs and M. V. D. Linden, "Tone Matters: Effects of Exposure to Positive and Negative Tone of Television News Stories on Anti-Immigrant Attitudes and Carry-Over Effects to Uninvolved Immigrant Groups," Int. J. Public Opin. Res., vol. 30, no. 2, pp. 211–232, Sep. 2018, doi: 10.1093/IJPOR/EDW036.

[10] C. S. Czymara and D. Stephan, "Mass Media and Concerns about Immigration in Germany in the 21st Century: Individual-Level Evidence over 15 Years," Eur. Sociol. Rev., vol. 34, no. 4, pp. 381–401, 2018.

[11] L. Quillian, "Prejudice as a Response to Perceived Group Threat: Population Composition and Anti-Immigrant and Racial Prejudice in Europe," Am. Sociol. Rev., vol. 60, no. 4, pp. 586–611, 1995, doi: 10.2307/2096296.

[12] H. G. Boomgaarden and R. Vliegenthart, "How news content influences anti-immigration attitudes: Germany, 1993–2005," Eur. J. Polit. Res., vol. 48, no. 4, pp. 516–542, Sep. 2009, doi: 10.1111/J.1475-6765.2009.01831.X.

[13] J. Kim, L. Pollacci, G. Rossetti, A. Sîrbu, F. Giannotti, and D. Pedreschi, "Twitter Data for Migration Studies," in Data Science for Migration and Mobility, A. A. Salah, E. E. Korkmaz, and T. Bircan, Eds., British Academy, 2022, p. 0. doi: 10.5871/bacad/9780197267103.003.0008.

[14] A. Sîrbu *et al.*, "Human migration: the big data perspective," Int. J. Data Sci. Anal., vol. 11, no. 4, pp. 341–360, Sep. 2021, doi: 10.1007/s41060-020-00213-5.

[15] E. Zagheni, V. R. K. Garimella, I. Weber, and B. State, "Inferring international and internal migration patterns from Twitter data," in Proceedings of the 23rd International Conference on World Wide Web, in WWW '14 Companion. New York, NY, USA: Association for Computing Machinery, Apr. 2014, pp. 439–444. doi: 10.1145/2567948.2576930.

[16] J. Kim, A. Sîrbu, F. Giannotti, G. Rossetti, and H. Rapoport, "Origin and destination attachment: study of cultural integration on Twitter," EPJ Data Sci., vol. 11, no. 1, Art. no. 1, Dec. 2022, doi: 10.1140/epjds/s13688-022-00363-5.

[17] S. Gil-Clavel, A. Grow, and M. J. Bijlsma, "Migration Policies and Immigrants' Language Acquisition in EU-15: Evidence from Twitter," Popul. Dev. Rev., vol. 49, no. 3, pp. 469–497, 2023, doi: 10.1111/padr.12574.

[18] C. Arcila-Calderón, D. Blanco-Herrero, M. Frías-Vázquez, and F. Seoane-Pérez, "Refugees welcome? Online hate speech and sentiments in Twitter in Spain during the reception of the boat Aquarius," Sustainability, vol. 13, no. 5, p. 2728, 2021.

[19] F. Rowe, M. Mahony, E. Graells-Garrido, M. Rango, and N. Sievers, "Using Twitter to track immigration sentiment during early stages of the COVID-19 pandemic," Data Policy, vol. 3, 2021, doi: 10.1017/dap.2021.38.

[20] M. A. Walker and E. F. Boamah, "The digital life of the #migrantcaravan: Contextualizing Twitter as a spatial technology," Big Data Soc., vol. 7, no. 2, 2020, doi: 10.1177/2053951720978485.

[21] D. Toudert, "Migrant caravan crisis: Some realities about the public discourse on twitter," Migr. Int., vol. 12, 2021, doi: 10.33679/rmi.v1i1.2172.

[22] M. McCombs and S. I. Ghanem, "The convergence of agenda setting and framing," in Framing public life, Routledge, 2001, pp. 83–98.

[23] B. Cohen, "Press and foreign policy. Princeton university press," Princet. NJ, 1963.

[24] D. A. Scheufele and D. Tewksbury, "Framing, Agenda Setting, and Priming: The Evolution of Three Media Effects Models," J. Commun., vol. 56, no. 4, pp. 864–866, Sep. 2006, doi: 10.1111/j.1460-2466.2006.00326.x.

[25] H. Blumer, "Race prejudice as a sense of group position," Pac. Sociol. Rev., vol. 1, no. 1, pp. 3–7, 1958.

[26] G. W. Allport, K. Clark, and T. Pettigrew, "The nature of prejudice," 1954.

[27] K. Mirwaldt, "Contact, conflict and geography: What factors shape cross-border citizen relations?," Polit. Geogr., vol. 29, no. 8, pp. 434–443, Nov. 2010, doi: 10.1016/j.polgeo.2010.10.004.

[28] A. Menshikova and F. van Tubergen, "What Drives Anti-Immigrant Sentiments Online? A Novel Approach Using Twitter," Eur. Sociol. Rev., Sep. 2022, doi: 10.1093/esr/jcac006.

[29] D. Galla and J. Burke, "Predicting Social Unrest Using GDELT," in Machine Learning and Data Mining in Pattern Recognition, P. Perner, Ed., Cham: Springer International Publishing, 2018, pp. 103–116.

[30] F. Qiao, P. Li, X. Zhang, Z. Ding, J. Cheng, and H. Wang, "Predicting Social Unrest Events with Hidden Markov Models Using GDELT," Discrete Dyn. Nat. Soc., vol. 2017, p. 8180272, 2017, doi: 10.1155/2017/8180272.

[31] E. Boudemagh and I. Moise, "News Media Coverage of Refugees in 2016: A GDELT Case Study," Proc. Int. AAAI Conf. Web Soc. Media, vol. 11, no. 1, pp. 743–750, May 2017, doi: 10.1609/icwsm.v11i1.14981.

[32] E. S. Negara, D. Triadi, and R. Andryani, "Topic Modelling Twitter Data with Latent Dirichlet Allocation Method," in 2019 International Conference on Electrical Engineering and Computer Science (ICECOS), Batam Island, Indonesia: IEEE, Oct. 2019, pp. 386–390. doi: 10.1109/ICECOS47637.2019.8984523.

[33] S. Qomariyah, N. Iriawan, and K. Fithriasari, "Topic modeling Twitter data using Latent Dirichlet Allocation and Latent Semantic Analysis," AIP Conf. Proc., vol. 2194, no. 1, p. 020093, Dec. 2019, doi: 10.1063/1.5139825.

[34] J. Cao, T. Xia, J. Li, Y. Zhang, and S. Tang, "A density-based method for adaptive LDA model selection," Neurocomputing, vol. 72, no. 7, pp. 1775–1781, 2009, doi: https://doi.org/10.1016/j.neucom.2008.06.011.

[35] H. Parveen and S. Pandey, "Sentiment analysis on Twitter Data-set using Naive Bayes algorithm," in 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Bangalore, India: IEEE, 2016, pp. 416–419. doi: 10.1109/ICATCCT.2016.7912034.

[36] N. Zainuddin and A. Selamat, "Sentiment analysis using Support Vector Machine," in 2014 International Conference on Computer, Communications, and Control Technology (I4CT), Langkawi, Malaysia: IEEE, Sep. 2014, pp. 333–337. doi: 10.1109/I4CT.2014.6914200.

[37] X. Fei, H. Wang, and J. Zhu, "Sentiment word identification using the maximum entropy model," in Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering(NLPKE-2010), Beijing, China: IEEE, Aug. 2010, pp. 1–4. doi: 10.1109/NLPKE.2010.5587811.

[38] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: LIWC 2001," Mahway Lawrence Erlbaum Assoc., vol. 71, no. 2001, p. 2001, 2001.

[39] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining.," in Lrec, in 2010, vol. 10. 2010, pp. 2200–2204. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf

[40] C. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," Proc. Int. AAAI Conf. Web Soc. Media, vol. 8, no. 1, pp. 216–225, May 2014.

[41] P.-Y. Hsu, H.-T. Lei, S.-H. Huang, T. H. Liao, Y.-C. Lo, and C.-C. Lo, "Effects of sentiment on recommendations in social network," Electron. Mark., vol. 29, no. 2, pp. 253–262, Jun. 2019, doi: 10.1007/s12525-018-0314-5.

[42] S. Stieglitz and L. Dang-Xuan, "Emotions and Information Diffusion in Social Media—Sentiment of Microblogs and Sharing Behavior," J. Manag. Inf. Syst., vol. 29, no. 4, pp. 217–248, Apr. 2013, doi: 10.2753/MIS0742-1222290408.

[43] T. L. Griffiths and M. Steyvers, "Finding scientific topics," Proc. Natl. Acad. Sci., vol. 101, no. suppl_1, pp. 5228–5235, 2004, doi: 10.1073/pnas.0307752101.

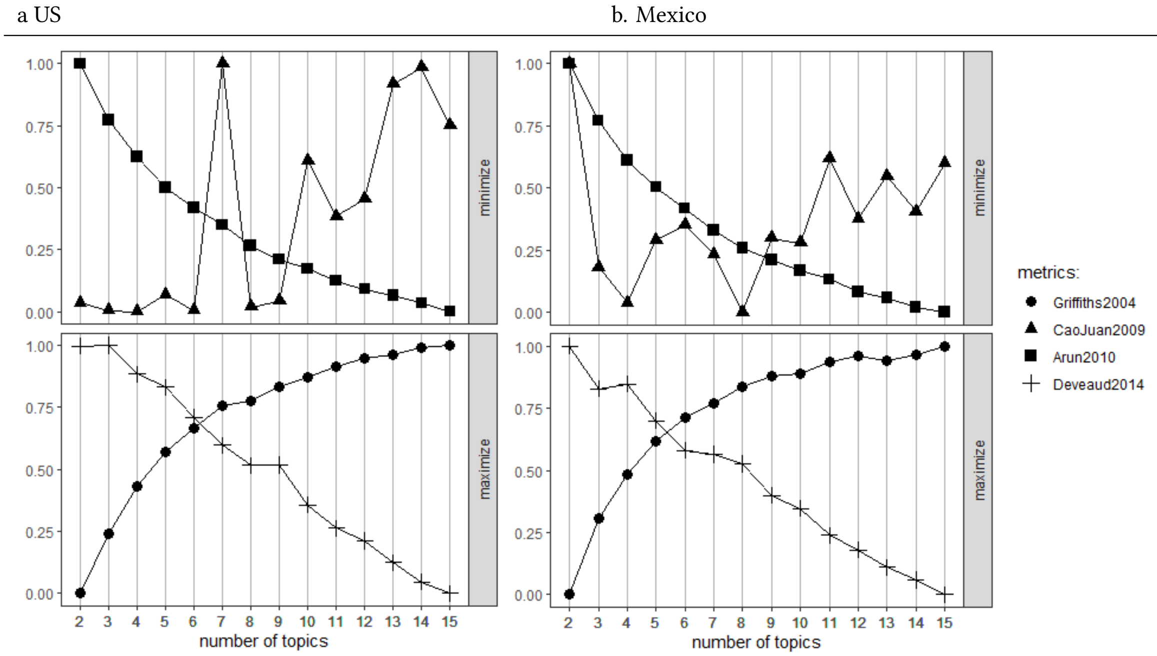
[44] R. Deveaud, E. SanJuan, and P. Bellot, "Accurate and effective latent concept modeling for ad hoc information retrieval," Doc. Numér., vol. 17, no. 1, pp. 61–84, 2014.

[45] R. Arun, V. Suresh, C. E. V. Madhavan, and N. Murthy, "On finding the natural number of topics with latent dirichlet allocation: Some observations," in Pacific-Asia conference on knowledge discovery and data mining, 2010, pp. 391–402.

APPENDIX

APPENDIX A. Choosing the number of Topics

This figure shows four different measures to find the optimal number of topics. All measures run from 0 to 1. The aim for Griffiths et al. (2004) [43] and Deveaud et al. (2014) [44] is maximization, whereas for Cao et al. (2009) [34] and Arun et al. (2010) [45] it is minimization. We used Cao et al. (2009) [34] to find the optimal number. For a range of 2 to 15 number of topics, a Latent Dirichlet Allocation model for each number of topics was calculated. Then, the metric Cao et al. (2009) [34] was calculated for the different fitted LDA models, and the best number of topics was found by taking the minimum value.


a US

b. Mexico