Multidimensional Analysis of Social Discourse on Twitter around Venezuelan Migration in Colombia

Joseph Martínez [1,*], Melissa Miller-Felton [1], Jose J. Padilla [1], Erika Frydenlund [1], Katherine Palacio [2]

[1] Virginia Modeling, Analysis, and Simulation Center, Old Dominion University, Suffolk, Virginia, USA
[2] Department of Industrial Engineering, Universidad del Norte, Barranquilla, Colombia
[*] Corresponding author
Email: jmart130@odu.edu

Introduction

The outflow of refugees and migrants from Venezuela is one of the largest displacement crisis in the world, with almost 7.7 million migrants and refugees as of November 2023 from which 6.5 million (84%) are currently in Latin America and the Caribbean, and 2.88 million (37%) in Colombia, making it the country that received the highest inflow [1]. Despite having a shared language, religion, and broad cultural heritage, this mass migration across Latin America has coincided with a decrease in migrant sentiment [2], and the formation of xenophobic, sexist, and discriminatory stances of the locals against the Venezuelan migrants [3]. These stances are influenced by ambivalent and even contradictory opinions and information from social networks, triggering xenophobic behaviors and actions [3].

Understanding the perception of the locals towards the migrants is vital because (1) it can help the policy development towards inclusion effectively through the public's views and concerns, linked to the creation or adjustment of immigration policies, social integration programs, and support services. (2) Analyzing sentiment helps identify areas where community relations might be strained and where interventions might be needed to promote integration of the migrants. (3) It can help address misinformation and stereotypes about migrants, thereby improving social cohesion. We want to study the conversation among the components of time, space, social interaction, and the usage of xenophobic terms.

We want to study it through time because we can know how much the conversation changed over time and whether it was driven by events, policies, or trends. The phenomenon has roots before 2015 but since that time the number of migrants increased dramatically. Events like COVID-19, The Presidential Elections in 2018, and governmental policies such as the ETPV can shape the conversation.
We want to study it through space because we can know whether the location of the people (or the characteristics of the people residing there) makes a difference when understanding the conversation.
We want the study the xenophobic terms because it has been widely used, but they were not originally used in that sense. We want to understand the sentiment and patterns of usage over time in terms of frequency and tone

Words or colloquialisms describing groups of people are not generally created by them but can often be reclaimed later as a method of empowerment. When a group is rendered physically or socially powerless—as in the case of Venezuelan migrants forced to leave their home country—the dominant culture describes the people and their predicament in words and phrases that capture out-group stereotypes. One scholar explains of the destructive nature of cultural expressions, "As a medium of communication, language expresses hidden notions of power, although, at a superficial level, the ideas and meanings contained in ordinary words are often assumed to be universally accepted by those who speak the language" [4]. In this paper, we explore how historically normalized words for Venezuelan migrants in Colombia, namely *veneco* and *veneca,* have transformed through social media to become markers of hate speech and xenophobic rhetoric on Twitter.

One striking and worrying aspect of the study is the prevalence of sexism among the populations, which translates into certain stereotypes about migrant women. Close to half of the people consulted in the three countries think that

migrant women will end up engaging in prostitution; at the same time, sexist roles are being reproduced, leaving women overburdened with care responsibilities, which increases the likelihood of their rights being infringed [3] *Veneco* slang for a Venezuelan man is used as an insult and *veneca* has out right become a synonym for prostitute *[5]*.

Background and Literature

Venezuelan Migration

Other studies have studied the perception of locals about Venezuelan migrants
The Barómetro de Xenophobia [6] analyzed the conversation with xenophobic content on Twitter. However, this was only done in 2021, and for the five cities Bogotá, Cali, Medellín, Barranquilla, and Cúcuta.

Use of Twitter and NLP for this analysis

However, LIMITATIONS OF OTHER STUDIES

- Sample size is small
- Not in Spanish
- Do not have a full picture
- Not that many tweets with geolocation activated

That is why in this paper we propose a multidimensional analysis to study this phenomena under different lenses.

- Sample size is bigger.
- Tweets in Spanish, and we trained the Spanish versions of the model
- Different dimensions give a unique perspective that compared with each other provides a better perspective
- Using the model, we could run analysis on tweets without geolocation activated.

Pejorative terms

Since the 1970s, Venezuela has been a safe haven for Colombians fleeing widespread civil conflict in hopes of work opportunities and a better life [7]. This amicable relationship and relatively porous border allowed Colombians to traverse between border towns with little immigration enforcement. During the 1970s when Venezuelans hosted millions of Colombians, the words *veneco* (masculine) and *veneca* (feminine) “weren’t always taken as an offense by Venezuelans.” Sergio Chacón, Master in Linguistics and Spanish, explains that this word arose from a kind of cultural and linguistic syncretism born within the border between both countries” [8]. The term, in its inception described Colombians that lived in Venezuela and developed a new accent, whereas now, it is a descriptor for Venezuelans migrating to Colombia [3].
Mid-2015 marked a reversal in migrant patterns. Venezuelans, and Colombians who had been in Venezuela for decades, began migrating to Colombia. “An estimated 4.5 million Venezuelans have fled their nation’s economic and humanitarian catastrophe in recent years, according to the U.N. About half of those are now residing in just two countries: Colombia and Peru”[4]. Corruption and a failed economy, often scapegoated as a botched socialism experiment, have led to an economic and political collapse that thrust Venezuelan citizens into the middle of a humanitarian crisis with little access to basic necessities to sustain life [9]. This has led to a major outflow of Venezuelans migrating into neighboring countries like Colombia (see Figure 1).

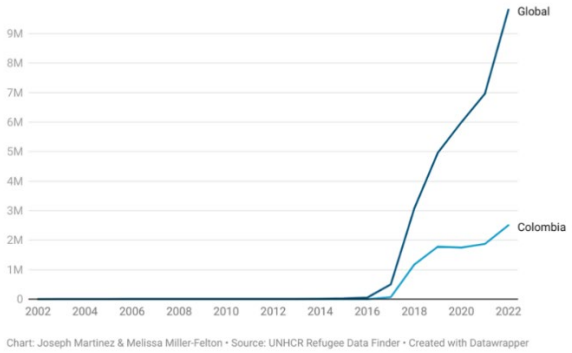


Figure 1. Migration of Venezuelans Over the Past 20 Years.

.

While it can seem like an exaggeration to refer to slang, like *veneco/a*, as discriminatory xenophobic rhetoric, it is appropriately classified as hate speech by meeting the following markers: (1) it provokes hatred and violence to large populations; (2) uses social media as medium; (3) is used broadly to describe groups; and (4) often used in rhetoric [10]. Scholars of linguistics, culture, and communication have started to take notice and offer an explanation for this shift. A sociologist at Universidad del Norte explains that when people are grouped with an adjective, even if it is a word derived from their nationality, it serves as a conduit of stigmatization and rejection. Now that there is a negative image about Venezuela, referring to Venezuelans even with a word like *veneco* derived as a shortened term for their nationality, will reproduce a negative association [8]. In this way, the word changes from a colloquialism without predetermined connotation to a pejorative.

Analyzing this 'othering' of Venezuelan migrants by Colombians serves to create a hierarchal power structure between two socio-cultural groups; othering by Colombians situates Venezuelan migrants as separate and below them in their societal order. This phenomenon of intergroup othering is not new or isolated to this particular case. Situating migrants as 'invaders' has been used to legitimize state-sanctioned violence in places like India, for decades where women who migrate from other regions are assumed to be involved in sex work, out of desperation for resources [11]. Joysheel Shrivastava [12] examines the prevalence of this 'othering' amongst women in India, where she describes women as the primary perpetuators in continuing a traditional violent language that is rooted in patriarchy.

Method

The proposed approach consists of seven major steps summarized in Figure 2. The first four steps encompass (1) the extraction of the tweets, (2) geolocation to identify whether a Tweet is from Colombia or not, (3) validation to only select the Tweets related to the Venezuelan migration and discard the ones about other topics, and (4) the identification of tone and topic. The following three steps contain the (5) temporal and (6) spatiotemporal analysis of the tone of the conversation, and (7) a temporal analysis of the pejorative terms *veneco* and *veneca.*

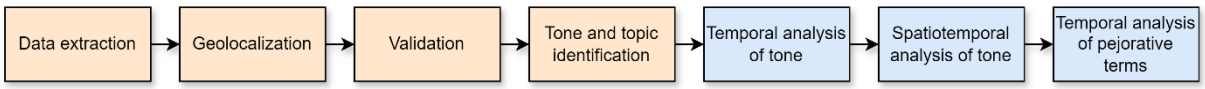


Figure 2. Methodology.

1. Data extraction

Using the Twitter Streaming API, we ran a search query containing keywords related to the Venezuelan migration accounting for their contracted versions and plurals. In addition, we included the variations by grammatical gender (feminine and masculine) since in Spanish the nouns have gender. We extracted the queries containing the keywords, but also the Tweets they referenced either as a retweet or a mention. In total, 7.4 million tweets were

extracted using the query in Figure 2, in the range of January 2014 and June 2022. We did not remove duplicates to keep the additional effect that one tweet can have when appearing multiple times.

```
veneco OR veneca OR venecos OR venecas OR ((venezolano OR venezolana OR venezolanos
OR venezolanas OR venezuela OR vzla OR vnzla) AND (migrante OR migrantes OR migracion))
```

Figure 3. Tweets extraction query.

2. Geolocalization

From the extracted tweets, in 2.35 million the geolocation was activated by the user, and in 5.1 million it was deactivated. Such geolocation, however, was mentioned as a raw text with no discrimination between different geographic levels like country or cities. For instance, "Bogotá" will not explicitly mention that is within the country of Colombia. Using the Nominatim API [13], we automatically converted the plain text to a country location. From the 2.35 million, the most frequent countries were Venezuela (47%), Colombia (15.1%), United States (6.4%) and Spain (4.5%). As we wanted to capture the whole conversation in Colombia, we decided to create a classification model whose purpose was to identify whether a tweet was from Colombia or not based on the tweet's text. With this, even when a Tweet had no geolocation activated, we could identify if it is from Colombia.
We fine-tuned three models: spaCy es_core_news_sm [14] and the transformer BETO [15], a BERT model trained on a sizable Spanish corpus (since all our tweets were in Spanish). Fine-tuning involves copying the weights from a pre-existing model and adjusting them for a new task [16]. To create the training and testing dataset we randomly selected the text and country from 135,000 tweets from Colombia and another 135,000 from any other country, resulting in a total dataset of 270,000 tweets. Using a 70/30 split we created the training (189,000) and testing (81,000) datasets, respectively. This ratio was used given that is commonly used for classification tasks with neural networks, [17] shown that it provides the best performance compared to other ratios in their task. The training results are presented in Table 1.

Table 1. Performance of the text geolocation models.

Architecture	Accuracy
es_core_news_sm	0.79
BETO	0.50

spaCy es_core_news_sm was selected to automatically label the entire dataset due to its highest accuracy score. After inference, from the 5.1 million Tweets with geolocation deactivated, 1.65 million where geolocated to Colombia based on its text.

3. Validity, tone, and topic identification

A classification model was trained to identify whether a tweet is relevant (about the Venezuelan migration) or not. We took a subset of 4,684 tweets and manually labeled them for relevance. We categorized each tweet to be about the Venezuelan migration or not (binary). The result was 2,991 relevant tweets and 1,693 irrelevant/unrelated. Upon the drop of duplicates from the relevant tweets and through random equal balancing, 1,554 tweets per category (True or False) were obtained. We split them with a 70/30 proportion for training (2,170) and testing (939) datasets. Three models were fine-tuned with the training dataset to classify whether a tweet is relevant: the two previously utilized models and the transformer BERT-base-multilingual-uncased-sentiment [18]. The performance accuracy metrics are presented in Table 1. The spaCy es_core_news_sm model was selected to label the entire dataset due to its highest accuracy score in both tasks (0.79 and 0.90, respectively).

Table 2. Performance of the Validation models.

Architecture	Accuracy
es_core_news_sm	0.90

BETO	0.81
BERT-base multilingual-uncased-sentiment	0.84

4. Tone and topic identification

Sentiment analysis determines whether data is positive, negative, or neutral.

The 2,991 tweets identified as relevant from the previous section were manually labeled to classify the subject of the tweet and tone of the terms *veneco* and *veneca*. Two coders labeled tweets separately, the degree of agreement between the two labelers was measured to be 60%, and one of the two labeled datasets was randomly selected for subsequent analysis. 60% is a fair value considering the difficulty of labeling sentiment expression as subjective interpretations can affect this agreement. The results were built with the labels of that selected dataset.
Since we are only interested in tweets that are related to migrants, we classified the subject of the tweet into one of the seven categories: *geopolitics, government, migrants, locals, media, migration*, and *other*. Filtered only for *migrants,* only 1,531 tweets were identified to be actually about them.
The tone was manually classified along a Likert scale [-3 (extremely negative), 3 (extremely positive)], where 0 represents neutral. The labeling criteria that our team established through discussion and consensus are shown in Figure 3.

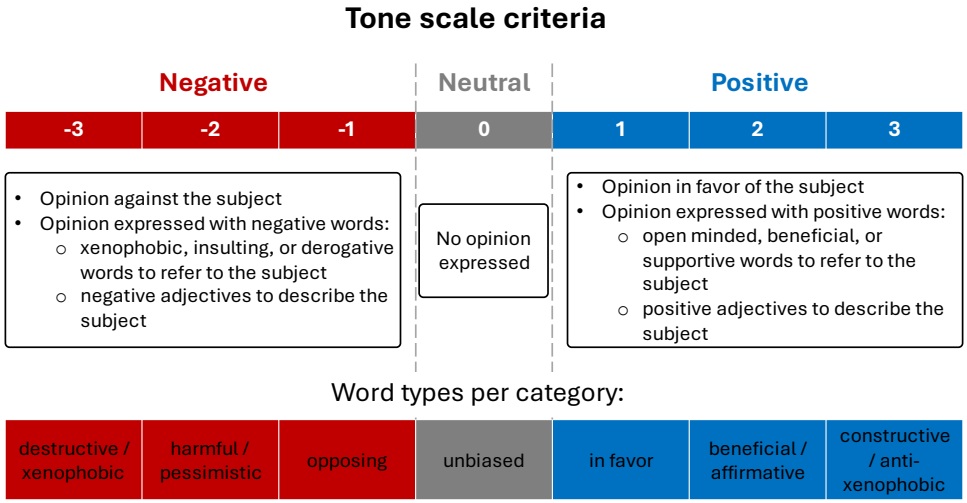


Figure 4. Criteria used for the tone scale labeling.

After the filtering, this table shows how many tweets were used for every model training and testing. We used 70% of them for training and 30% for testing.

Model	Training size	Testing size	Total size
Subject	824	356	1180
Tone	1344	576	1920

Results

1. Temporal analysis of tone

Figure 5 shows the proportion of Tweets with every type of tone where the average proportion was higher for negative tweets (44%), followed by neutral (35%) and positive (21%). From the total number of tweets, the proportion of retweets is 61%, and Figure 6 shows the proportion only for the unique tweets, discarding the retweets. The mean proportion of negative tweets increased to 55%, neutrals decreased to 25% and positives remained in 20%. The impact of retweets in the tone proportion can be noticeable given the portion of negative tweets and was driven by an increase in the portion of neutral tweets. This suggests that users more often retweet positive or neutral

content and that increase in the total number of Tweets counters the effect of negative Tweets. This is particularly true when we identify that 67% of retweets had a non-negative tone.
The frequency of Tweets with specific tones (Figure 7) shows us that while in mid-2015 and the beginning of 2017 experienced peaks in negative tone, the conversation was smaller in comparison with the period of 2018-2022. There, again the negative tweets were predominant in the majority of months with the exception of the period September 2018 (also the month with the highest number of tweets of the whole timeframe) to June 2019 where the negative was more frequent. However, when the retweets are discarded (Figure 8), the negative tone exhibits the highest frequency for all months and the negative and positive are noticeably similar. Additionally, in November 2019 the highest negative tone was achieved.

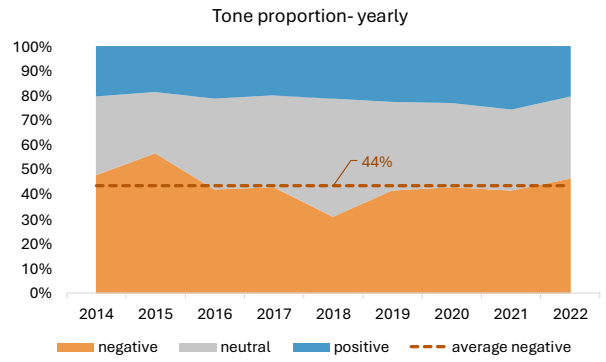


Figure 5. Tone yearly proportion.

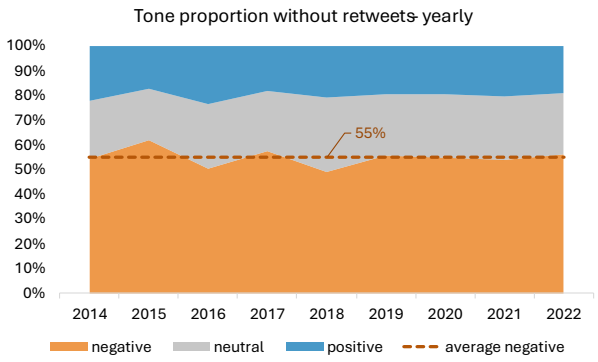


Figure 6. Tone yearly without retweets.

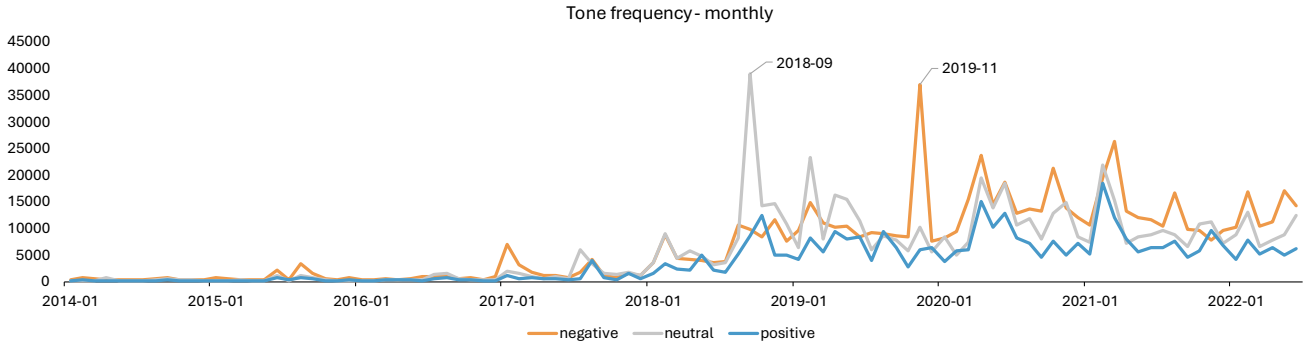


Figure 7. Tone monthly frequency.

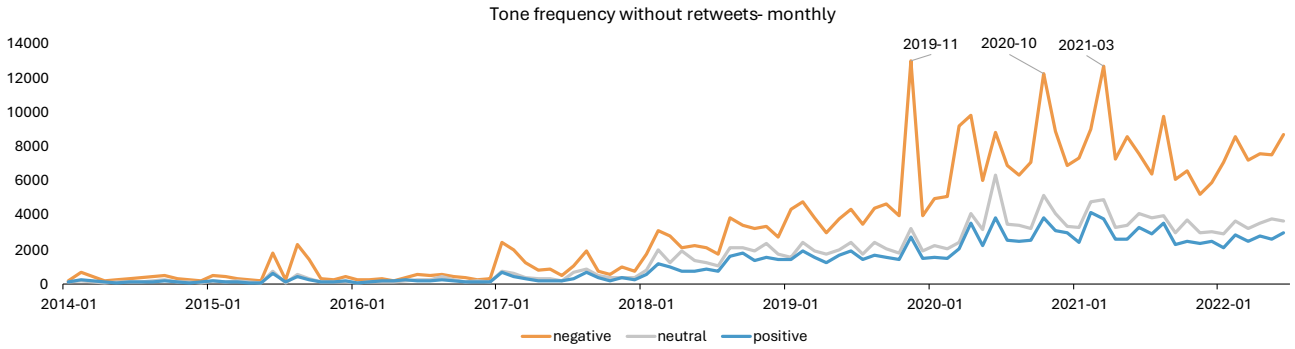


Figure 8. Tone monthly frequency without retweets.

For the tweets, frustration frequency was captured as the number of tweets with a negative tone in the specific timeframe. Figure 8 and Figure 9 show the frequency of negative tweets by every category. The Tweets about

migrants were predominant during all years, followed by government and geopolitics. Unusual increases in Tweets about the Colombian government happened in January 2019 and June 2021. The same during March 2020 and October 2020 for migrants. Comparing them with the hate speech frequencies they follow the same pattern, showing the

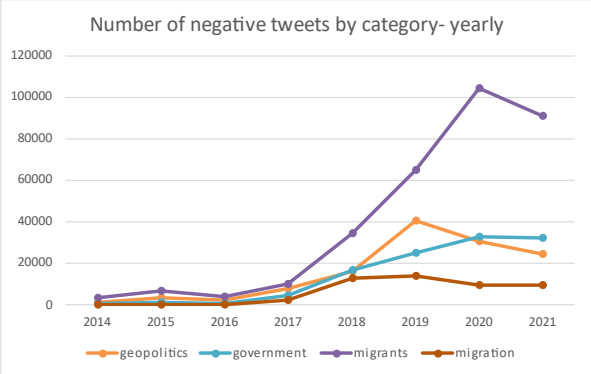


Figure 9. Yearly frequencies by frustration type

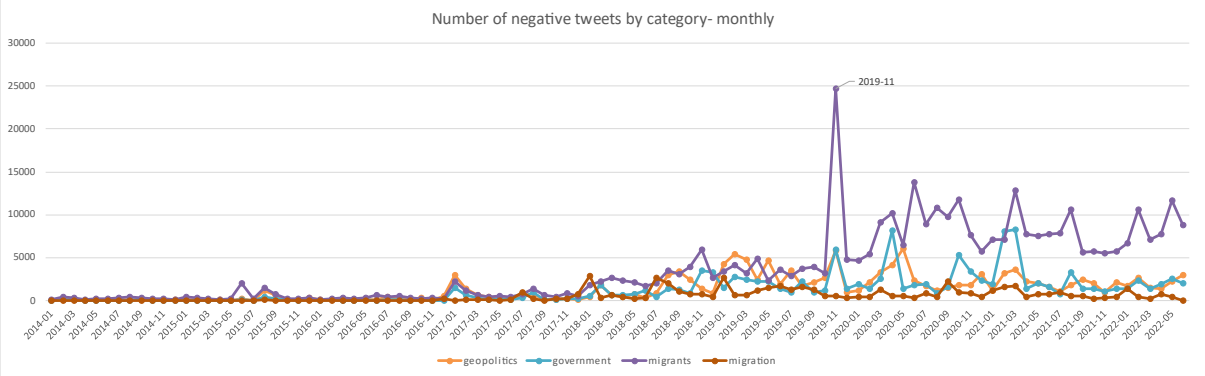
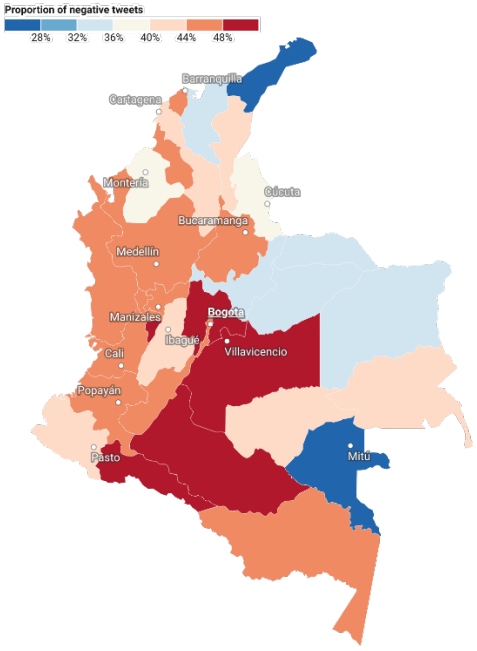
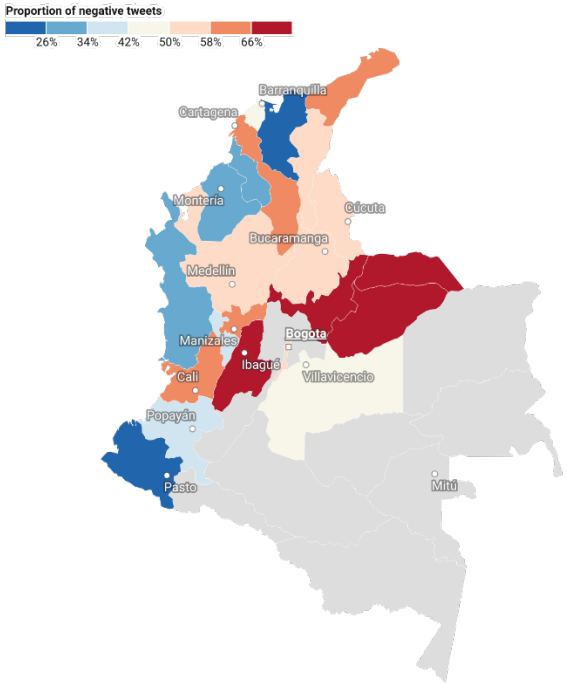


Figure 10. Monthly frequencies by frustration type

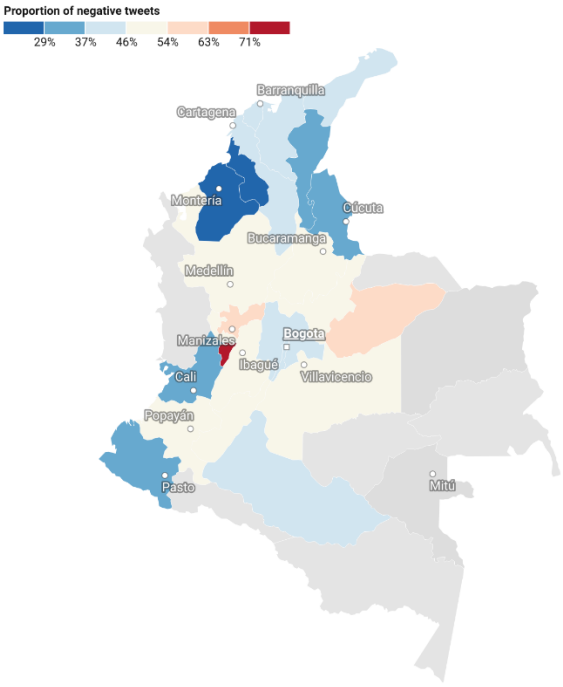
2. spatiotemporal analysis of tone



Proportion of negative tweets (2014)



Proportion of negative tweets (2016)



3. Temporal Analysis of Pejorative Terms

a. Terms usage estimation

The monthly and yearly distribution of tweets of the terms *veneco* and *veneca* are presented in Figure 9.

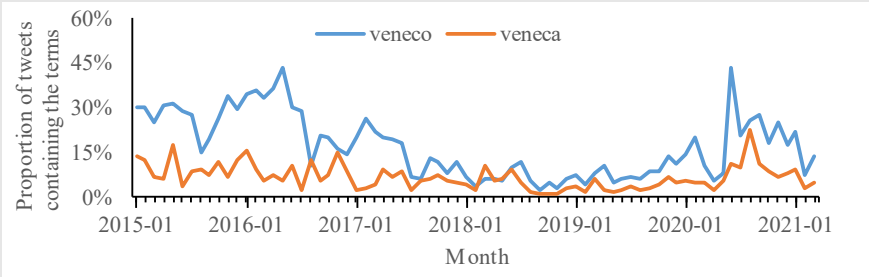


Figure 11. Monthly proportion of tweets containing the terms veneco and veneca.

For the majority of periods, the frequency of tweets containing *veneco* is higher than those containing *veneca*. This could be a result of masculine plurals including mixed male/female groups and is thus more encompassing of a term. Additionally, the proportion of tweets that contain either term is relatively higher in some years than others. Specifically, in 2015, the proportion was at its maximum for *veneco* (25%) and *veneca* (8%), followed by a decrease until 2018 to 5% and 2.5%, respectively. This also corresponds to a time where we observe an increase in the number of Venezuelans arriving in Colombia (see Figure 1), but the rate of arrival tapered off in later years. In 2021, the proportions rose again, possibly related to strained resources globally arising from Covid-19 and pandemic shutdowns. These strained resources and social tensions during Covid-19 could result in scapegoating of migrants, thereby leading to increased hate speech towards Venezuelan migrants at that time.
If we relate the proportion of the term's usage to the number of Venezuelan migrants in Colombia, it is rather unusual that the proportion during the period from 2015 to 2018 is significantly high. Given that a substantial increase in the number of Venezuelan migrants does not occur until 2018.

The monthly proportion of negative tone for both terms is presented in. In the given context, the tones refer to the emotional tone or sentiment expressed in the tweets containing the terms. The tones can be informative as they provide insight into how the public perceives the terms and the events related to them. For instance, a high proportion of negative tones in tweets containing these terms could suggest negative attitudes towards Venezuelans or the humanitarian crisis.

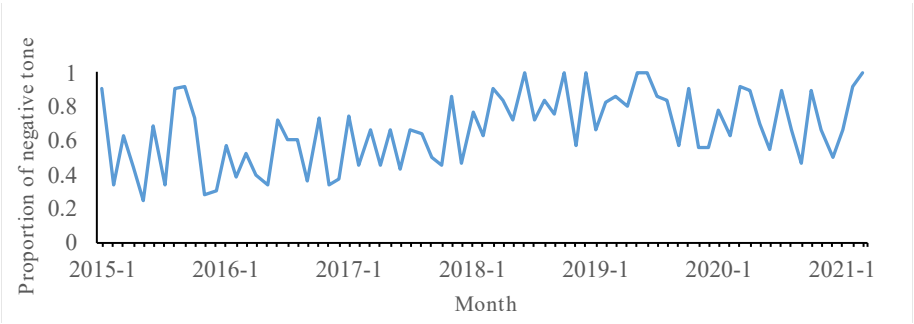


Figure 12. Monthly proportion of negative tone of the term veneco.

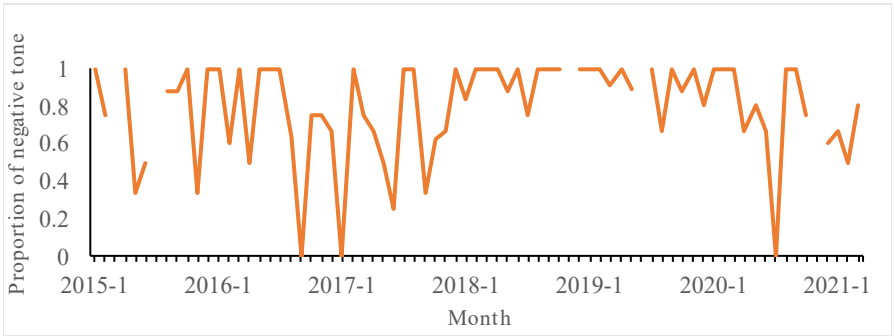


Figure 13. Monthly proportion of negative tone of the term veneca.

The two terms exhibit a predominantly negative tone, and it is noteworthy that the average negative tone associated with *veneca* is higher compared to that of *veneco*. Specifically, *veneca* has a 79% negative tone, whereas *veneco* reflects a 66% negative tone. In interviews we conducted with subject matter experts (linguists, sociologists, and communications scholars) who study derogatory language in Colombia in 2023, one possible explanation for this is that the term *veneca* has always had a negative connotation, well beyond the recent humanitarian migration of Venezuelans. We describe this phenomenon in more detail below, as the related themes are reflected in our dataset. The series of instances involving *veneco* initially displayed a balanced distribution between negative and non-negative sentiments until 2018, coinciding with the significant increase in the number of Venezuelan migrants (see Figure 1). This observation shows that despite the highest usage of *veneco* and *veneca* occurring between 2015 and 2018, the associated predominant tone was not necessarily negative. This finding also aligns with statements in the literature, which suggest that the term *veneco* was originally used as a descriptor but has evolved over time to become a pejorative term.

On the other hand, the series of instances involving *veneca* exhibits high variability due to the limited number of observations. This variability may affect the reliability of the negativity measurement

References

1. R4V Plataforma de Coordinación interagencial para Refugiados y Migrantes de Venezuela, *R4V América Latina y el Caribe, Refugiados y Migrantes Venezolanos en la Región - Nov. 2023*. 2023: https://www.r4v.info/es/document/r4v-america-latina-y-el-caribe-refugiados-y-migrantes-venezolanos-en-la-region-nov-2023.
2. Lebow, J., et al., *Migrant Exposure and Anti-Migrant Sentiment: The Case of the Venezuelan Exodus.* Available at SSRN 3660641, 2023.
3. Oxfam, *Yes, but Not Here: Perceptions of Xenophobia and Discrimination towards Venezuelan Migrants in Colombia, Ecuador and Peru*. 2019: https://www.oxfam.org/en/research/yes-not-here.
4. Wanitzek, U., *The Power of Language in the Discourse on Women's Rights: Some Examples from Tanzania.* Africa Today, 2002. **49**: p. 19 - 3.
5. Beach, C., *Frontera combustible: Conceptualising the state through the experiences of petrol smugglers in the Colombian/Venezuelan borderlands of Norte de Santander/Táchira.* Journal of Extreme Anthropology, 2018. **2**(2): p. 42-60.
6. Baldrich Luna, E.E., et al., *Informe final capstone-Barómetro de xenofobia hacia migrantes venezolanos.* 2022.
7. Bank, T.W., *Supporting Colombian Host Communities and Venezuelan Migrants During the COVID-19 Pandemic*, in *The World Band-Results Brief*. 2021.
8. Project, V.M., *Is 'Veneco' an insult or is it an inclusive word?*, in *Veneuzuela Migration Project-Education*. 2021.
9. Press, A., *Mounting Venezuela Exodus Sparks Fears of Rising Xenophobia*, VOA, Editor. 2019.
10. Cervone, C., M. Augoustinos, and A. Maass, *The Language of Derogation and Hate: Functions, Consequences, and Reappropriation.* Journal of Language and Social Psychology, 2021. **40**(1): p. 80-101.
11. Bilewicz, M. and W. Soral, *Hate Speech Epidemic. The Dynamic Effects of Derogatory Language on Intergroup Relations and Political Radicalization.* Political Psychology, 2020. **41**.
12. Shrivastava, J., *The Violence Of Language: A Feminist Take On The 'Culture' Of Abuses*, in *Feminism In India*. 2020.
13. osm-search, *Nominatim.* GitHub, 2023.
14. SpaCy, *es_core_news_sw*. 2023.
15. Cañete, J., et al., *Spanish Pre-Trained BERT Model and Evaluation Data*. 2020.
16. Houlsby, N., et al., *Parameter-Efficient Transfer Learning for NLP*, in *Proceedings of the 36th International Conference on Machine Learning*, C. Kamalika and S. Ruslan, Editors. 2019, PMLR: Proceedings of Machine Learning Research. p. 2790--2799.
17. Nguyen, Q.H., et al., *Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil.* Mathematical Problems in Engineering, 2021. **2021**: p. 4832864.
18. nlptown, *bert-base-multilingual-uncased-sentiment*. 2022.