BEHIND DEROGATORY TERMS FOR VENEZUELAN MIGRANTS IN COLOMBIA: XENOPHOBIA AND SEXISM IDENTIFICATION WITH TWITTER DATA AND NLP

Joseph Martínez1[0000-0002-2244-8274], Melissa Miller-Felton1[0009-0004-2943-4336], Jose J. Padilla1, Erika Frydenlund1[0000-0002-7694-7845], Katherine Palacio2[0000-0001-6134-7700]

1 Virginia Modeling, Analysis, and Simulation Center, Suffolk, Virginia, USA

2 Department of Industrial Engineering, Universidad del Norte, Barranquilla, Colombia  
jmart130@odu.edu

**Abstract.** In this paper, we examined the transformation of colloquial words used to describe Venezuelan migrants in Colombia, namely *veneco* and *veneca*, and how they have become markers of hate speech and xenophobic rhetoric on Twitter. Using computational models, we analyze these pejorative terms’ sentiment and usage patterns over time. We also explore the role of gender stereotypes in social media hate speech toward migrants, using a Critical Feminist lens to offer specific context about how the use of these terms vary by gender and sex. This perspective also shows how stereotypes are perpetuated about Venezuelan women by those outside of this group as a means to dehumanize and foster inferiority. The study reveals a shift in linguistic interpretations, as these once more commonly used terms have acquired negative connotations and are now frequently used in hate speech toward Venezuelan migrants. Through the analysis of 5.7 million tweets, we estimated the frequency of term usage and manually labeled a sample of 2,991 tweets to assess the negative tone associated with the terms. Additionally, qualitative analysis of a subset of tweets further examines the content and themes. The findings highlight the pervasiveness of othering and xenophobia towards Venezuelan migrants, driven by macro-level rhetoric and perpetuated on social media platforms at the micro level. This research contributes to a better understanding of the language dynamics and power structures surrounding migration and provides insights into the need for addressing xenophobic attitudes and promoting inclusive discourse.

**Keywords:** Migration, xenophobia, Natural Language Processing, social media.

1. Introduction

Language is the primary medium for communication and evolves with the people who speak it. 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” [1]. 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.

We focus in this study on those two pejorative terms in reference to Venezuelan migrants in Colombia to try to understand the sentiment and patterns of usage over time in terms of frequency and tone. We use big data methods and train computational models to reveal how certain gender stereotypes play out in social media hate speech toward migrants, which we explain using a Critical Feminist lens.

1. Background and Literature

Since the 1970s, Venezuela has been a safe haven for Colombians fleeing widespread civil conflict in hopes of work opportunities and a better life [2]. 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” [3]. 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 [4]. This has led to a major outflow of Venezuelans migrating into neighboring countries like Colombia (see Fig 1).

Chart, line chart

Description automatically generated

Fig 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 [5]. 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 [3]. In this way, the word changes from a colloquialism without predetermined connotation to a pejorative.

Analyzing this from a Critical Feminist lens, 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 [6]. Joysheel Shrivastava [7] 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.

Here, we focus on the evolution of othering terms (*veneco/a)* in social media, specifically Twitter. In the following sections, we explain our research methodology to extract tweets using the terms. The content, sentiment, and frequency distributions of these terms reinforce the notions of othering and justification for a cultivating culture of xenophobia, at micro and macro levels.

1. Methodology

The study is divided into three parts. First, 5.7 million tweets from January 2015 to March 2021 were extracted and filtered via Natural Language Processing (NLP) models to estimate the total frequency of the usage of the pejorative terms (*veneco*/*a*). Second, from the 5.7 million tweets, we manually labeled a sample of 2,991 tweets to analyze the proportion of the negative tone associated with the two terms. Third, we coded a stratified sample of 20% of the 2,991 tweets in a qualitative coding software to do a qualitative analysis of the tweets’ contents. This sampling method was used to ensure that the sample accurately represents the different subgroups.

* 1. Terms usage estimation with NLP

The Twitter API was used to extract tweets based on specific query parameters, such as Boolean conditions and tags. Multiple query variations were tested, but the number of extracted tweets did not significantly change. We extracted 5,710,988 tweets from January 2015 to March 2021 using at least one of the pejorative terms.

For geolocation, we fine-tuned a classification model whose purpose was to identify whether a tweet was from Colombia or not based on the tweet’s text. Fine-tuning involves copying the weights from a pre-existing model and adjusting them for a new task [8].We used the tweet author’s account location (if users share their location in their public profile) from a subset of four million tweets to build the model’s training and testing dataset. Using the Nominatim API [9], we automatically detected the author’s location from their bio text, which the API converted to a country location. 135,000 were from Colombia; we randomly selected another 135,000 from other countries to create the “not from Colombia” label resulting in a total dataset of 270,000 tweets. Using a 70/30 split we created the training (189,700) and testing (81,300) datasets, respectively. This ratio was used given that is commonly used for classification tasks with neural networks, [10] shown that it provides the best performance compared to other ratios in their task.

For the geolocation task, the two fine-tuned models were: spaCy es\_core\_news\_sm [11] and the transformer BETO [12], a BERT model trained on a sizable Spanish corpus (since all our tweets were in Spanish).

Another classification model was used 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 [13]. 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 1. Performance of the models

|  |  |  |
| --- | --- | --- |
| Architecture | Geolocation model accuracy | Relevance model accuracy |
| es\_core\_news\_sm | **0.79** | **0.90** |
| BETO | 0.50 | 0.81 |
| BERT-base multilingual-uncased-sentiment |  | 0.84 |

spaCy es\_core\_news\_sm was selected to automatically label the entire dataset due to its highest accuracy score in both tasks (0.79 and 0.90, respectively).

* 1. Terms’ tone analysis with manual coding

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 (Section 4) 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 in 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 **Error! Reference source not found.**. The same two labelers were involved in this labeling process.

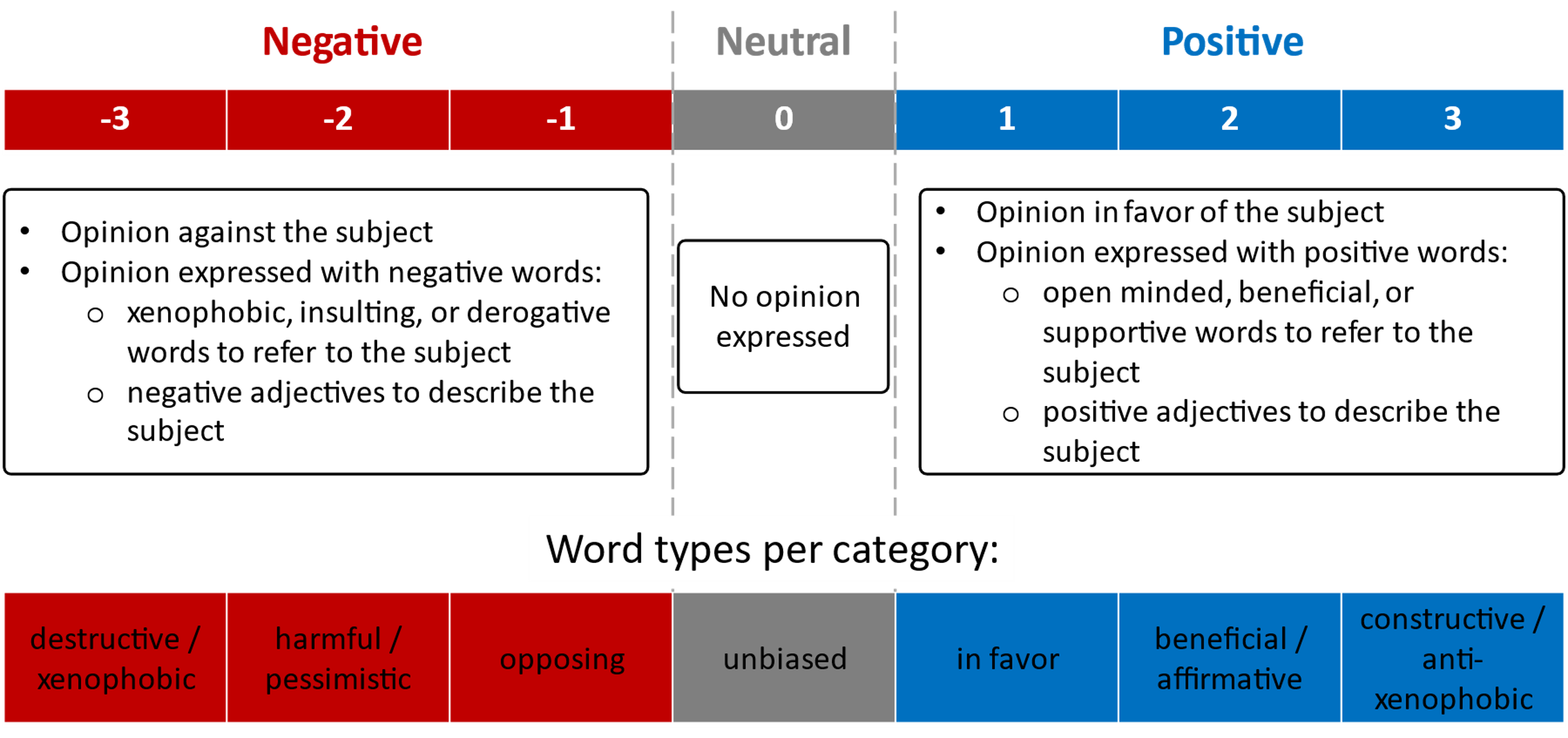


Fig 2. Tone scale labeling criteria.

* 1. Qualitative analysis of the terms

Of the 1,531 manually coded tweets referring to migrants, 663 contained the term *veneco* and 239 the word *veneca* and their plurals; 3 contained both*;* 629 contained neither term. It should be noted here that the plural of the masculine *venecos* can refer to a mixed male and female group, but the feminine plural is only female.

Since the focus of this paper is specifically on the use of the two terms, we excluded the 629 tweets that did not contain them from further analysis. We selected a 20% stratified sample of these 663 tweets along the tone (see Table 2). These 181 tweets were manually coded in qualitative software to analyze the tweets’ contents.

Table 2. Stratified sampling distribution.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Contains the term | Tone | | |  |
| Positive | Neutral | Negative | Total |
| *veneco/s* | 25 | 24 | 84 | 133 |
| *veneca/s* | 5 | 5 | 38 | 48 |
| Total | 30 | 29 | 122 | **181** |

Each tweet was associated with a numerical tone value ranging from -3 to 3. We grouped them as negative, neutral, and positive tweets. By doing this, we were able to identify what types of comments fall into each tone category. First, we organized the tweets by gender. Next, we filtered the tweets by the tone’s numerical value. Then, we identified major themes within each tone category. Finally, we identified theme patterns that crossed gender lines and ones that were exclusive to a particular gender.

1. Results
   1. Terms usage estimation with NLP

The monthly and yearly distribution of tweets of the terms *veneco* and *veneca* are presented in Fig 3.

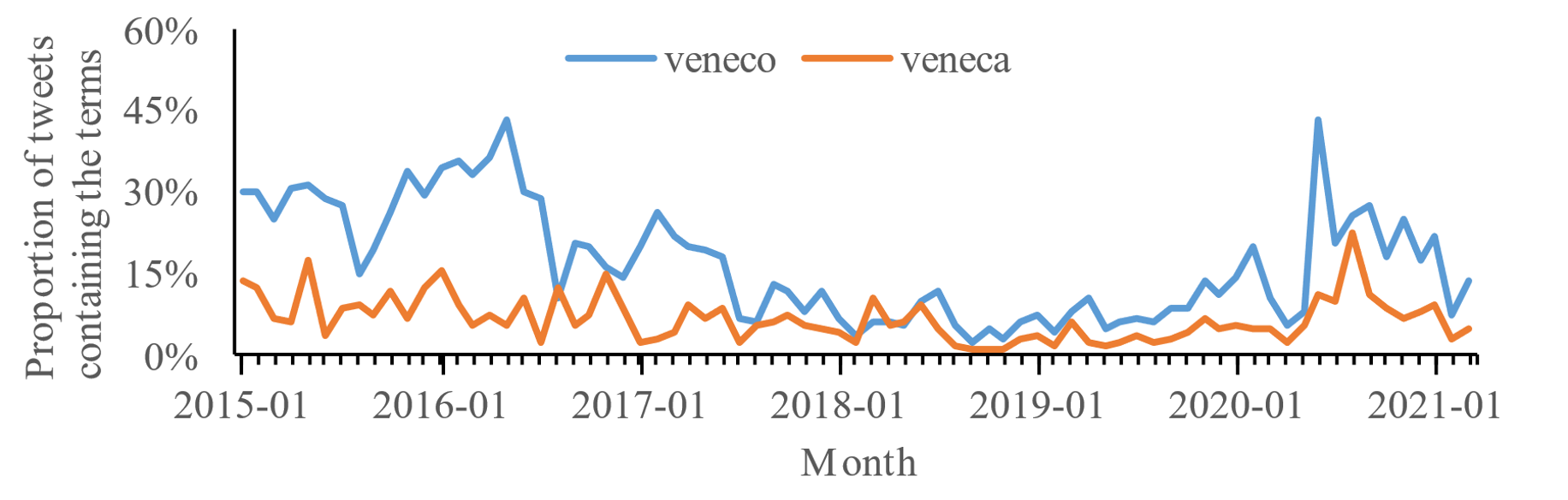


Fig 3. Monthly proportion of tweets containing the terms *veneco* and *veneca*.

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 Fig 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.

* 1. Terms’ tone analysis with manual coding

The monthly proportion of negative tone for both terms is presented in Fig 4 and **Error! Reference source not found.**. 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.

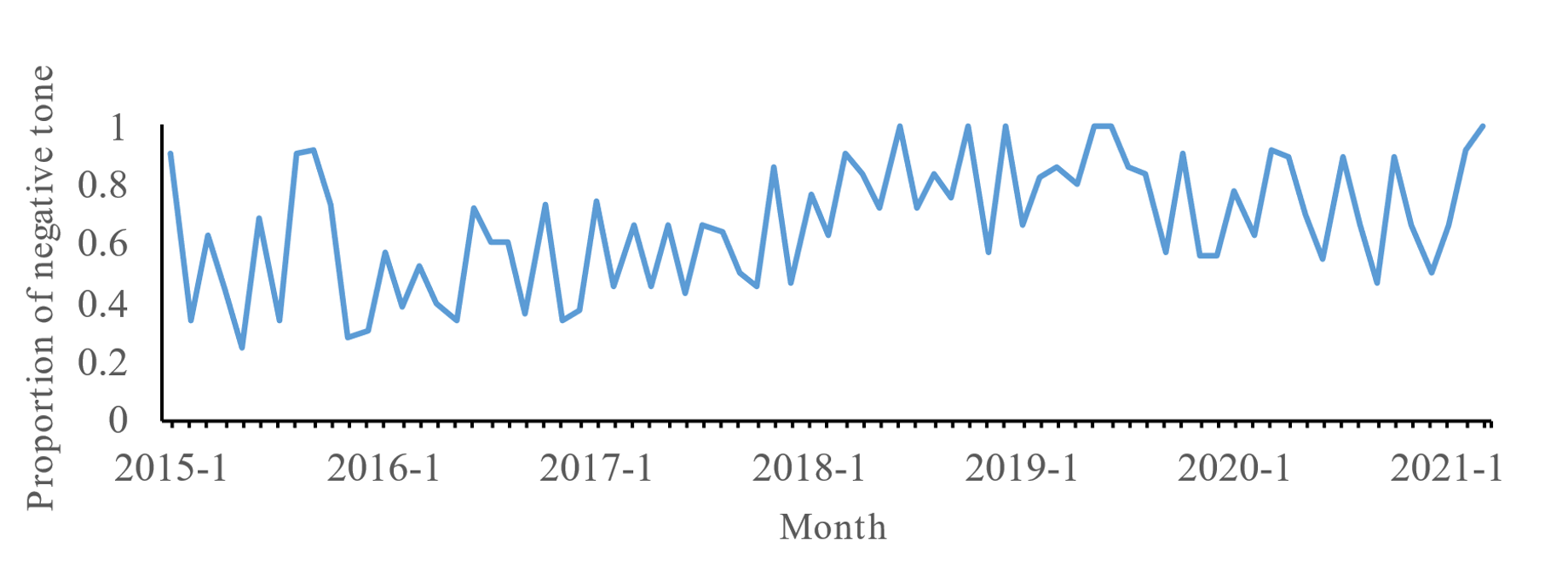


Fig 4. Monthly proportion of negative tone of the term *veneco*.

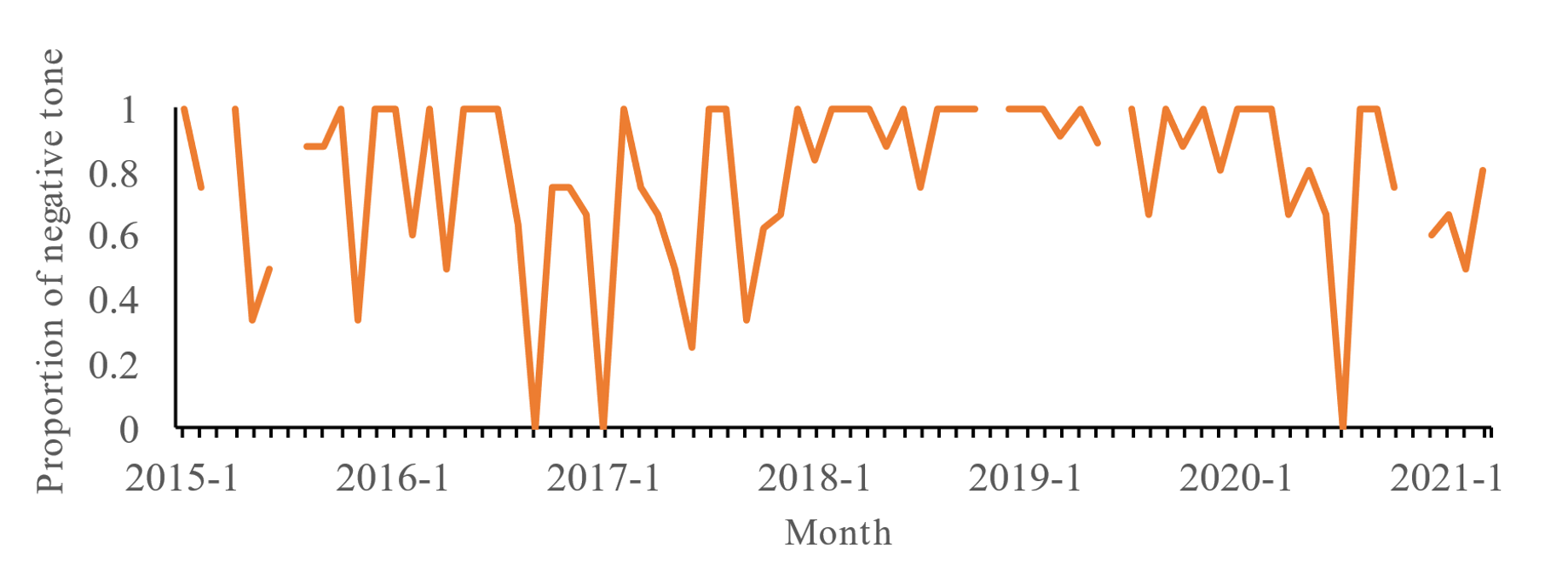


Fig 5. Monthly proportion of negative tone of the term *veneca*.

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 Fig 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.

* 1. Qualitative analysis of the terms

The qualitative analysis revealed major themes. Negative tones had four general themes: derogatory name calling, political outrage, deprivation mocking, and perceived criminal behavior/sexual deviants. Perceived criminal behavior was exclusive to men, while sexual deviancy was attributed to women. Negative tones shared the other three categories across gender lines. Neutral tones remained the same for both terms with a central theme of “descriptor” of Venezuelans by Colombians. Descriptor refers to an ethnic identifier. while positive tone tweets shared a theme of self-descriptor, or a way that Venezuelan men and women refer to themselves; for *veneco* there were sentiments of unity between citizens of both country and words of welcome; and for *veneca*, positive tone tweets referred to favorable physical attributes, such as beauty and body features.

The term *veneco* has been negatively influenced by various events, including soccer matches, political incidents, and politicians’ comments. Specifically, soccer matches between the Colombian and Venezuelan teams have been a source of negative remarks. In 2015, conversations heavily focused on the two governments due to the increase in deportations of Colombians by Venezuela, which led to comments of rejection and hatred of Colombians towards the Venezuelan government. Additionally, in 2017, the Vice-President of Colombia said he would not allocate houses for Venezuelans, which also increased the proportion of negative comments, focusing on Venezuelans as sapping available housing from Colombian citizens.

The types of comments that generalized all Venezuelan migrants as prostitutes, or that they are only useful for sex, were frequent and drove the high proportion of negative tone. On the surface, it appears to be a blanket judgment about Venezuelan women, but it actually speaks to the perceived desperation that women may be facing by leaving their country of origin with little to no resources and their willingness to perform sexual acts as a survival mechanism. Regardless of context, it is a form of gender-based violence that not only perpetuates harmful stereotypes but also contributes to the marginalization and stigmatization of Venezuelan women. Another study showed that the stigma towards the Venezuelan migrants results in real and perceived threats to their safety, well-being, and integration [14].

1. Conclusions

This study was structured into three main parts. Firstly, we employed NLP models to estimate the usage of the terms *veneco* and *veneca.* Secondly, we conducted a manual coding process to analyze the tone associated with these terms. Lastly, we performed a qualitative analysis of the terms.

The usage estimation of the terms shows that the frequency of tweets containing *veneco* is generally higher than those containing *veneca* for the majority of periods. This might be due to the broader usage of *veneco* for all Venezuelan migrants, while *veneca* specifically targets women. The proportion of tweets containing either term is higher in certain years, corresponding to an increase in the number of Venezuelan migrants arriving in Colombia. A resurgence of tweets in 2021 is possibly related to strained resources globally due to the COVID-19 pandemic, leading to tension from competition for scarce resources between Venezuelan migrants and Colombians.

The NLP models were utilized to filter and refine the dataset in this first part. For geolocation, a classification model was fine-tuned to identify whether a tweet was from Colombia or not based on the tweet's text. Another classification model was trained to determine the relevance of tweets. The spaCy es\_core\_news\_sm model [11] was selected to auto-code relevant tweets due to its highest accuracy score.

The analysis of the tweets’ tone reveals that both terms are predominantly used in negative ways, but the average negative tone associated with *veneca* is higher. The series of instances involving *veneco* initially displayed a balanced distribution between negative and non-negative sentiments until 2018. This finding suggests that the term *veneco* might have evolved over time from a neutral descriptor to a pejorative term. On the other hand, the analysis of the tone for *veneca* exhibits high variability due to the limited number of observations, which may affect the reliability of the negativity measurement.

Qualitative analysis of the tweets provided greater context for the various categories of tone. For example, a negative tone was not exclusively representative of xenophobic sentiments, criminality, or disparaging comments about Venezuelans. Rather it was representative critical or disapproving language, typically immediate responses to current events discussed on social media. In instances where derogatory name calling or mocking deprivation was tweeted, it was done in a dehumanizing way to (i.e., referring to poor physical/hygiene state, lack of food, etc.) to establish and reinforce social hierarchies and in group othering.

References

1. Wanitzek, U., *The Power of Language in the Discourse on Women's Rights: Some Examples from Tanzania.* Africa Today, 2002. **49**: p. 19 - 3.

2. Bank, T.W., *Supporting Colombian Host Communities and Venezuelan Migrants During the COVID-19 Pandemic*, in *The World Band-Results Brief*. 2021.

3. Project, V.M., *Is 'Veneco' an insult or is it an inclusive word?*, in *Veneuzuela Migration Project-Education*. 2021.

4. Press, A., *Mounting Venezuela Exodus Sparks Fears of Rising Xenophobia*, VOA, Editor. 2019.

5. 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.

6. Bilewicz, M. and W. Soral, *Hate Speech Epidemic. The Dynamic Effects of Derogatory Language on Intergroup Relations and Political Radicalization.* Political Psychology, 2020. **41**.

7. Shrivastava, J., *The Violence Of Language: A Feminist Take On The ‘Culture’ Of Abuses*, in *Feminism In India*. 2020.

8. 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.

9. osm-search, *Nominatim.* GitHub, 2023.

10. 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.

11. SpaCy, *es\_core\_news\_sw*. 2023.

12. Cañete, J., et al., *Spanish Pre-Trained BERT Model and Evaluation Data*. 2020.

13. nlptown, *bert-base-multilingual-uncased-sentiment*. 2022.

14. Vick, K., *'You Don’t Have to Be Rich to Do the Right Thing.' Colombia's President Iván Duque on Welcoming Venezuelan Refugees*, in *TIME*. 2021.

Acknowledgments

This research is funded by grant number N000141912624 by the Office of Naval Research through the Minerva Research Initiative. This research is also funded by grant number 22RT0286 by the Air Force Office of Scientific Research through the Minerva Research Initiative. None of the views reported in the study are those of the funding organization.