doi:10.5477/cis/reis.172.21

Rejection and Hate Speech in Twitter: Content Analysis of Tweets about Migrants and Refugees in Spanish

Rechazo y discurso de odio en Twitter: análisis de contenido de los tuits sobre migrantes y refugiados en español

Carlos Arcila Calderón, David Blanco-Herrero and María Belén Valdez Apolo

Key words

Sentiment Analysis
- Big Data
- Hate Speech
- Immigration
- Refugees
- Twitter

Abstract

We use Twitter to study the verbal rejection towards migrants and refugees as a potential hate speech predictor with two content analysis of tweets in Spanish collected with Twitter's API: the first analysis, manual, with 1,469 tweets; the second, automatic, uses big data techniques to study 337,116 new tweets. In the first one rejection was predominant over acceptance and neutrality. Rejection was smaller in the second one, showing how fluctuant these expressions are depending the media context. In both cases rejection toward migrants was significantly bigger than over refugees, as it had already been observed in international contexts. This work also created a training corpus about immigrant rejection, valid for future studies, and observed the negative aspects most frequently associated to rejection of immigrants, as well as the relationship existing between this and the fact of tweets being information or opinion.

Palabras clave

Análisis de sentimientos
- *Big data*
- Discurso de odio
- Inmigración
- Refugiados
- Twitter

Resumen

Se analiza el rechazo verbal al extranjero como potencial detector de discurso de odio a través de dos análisis de contenido de tuits en español recogidos con la API de Twitter: el primero, manual, a 1.469 tuits; el segundo, automático, analiza otros 337.116 tuits utilizando técnicas de *big data*. El rechazo fue predominante en el primer análisis y minoritario en el segundo, mostrando la fluctuación que experimentan estas expresiones en función del contexto mediático. En ambos casos el rechazo hacia los migrantes fue significativamente mayor que hacia los refugiados, como se había observado ya en contextos internacionales. El trabajo también generó un corpus de entrenamiento sobre rechazo al extranjero y observó los aspectos negativos asociados más frecuentemente al rechazo, así como la relación entre este y la condición informativa u opinativa del tuit.

Citation

Arcila Calderón, Carlos; Blanco-Herrero, David and Valdez Apolo, María Belén (2020). "Rejection and Hate Speech in Twitter: Content Analysis of Tweets about Migrants and Refugees in Spanish". *Revista Española de Investigaciones Sociológicas,* 172: 21-40. (http://dx.doi.org/10.5477/cis/reis.172.21)

Carlos Arcila Calderón: Universidad de Salamanca | carcila@usal.es

David Blanco-Herrero: Universidad de Salamanca | david.blanco.herrero@usal.es

María Belén Valdez Apolo: Universidad del Azuay (Ecuador)| mariabelenvaldezapolo@gmail.com

Introduction[1]

In a global context shaped by international migrations, several anti-immigration and xenophobic political options have gained power in countries around the whole world. At the same time, we can see an increase of hate against the *other* in digital platforms (Müller and Schwarz, 2018), what leads to an increase of hate speech in social media (Bartlett *et al.,* 2014), and therefore, to a potential increase of the attacks against immigrants.

In this line, the most frequent option has been to measure the attitudes towards immigrants using surveys as a tool, a method that, in this case, can be problematic due to the social desirability bias, according to which a person would hardly consider him or herself or his or her expressions as racist or xenophobic (Cea D'Ancona, 2009). It is not the goal of this paper to question the reliability of surveys as a method to measure rejection against immigrants, but to offer complementary information that allows to observe the phenomenon in all its dimensions, highlighting the analysis of the public opinion expressed in social media allows the visibilization of dominant opinions.

Considering rejection as a socially built category (Berger and Luckman, 1966), and taking into account the capacity of social media to show in the public sphere rejection attitudes of individuals against the *exogroup*, this paper seeks to determine at small and large-scale the presence of expressions of verbal rejection against migrants and refugees in social media as a potential base for other types of rejection of a bigger scale. Also, we pretend to show whether tweets in Spanish referring to migrants have a more negative connotation than those referring to refugees. For that, a manual content analysis and a large-scale text classification analysis were conducted. Additionally, with an exploratory and complementary intention, we seek to know what are the main issues associated with rejection towards these groups of people, as well as discovering whether the type of message (information/opinion) relates to the expression of rejection.

The present work broadens the existing knowledge about rejection towards migrants and refugees in social media using traditional and computational techniques. More specifically, this paper analyses the content of Twitter so that we can study the presence of expressions of verbal rejection towards immigrants in the Spanish speaking world, complementing and updating previous works, that have studied the approach towards the Refugee Crisis in different countries (Gualda and Rebollo, 2016), or that have measured more specific aspects of the rejection towards immigrants, such as gender-related stereotypes (Gallego *et al.,* 2017) or hate speech (Ben-David and Matamoros-Fernández, 2016), but that have not focused on the presence of verbal rejection.

The relevance of the public discourse in social media makes this kind of analysis an important tool to capture the feeling of societies around particular topics, as well as to predict future behaviors (Kalyanam *et al.,* 2016). This way, the implementation of analysis of rejection of immigrants in Twitter or of works that take social media as a source of data has a great potential, especially in sensitive topics that are guided by social desirability, such as this one. In this sense, the main methodological contribution of the study is the creation of a corpus of examples with samples of acceptance/neutrality and rejection of foreigners, which can be used to train future models of download and automatic analysis in Spanish.

[1] Authors would like to thank the support and the resources provided by the project Preventing Hate Against Refugees and Migrants (PHARM), funded by the European Union in the frame of the Rights, Equality and Citizenship programme (REC-RRAC-RACI-AG-2019 (GA no. 875217); and by the project Desarrollo y evaluación de un detector del discurso de odio en línea en español (STOP-HATE), funded by the Fundación General de la Universidad de Salamanca in the frame of the Plan TCUE 2018-2020 (PC-TCUE18-20_016).

Contextualization of the study

New forms of measuring rejection against migrants or refugees

Although it is the most common tool, researchers have deepened in the validity of surveys as a method to measure attitudes towards migrants and refugees due to the aforementioned social desirability bias (Cea D'Ancona, 2009; Díez Nicolás, 2009). Also, the studies based on social media are gaining relevance, as it is in them where a great part of the public discourse of contemporary societies takes place. As Schäfer and Schadauer (2019) observed, fake news and disinformation spread online are often behind contents that promote rejection against migrants and refugees. And, given that the phenomenon of fake news cannot be separated from social media (Bakir and McStay, 2018), the study of these is of particular interest. That is why Twitter has become one of the platforms that most popularity has gained for scientific researches. Focused in the attitudes towards migrants and refugees, Chaudhrey (2015) has proven the capacity of this platform to track online racism. More recent studies have tried to stablish correlations between hate speech in social media and cases of violence, like Müller and Schwarz (2018), who studied the connection between social media and hate crimes using data from Facebook and Twitter. In general terms, using digital tools like Twitter in order to download and process great volumes of data and analyze attitudes towards migrants and refugees is gaining relevant, as shown by a study with 862,999 tweets of Gallego *et al.* (2017), which includes a gender perspective to the study of the representation of refugees. Rebollo and Gualda (2017) conducted a similar study with a sample of 151,294 tweets in Spanish and, despite having different goals, this method was also used by Gualda *et al.* (2015). The present work also follows the steps of researches such as the modelling of online hate speech in Twitter conducted by Burnap and Williams (2015).

In this paper we have followed these previous efforts and we have chosen Twitter because, even though it is not a representative platform of all citizens, its easy viralization of contents, popularity and the rapidity of communication are of great interest for the monitoring and analysis of this medium. Also, this platform becomes an open register of sentiments and opinions around all kinds of topics, including hate speech and rejection, which are expressed freely and without the borders that usually appear in the offline territory.

Online hate speech against migrants and refugees

Discussion around prejudice and rejection towards the exogroup is broad in Social Sciences (Brewer, 1999; Peherson *et al.,* 2011). In our field of interest, Bourkis and Dayan (2004) point that a strong national identity relates with negative attitudes towards immigrants, something in which Verkuyten and Brug (2004) agree. However, there are still important limitations to the study of discrimination of the *different* (Billig, 2002). Specifically, Brown (2000) points out that, from the perspective of social identity, rejection of the other can go from verbal rejection to genocide. Similarly, the Perceived Ethnic Difference Questionnaire (PEDG) of Contrada *et al.* (2001) identifies verbal rejection as the most basic form of discrimination. The present paper seeks to analyze the problem from its base, studying the verbal expression of the most general forms of rejection towards migrants and refugees, becoming a starting point for future works that dig into other more specific forms of rejection.

From a theoretical point of view, there is an intense connection between the use of

language and the transmission of prejudices against the *other*, the *exogroup* (Maass *et al.,* 1989). However, and contrary to the paradigm of the hypothesis of the intergroup linguistic bias theory (Whitley and Kite, 2010; Gorham, 2006), in the transmission of prejudice through social media, the most recent empirical evidence shows that negative descriptions (such as *rejection*) towards the exogroup stop being vague or abstract and become specific, visible and measurable when they are supported by official discourses (Crandall *et al.,* 2018) or in networks with possibility of anonymity (Fox *et al.,* 2015).

At the same time, it has been observed that rejection towards immigrants, promoting their expulsion or forbidding their entrance, by influential people or opinion leaders can also lead to a potential increase in hate speech (Gualda and Rebollo, 2016). These negative descriptions and shows of rejection are, precisely, the main support of hate narrative towards highly stereotyped and vulnerable groups.

Hate speech involves the promotion of messages encouraging rejection, underestimation, humiliation, bullying, discredit or stigmatization of individuals or social groups based on attributes that go from nationality to sexual orientation. The European Commission against Racism and Intolerance (ECRI), via their General Recommendation no. 15 (2016), specifies that this speech can be motivated by race, skin color, ancestry, national or ethnic origin, age, disability, language, religion or beliefs, sex, gender, gender identity, sexual orientation or other personal characteristic or conditions. The Council of Europe, with its Recommendation no. 97 (1997), adds that it must be an expression that "spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance".

In a context in which digital and social media allow a faster and greater creation and diffusion of these contents, the relevance of hate speech comes, mainly, from its role as trigger of hate crimes. Muller and Schwarz (2018) suggest that there is a significant relationship between online hate speech and real attacks and also that "online hate speech can act as a propagating mechanism for violent crimes" (p. 24). With this basis, authors like Kreis (2017) have analyzed hate speech in Europe towards migrants and refugees in Twitter, something that Chaudhry (2015) did in Canada.

With this, we believe that it is important to determine to what extent the expressions of *rejection* towards the vulnerable group of foreigners are predominant in social media in Spanish, as well as analyzing the problematics and negative aspects associated with them and the type of message in which this rejection is expressed, so that, with that basis, more effective strategies against more direct hate speech and rejection in general can be articulated. That is how the next research questions appear:

RQ1: Which is the dominant opinion in terms of *acceptance* or *rejection* towards migrants and refugees in Twitter messages in Spanish?

RQ2: Which are the negative aspects with which the expressions of *rejection* towards migrants and refugees is associated in Twitter messages in Spanish?

RQ3: In what type of messages (information or opinion) is it more common to find *rejection* against migrants and refugees in Twitter in Spanish?

Differences between migrants and refugees

Although in the daily discourse many people might use these terms indistinctly and in the great people movements we usually find people with both profiles, it is convenient to differentiate between 'migrant(s)'

and 'refugee(s)'[2]. The 1951 Refugee Convention of Geneva says that a 'refugee' is a person who

> […] owing to well-founded fear of being persecuted for reasons of race, religion, nationality, membership of a particular social group or political opinion, is outside the country of his nationality and is unable or, owing to such fear, is unwilling to avail himself of the protection of that country.

Migrants, on the other side, choose to migrate not because of a direct death or persecution threat, but looking for an improvement in the quality of their lives, mainly for social or economic reasons —which, however, might be equally pressing, despite not having the refugee status—.

The importance of this differentiation lays, therefore, in the urgent international protection and the asylum that refugees demand, which, according to international treaties, such as the 1951 Refugee Convention, must be issued by national organisms of a country with the support of supranational organizations such as UNHCR. Even though granting somebody the refugee status might obey legislation or criteria that do not always correspond with the reality of each individual, the difference between both groups is also visible in the approach to the phenomenon by hosting societies, who tend to show greater support to those who they believe to have migrated unwillingly —such as refugees— rather than willingly —the case of migrants— (Verkuyten, 2014). O'Rourke and Sinnott (2006) and Murray and Marx (2013) support this distinction too, because in general terms people tend to be less hostile to refugees than to migrants, whether if they are "legal" or "illegal" (Murray and Marx, 2013).

With this, it is very likely that the intergroup linguistic bias theory would explain some of the differences in the ways by which prejudice is transmitted when publics with different empathic charges produced by a positive media coverage are compared (Park, 2012). This is, when empathy is bigger —as it is with refugees due to a victimization media treatment—, it is likely for to hate expression to be vaguer or more abstract, whereas in the cases of less empathy —as it is with migrants, whose media treatment is the one of a negative burden for the countries— rejection will be more evident and manifest.

This has been proven in countries such as the United States or the Netherlands, observing how the hosting society tends to see refugees as not having any alternative —and, therefore, innocent victims— and to reject them less and support them more than migrants, as it is believed that they migrate voluntarily (Verkuyten *et al.,* 2018; O'Rourke and Sinnott, 2006). These studies allow us to assume that also in the Spanish-speaking setting:

H1: Rejection in Twitter is more frequent towards migrants than towards refugees.

Method

This quantitative study has a descriptive and correlational reach and it is based in the content analysis and automated classification of texts using supervised machine learning. These techniques were applied to messages from Twitter, so that each tweet made one unit of analysis. As for the part of automated text classification, it is big data technique that uses classification algorithms to produce predictive models bas-

[2] In this text the term 'immigrant(s)' will be also included together with the term 'migrant(s)', as both of them refer to people without the refugee status and, therefore, are seen by the general public as voluntary migrants.

Beyond the inclusion of the term in the group of 'migrants' in relation to Hypothesis 1, along the article the term 'immigrant(s)' will be used together with the term 'foreigner(s)' to refer to migrants and refugees as a joint group, as both collectives, despite their different legal status, are immigrants and foreigners in the hosting country.

ing on a set of examples previously classified with different techniques, including content analysis itself. This work has two phases: the first, of manual content analysis, allowed us to answer the three research questions and the hypothesis, as well as to build the text classification model used in the second phase of large scale automated analysis. In this second phase, we answer, with a far greater volume of data, to RQ1 and H1, which are the two main elements of the study, allowing as to complement and compare the results of both phases. Given its exploratory and complementary nature, RQ2 and RQ3 were answered just in the first phase.

Manual content analysis

Sample and procedure

The first collection of tweets was executed using the Application Programming Interface (API) of Twitter, which allows the download from the real-time streaming of Twitter or from the rest of the flow. In this case we used the streaming API, which downloads all tweets published in the network anywhere in the world in the selected language[3] containing a specific keyword during the time the tool is active. During the months of April and May 2018 a random sample of 4,000 tweets in Spanish were collected with the only requisite of including the terms: 'refugiado' / 'refugiados' / 'migrante' / 'migrantes' / 'inmigrante' / 'inmigrantes'.

[3] The tool detects the language declared in the JSON —the indicated language of the tweet— and downloads those contents that fulfill this condition and that include the specified keywords either in the boy of the tweet or in the elements it includes (images, links, etc.). Therefore, it is possible that a user who has his/her account configured in Spanish introduces one of the selected keywords in a message that is written in a different language or dialect. These messages, also downloaded by the tool, were removed in the subsequent filtering phase.

Initially, the 4,000 tweets were filtered, removing those that used those words in contexts different from the migration of people, the repeated tweets, those without any logic sense, those whose meaning depended on a link or an image, those written in different languages and those that only included emojis or mentions to other users. The final sample had 1,469 tweets.

Measures

Tweets were manually classified by two trained coders following these measures:

a) *Meaning or Presence of expressions of rejection*: This classification, the most important of the study, demands the comprehension of the meaning of the tweet in order to discover the attitude towards migrants/refugees, specially the *rejection*, main variable of the study. In was coded in the categories *acceptance*, *rejection* and, if there is no position stated, *neutrality*, so that *acceptance* and *neutrality* mean lack of *rejection*. Even when the tweet is informative, it can lead to *acceptance* or *rejection* depending on its content if, for example, it shares some public statements in one or the other direction. Were coded as *neutral* those tweets in which it was not possible to discern whether the opinion or the information promotes or expresses the *acceptance* or the *rejection* of the migrant/refugee. It should be highlighted that there are tweets in which solidarity or compassion is shown; however, if there was no assumption of a defensive or welcoming position, or if rights or actions in that line were not demanded, it was not considered *acceptance* but *neutral*, as compassion does not necessarily mean the *acceptance* of the immigrant, but his/her victimization, because compassion does not avoid by itself the consideration of the immigrant as a burden for

the State. This way, under *acceptance* were only included tweet that showed welcoming or defense. Under *rejection* fell tweets that reflect a non-acceptance of migrants or refugees, or the association of them with negative aspects, such as connecting them with delinquency, an economic burden, an invasion or an avalanche, with impoverishment, etc. It is also *rejection* when the word 'refugee', 'migrant' or 'immigrant' is used in a pejorative way or as an insult. According to the Perceived Ethnic Difference Questionnaire (Contrada *et al.,* 2001), these are contents expressing rejection via offensive comments towards a person or group o via the use of pejorative or demeaning terms, but without the need for them to constitute hate speech of offensive language (Davidson *et al.,* 2017).

b) *Negative association that explain rejection*. This category digs into the reason or reasons associated with the rejection of migrants or refugees. It was measured the presence or absence of six indicators of rejection, built ad-hoc for this study using items from Díez Nicolás (2009), Wike *et al.* (2016) and Cea D'Ancona (2009) in their studies, and compared with the ideas that Gualda and Rebollo (2016) and Rebollo and Gualda (2017) observed that hosting societies associate with immigrants. This way, these indicators become a combination and synthesis of previous works, mixing in six categories the possible expressions of rejection towards foreigners. The goal is to observe, from these six big associations, which ones are more frequent and, therefore, in what aspects should the focus be in order to reduce rejection. We tried to give priority to the predominant indicator in each tweet, but, given the impossibility to choose only one, in some texts more than one argument was selected, what made the number of indicators bigger than the one of rejection messages. The indicators we used are the following ones:

— *Economic burden* (1). If it points out that foreigners mean an economic effort for the State or for its citizens, taking away social benefits or jobs that should be for the nationals of that country, if it rejects the idea of granting immigrants help or supports, or when it is considered that foreigners are in a privileged position compared to the national population when receiving State support.

— *Security threat* (2). If it is considered that immigrants are violent, blameworthy for insecurity or if they represent any kind of danger, especially terrorism.

— *Identity threat* (3). Immigrants are believed to pose a danger to the culture of the country, forcing the one of their origin countries and it is feared that immigration can make the destination country to lose its identity; or when there is annoyance with the "imposition" of religious praxis and beliefs of the immigrants. It also applies to the mentions of multiculturalism under a negative light.

— *Invasion threat* (4). It is detected with words such as 'manada', 'ola', 'miles', 'millones', 'invasión', 'avalancha'[4], referring to the huge amount of migrants and refugees. It is considered that there are a lot or too many migrants, and that they should be expelled or that the borders should be strengthened, but not for any particular reason, but for the fear of being invaded.

[4] Herd, wave, thousands, millions, invasion, avalanche.

TABLE 1. *Reliability of the measures*

Variable	Cohen's Kappa	Krippendorff's Alpha
Meaning	0.778	0.777
Explanation: Economic burden	0.754	0.753
Explanation: Security threat	0.784	0.784
Explanation: Invasion threat	0.680	0.681
Explanation: Identity threat	0.688	0.688
Explanation: Explicit rejection	0.687	0.687
Explanation: Social prejudice	0.580	0.580
Type	0.766	0.766
Average	**0.715**	**0.715**

Source: Developed by authors.

- — *Explicit rejection* (5). When the words show explicit rejection or hostility, without any further reason and/or when the words immigrant or refugee are used as an insult, in a pejorative or in a negative way, with expressions such as "Maldito refugiado", "inmigrante tenía que ser"[5], etc. This category, in which immigrants are rejected because of their condition of such, is the closest to the expressions that include hate speech and, therefore, it is the most dangerous one.
- — *Social prejudice* (6). When it is detected that the presence of immigrants harms the environment of a city or country, when it is mentioned the poverty status or the social class of the refugee or migrant, their education, abilities, etc. In line with Cortina's (2017) approaches, the foreigner is not rejected for being such, but for his/her social status.

c) *Type*. Formal category that distinguishes between information and opinion messages. As it happened in the previous case, they are both exploratory variables that pursue to broaden the main analysis; this category has special relevance due to the necessity of distinguishing facts from opinion in the current wave of post-truth, under which more importance is given to opinions than to facts when building reality (Oxford Dictionaries, 2016). If it is seen a news style, objective, it shares invitations or announcements, offers data or statistics, it was coded as Informative. If it has a personal style, with a subjective nature, with more adjectives, it was coded as Opinion. It was considered as opinion any expression from an account from an organization or association if in it a personal or subjective position is expressed, or when opinions about any information is made.

In order to ensure reliability of the measures an inter-coders test was conducted with a random sample of 150 messages (~10 % of the total sample). Cohen's Kappa and Krippendorff's Alpha (measured from 0 to 1, being 1 total agreement) were used. As it can be seen in table 1, the values of both tests are close or superior to 0.7, what shows an adequate reliability. Only the presence of social prejudice as an explanation for

5 Bloody immigrant, of course it was an immigrant…

rejention of migrants and refugees showed a bit smaller values, but as it was a less clear to rate variable that got close to 0.6 —what Neuendorf (2002) demands for exploratory research—, it was kept in the study.

Large-scale computational analysis

In a second stage, computational methods were used in order to increase the original sample and escalate the initial study with a big data approach, so that RQ1 and H1 can be answered with a stronger position. With that aim, the 1,469 messages classified in the first stage were used as examples to generate a predictive model with supervised machine learning techniques that would allow to estimate the probability of any new tweet to belong to the category *rejection* (45 %, according to the manual classification explained later in the Results section) or *acceptance/neutrality* (55 %). This division was made with the focus over the rejection category, as opposed to neutrality and acceptance, as that was the category whose measuring was interesting for the study, as it is the one that could potentially lead to hate speech or that would be even behind xenophobic or racist feelings.

With this goal, we used the NLTK and SciKit-Learn Python libraries in order to generate binary classification models with the presence of expressions of rejection as the reference category, using six algorithms that are usually applied for text classification —Original Naïve Bayes, Naïve Bayes for Multinomial Models, Logistic Regression, Lineal classifiers with Stochastic Gradient Descent training, Naive Bayes for Bernoulli's Multivariate models or Support vector machines—. Natural Language Processing techniques were also applied in order to collect the features of the sample of tagged messages.

With these tools, we started by cleaning the strange characters, such as emojis, non-linguistic symbols or letters from other alphabets, and all the text was turned into lower case letters. Then, a model in Spanish was trained for the tagging of parts of speech, using as an example the corpus *escast3lb* from the module *cess-esp* of the library NLTK, which allowed the selection of the words of the tweets that were adjectives, verbs or nouns. The 5,000 most repeated words were tokenized and became quantitative features (vectors) of the examples so that predictive models could be created.

The 1,469 messages were randomly divided in two groups: 70 % for the training corpus and 30 % for the test corpus. Optimized classifiers were generated for each of the six aforementioned algorithms and they were implemented over the training corpus with the goal of generating six classifying models. With this, a classifier based on the vote of each of the six models was created; it included a confidence indicator based on the degree of agreement of the models for each prediction. This way, the classifier selects the category —*acceptance/neutrality* or *rejection*— that most models have predicted —in the case of a tie it does it randomly—, adding a confidence indicator based on the proportion the proportion of agreement (number of votes of the selected case/number of possible votes), what allowed us to establish a confidence threshold of 0.8 (80 %) for each prediction.

Each of the six classifiers, beside the one based on the votes of the other models, was evaluated using the test corpus in order to compare the original tags with the classifications made by the automated models. We used the classic evaluation metrics in supervised machine learning (Kelleher *et al.,* 2015): accuracy, precision, recall and F-score. In table 2 it can be seen how all values were significantly over the baseline of 55 %, especially the classifier based in the votes of the other models, as it had 76.19 %, what translates into an adequate predictive power of the models.

TABLE 2. *Evaluation metrics of the models*

Algorithms	Accuracy %	Precision %	Recall %	F-score %
Original Naïve Bayes	75.36	78.10	76.92	77.50
Naïve Bayes for Multinomial Models	74.64	79.88	72.23	75.86
Naive Bayes for Bernoulli's Multivariate models	72.15	68.80	90.62	78.22
Logistic Regression	74.53	73.56	84.05	78.46
Lineal classifiers with Stochastic Gradient Descent training	71.84	73.18	77.30	75.18
Support vector machines	75.36	76.11	80.68	78.32
Classifier based on the votes of the other models	**76.19**	**73.18**	**77.30**	**75.18**

Source: Developed by authors.

Results of each models were stored in a pickle format with the goal of escalating the original research and, like this, analyze a large-scale sample with computational methods. Specifically, like in the manual analysis collected, we used Twitter's streaming API and, connecting it with the text classification models trained with the previously manually tagged messages, we automatically collected all tweets in Spanish produced from 19th to 29th of July 2019 that included any of the keywords of the original download ('refugiado' / 'refugiados' / 'migrante' / 'migrantes' / 'inmigrante' / 'inmigrantes'). During this period, a total amount of 337,116 messages were downloaded and classified in real time.

Finally, using the classifier based on the votes of the six models generated after the manually tagged messages, and establishing a confidence level of 0.8 for each prediction, the analyses was conducted. From those 337,116 collected messages, 187,305 were classified as *rejection* or *acceptance/neutrality* with the minimum confidence level established for the study.

Results

Manual analysis

To preliminary answer RQ1 about the presence of rejection in Twitter in Spanish towards the collective of migrants and refugees (table 3), we found that, from the 1,469 coded tweets, the highest percentage (45 %) showed expressions of *rejection* towards migrants or refugees in any of its dimensions. In a 16.7 % of the tweets *acceptance* was shown towards the collective, whereas in 38.3 % of the tweets no meaning was detected, that is, they were *neutral* messages.

TABLE 3. *Coding according to the tweet's*

		Frequency	Percentage	Valid percentage
Meaning	Acceptance	245	16.7	16.7
	Rejection	**660**	**44.9**	**45.0**
	Neutrality	561	38.2	38.3
	Total	1,466	99.8	100.0
Lost		3	0.2	
Total		1,469	100.0	

Source: Developed by authors.

Regarding the second research question (RQ2) about the problematics or negative aspects associated to rejection towards migrants or refugees (table 4), we obtained the following results: in 35.6 % of the tweets in which rejection was shown, this appeared explicitly in a hostile way; in 30.2 % it was because migrants or refugees were associated with a security threat; in 26.1 % rejection was due to the perception of these collectives as an economic burden; the invasion threat was mentioned in 17 % of the tweets; 7.4 % of the texts showed a social prejudice; and 4.8 % of those showing rejection did it because they felt their identity threatened.

It is worth highlighting again the possibility of some tweets to include more than an explanation or association for the rejection of migrants and refugees, reason why the total percentage goes up to 121.1 %. This way, the most frequent was to mention two (57.1 %) or three (22.4 %) negative aspects in the same text. In 14 % of the rejection tweets it was not possible to clearly identify the existence of any of the predetermined problematics.

TABLE 4. *Coding of the associations that justify rejection*

		Frequency	Percentage	Valid percentage
Explanation	Economic burden	172	26.1	26.2
	Security threat	199	30.2	30.2
	Invasion threat	112	17.0	17.0
	Identity threat	32	**4.8**	**4.8**
	Explicit rejection	**235**	**35.6**	**35.7**
	Social prejudice	49	7.4	7.4
	Total	799	121.1	121.3

Source: Developed by authors.

Regarding the *type* of message, results showed how in 75.7 % of analyzed tweets an *opinion* or personal position about a topic was expressed, in contrast with the 24.3 % of tweets that were expressed in an *informative* or news-like way. Answering RQ3 that a clear association can be found between the *type* of the tweet and its *meaning* [$\chi^2(2,\ 1{,}464) = 208.972,\ p < 0.001$]: this way, *informative* tweets are more likely to be considered *neutral*; whereas expressions of *rejection* are more likely to be found in *opinion* posts. The typified residuals show that there is a greater probability for a message to be rejection when it is an opinion tweet ($13.6 > 3.29$) and that the possibility of being neutral ($13.1 > 3.29$) is significantly bigger when it is an informative tweet, something coherent with the predominance of messages coming from media which, mostly, show a neutral position. With this, the association between the type of tweet and the meaning is significant and weak: $|\Phi| = 0.378,\ p < 0.001$. In fact, if we only study the 1,108 tweets falling under the category of *opinion*, assuming that a great deal of informative tweets come from media and that hate speech takes it more from opinions than from data —in the frame of the contemporary wave of post-truth—, we find that 610 texts, this is, a 55 %, show *rejection*.

TABLE 5. *Acceptance or rejection according to the type of tweet*

			Type of tweet	
			Opinion	**Informative**
Meaning	Acceptance	Count	178.0	66.0
		% of total	16.1	18.5
	Rejection	Count	**610.0**	49.0
		% of total	**55.0**	13.8
	Neutral	Count	320.0	**241.0**
		% of total	28.9	**67.7**

Source: Developed by authors.

Regarding H1, it was observed a bigger presence of rejection when talking about migrants than refugees, whereas acceptance or the lack of an explicit feeling were more common towards the collective of refugees (table 6). We observe that 474 messages (32.3 % of the sample) referred to refugees —either singular or plural—, while 994 (67.7 %) referred to migrants —either singular or plural—. The statistical tests showed that the differences between the meaning of the tweet and the type of foreigner that they refer to were significant [$\chi^2(2, 1{,}465) = 145.815$, $p < 0.001$], being more likely for rejection tweets to be aimed at migrants than at refugees. This way, the typified residuals point out that there exists a greater probability for a message to contain expressions of rejection when migrants are mentioned ($12.1 > 3.29$) and that the probability of being neutral ($5.2 > 3.29$) or acceptance ($8.4 > 3.29$) is significantly bigger when they are about refugees. This way, the association existing between the condition of the foreigner and the volume of rejection is significant and weak: $|\Phi| = 0.315$, $p < 0.001$.

We can add that the perception of refugees as an economic burden [$t(171.135) = -2.977$, $p < 0.01$, $d = 0.46$] and as a security threat [$t(157.788) = -2.186$, $p < 0.05$, $d = -0.35$] is significantly smaller than when we are talking about migrants. In fact, migrants are associated with an economic burden in 28 % of the rejection tweets, while only 16 % of rejection tweets pointed at refugees refer to this condition. In its side, 32 % with expressions of rejection towards migrants include the security risk, in contrast with the 22 % of tweets focused on refugees.

TABLE 6. *Rejection or acceptance towards the different type of immigrants*

			Type of immigrant	
			Refugee	**Migrant**
Meaning	Acceptance	Count	114.0	131.0
		% of total	24.1	13.2
	Rejection	Count	106.0	**554.0**
		% of total	22.4	**55.9**
	Neutral	Count	**254.0**	306.0
		% of total	**53.6**	30.9

Source: Developed by authors.

Computacional analysis

In this second stage 337,116 tweets produced from 19th to 29th of July 2019 and referring to immigrants/migrants or refugees were collected and analyzed; 187,305 of them were classified with over an 80 % confidence level. This stage sought to broaden the initial analysis and to test in a different time period the main research question and hypothesis, that is, we wanted to know with a large sample what is the volume of tweets expressing rejection towards immigrants/migrants and refugees in Twitter in Spanish and discover whether rejection is greater towards migrants than towards refugees.

As we can see in table 7, and answering RQ1, the percentage of *rejection* towards migrants and refuges made a 9.19 % of the messages in which these groups were mentioned. It is a much smaller figure than the one discovered in the small-scale analysis fifteen months earlier, what shows the variations that triggering events or opinion climates can generate in this kind of messages. Additionally, we must take into account that by modelling the presence of expressions of rejection against the other categories and controlling only the type I error (false positives), the bias of the algorithm tends to classify as *rejection* only those messages if which it is completely sure. By analyzing the classification of the total of messages (N = 337,116), without taking into account the confidence level or the agreement between models, we can see how rejection goes up to 26.68 % of messages, what proves indeed how by choosing *rejection* as the reference category the classifier includes in this category only the clearest messages, whereas all the others would fall into the other category.

TABLE 7. *Large-scale classification of messages*

		Frequency	Percentage
	Acceptance/neutral	170,084	90.81
Meaning	Rejection	17,221	9.19
	Total	187,305	100.00
Lost		0	0
Total		187,305	100.00

Source: Developed by authors.

Afterwards, we conducted two classic statistical tests in order to discover whether an association between the expression of verbal rejection and the type of collective existed, answering like that H1. With that aim, we automatically added two variables to each message that showed, on one side, whether the tweet included (1) or not (0) any of the words 'migrante', 'migrantes', 'inmigrante', 'inmigrantes', and, on the other side, whether it included (1) or not (0) any of the words 'refugiado', 'refugiados'. These categorical variables were crossed using a contingency table with the meaning of the tweet, that is, whether it expressed *rejection* or *acceptance/neutrality*. The statistical tests proved that there is a significant association between rejection and the mention of migrants [$\chi^2(1, 187{,}305) = 9{,}828.634$, $p < 0.001$]; and rejection and the mention of refugees [$\chi^2(1, 187{,}305) = 3{,}138.518$, $p < 0.001$]. In the first association we can see how the typified residuals point out the existence of a greater probability for a

message to be rejection when migrants are mentioned ($140.8 > 3.29$); whereas from the second association we can infer that said probability is smaller if refugees are mentioned ($-56.0 < -3.29$). In both cases we can talk about a significant but weak association $|\Phi| = 0.325$, $p < 0.001$ and $|\Phi| = -0.129$, $p < 0.001$, respectively. Both tests support our research hypothesis with large-scale data and during a different period than the one of the first study.

Discussion and conclusions

This research has shown an important, although fluctuant, presence of tweets showing rejection towards migrants and refugees. It should be noted that the proportion of rejection found in the studied messages does not indicate that in 2018 a 45 % of Spanish-speaking people were racist or xenophobic, nor that in 2019 that volume had fallen to 9.19 %, but it does suggest two key conclusions: first, the huge variation that can be found in the expression of rejection or acceptance towards immigrants in social media according to the most recent media phenomena; and, second, the presence of rejection in social media, that sometimes might be explained by racist or xenophobic sentiments and/or attitudes and that sometimes is expressed using hate speech, especially when it finds support in official discourses (Crandall *et al.,* 2018) or in anonym networks (Fox *et al.,* 2015), which explains why it is important to continue studying this topic.

It has been observed, both in the manual study from 2018 and in the automated one from 2019, that rejection towards migrants is significantly superior than towards refugees, who are accepted or depicted in a neutral way significantly more frequently. This agrees with what O'Rourke and Sinnott (2006), Verkuyten *et al.* (2018) or Verkuyten (2014) said, as they observed how refugees are believed not to have another option and are less rejected and more supported than migrants, as they are believed to move willingly or not escaping from something as threatening as a war; this way, also the intergroup linguistic bias theories are supported, so the different empathy volumes received by specific groups —more empathy in the media cover of refugees than in the one of migrants—, leads to less rejection (Park, 2012). Nonetheless, many studies conducted nowadays to measure the attitudes towards immigration do not differentiate between these two groups, so the differences found in this research suggest the relevance of continuing analyzing it, as well as each group and their particularities separately.

At the same time, and with an exploratory goal, it was found that it is common that those people who reject foreigners do it in most of its dimensions —as the fact of having more than one argument in lots of the tweets proves—. From these dimensions, the most frequent ones are explicit hostility towards this collective, the perceived security threat or the association of these groups with an economic burden. Given the preliminary condition of this research, we must also highlight the need to keep increasing our knowledge about the causes behind rejection towards immigrants so that we can face it in the most fitting manner, both in the initial stadiums analyzed in this study and in the more harmful forms of rejection that can be expressed via hate speech or crimes.

Finally, it is important to underline that our results can also help in exploring computational mechanisms for hate speech detection by analyzing online verbal rejection, especially in the Spanish-speaking setting. This work looks into Twitter and, consequently, into other social media as valuable sources of information for the analysis of public opinion and of attitudes of the citizenship, especially in the cases in which

studies using a survey could be limited. In this line, one of the biggest contributions of this text is the generation of a corpus of examples of acceptance and rejection of migrants and refugees[6] with which models using supervised machine learning techniques can be trained, allowing the automatic and large-scale detection of these discourses.

Limitations and future research

Despite using information from Twitter, this study still uses a limited sample. Both the number of tweets and in its collection in two particular moments —between April and May 2018 and in July 2019— avoid an absolute extrapolation, because, as we have observed, the expression of rejection and acceptance suffer important fluctuations depending on specific phenomena that can affect the public opinion. However, together with the analysis of the situation in those two moments, this work also includes a preliminary analysis of the negative aspects associated with rejection and the relationship between the type of tweet and the expression of rejection.

At the same time, given that contents were not geo-tagged, it is impossible to conduct a more detailed analysis by country, something that, on the other hand, is never easy in the analysis that use Twitter as a source, as very few users make their location public —according to Gaffney and Puschman (2014), only 1 % of Twitter's flow is geo-tagged— and a great part of these data are protected by the medium. This way, the collected data are preliminary and general, but they offer a tool for future works to compare rejection towards foreigners in different Spanish-speaking contexts. At the same time, images, link and tweets that only had emojis were excluded from the analysis, limiting it to a text analysis.

The use of social media for scientific analysis involves also some limitations given the technical difficulties of the interfaces. One example of these difficulties, mentioned also by Chaudhry (2015), is Twitter's API, which offers to free developer Twitter accounts —like the one used in this study— access only to 1 % of all published tweets when downloading them with the *stream* or limits the *rest* accesses to the last seven days. However, the quantity of data is much bigger than the one of any other analogical data collection model. Also, given the composition of users that are present in social media and in Twitter in particular, it is impossible to generalize the conclusions of the studies using these platforms as data source, because some sociodemographic groups, especially older people, are hardly represented.

Bibliography

Arcila-Calderón, Carlos; Ortega-Mohedano, Félix; Jiménez-Amores, Javier and Trullenque, Sofía (2017). "Análisis supervisado de sentimientos políticos en español: clasificación en tiempo real de tuits basada en aprendizaje automático". *El profesional de la información,* 26(5): 973-982. doi: 10.3145/epi.2017.sep.18

Bakir, Vian and McStay, Andrew (2018). "Fake News and The Economy of Emotions". *Digital Journalism*, 6(2): 154-175. doi: 10.1080/21670811.2017.1345645

Bartlett, Jamie Reffin, Jeremy; Rumbale, Noelle and Williamson, Sarah (2014). *Anti-social media*. London: Demos.

Ben-David, Anat and Matamoros-Fernández, Ariadna (2016). "Hate speech and covert discrimination on social media: Monitoring the Facebook pages of extreme-right political parties in Spain". *International Journal of Communication*, 10: 1167-1193. Available at: https://ijoc.org/index.php/ijoc/article/view/3697/1585, access December 17, 2019.

[6] This corpus is available with open access in the link: https://github.com/carlosarcila/rejection

Berger, Peter L. and Luckmann, Thomas (1966). *The Social Construction of Reality*. New York: Random House.

Billig, Michael (2002). "Henri Tajfel's 'Cognitive aspects of prejudice' and the psychology of bigotry". *British Journal of Social Psychology,* 41(2): 171-188. doi: 10.1348/014466602760060165

Bourhis, Richard V. and Dayan, Joelle (2004). "Acculturation orientations towards Israeli Arabs and Jewish immigrants in Israel". *International Journal of Psychology,* 39(2): 118-131. doi: 10.1080/00207590344000358

Brewer, Marilynn B. (1999). "The Psychology of Prejudice: Ingroup Love and Outgroup Hate?". *Journal of Social Issues*, 55(3): 429-444. doi: 10.1111/0022-4537.00126

Brown, Rupert (2000). "Social Identity Theory: past achievements, current problems and future challenges". *European Journal of Social Psychology,* 30(6): 745-778. doi: 10.1002/1099-0992-(200011/12)30:6<745::AID-EJSP24>3.0.CO;2-O

Burnap, Pete and Williams, Matthew L. (2015). "Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making". *Policy & Internet*, 7(2): 223-242. doi: 10.1002/poi3.85

Cea D'Ancona, María Ángeles (2009). "La compleja detección del racismo y la xenofobia a través de encuesta. Un paso adelante en su medición". *Revista Española de Investigaciones Sociológicas (Reis),* 125: 13-45. Available at: http://reis.cis.es/REIS/PDF/REIS_125_011231144723167.pdf, access August 28, 2019.

Chaudhry, Irfan (2015). "Hashtagging hate: Using Twitter to track racism online". *First Monday*, 20(2). doi: 10.5210/fm.v20i2.5450

Contrada, Richard J.; Gary, Melvin L.; Coups, Elliot; Egeth, Jill D.; Sewell, Andrea; Ewell, Kevin; Goyal, Tanya M. and Chasse, Valerie (2001). "Measures of ethnicity-related stress: Psychometric properties, ethnic group differences, and associations with well-being". *Journal of Applied Social Psychology,* 31: 1775-1820. doi:10.1111/j.1559-1816.2001.tb00205.x

Cortina, Adela (2017). *Aporofobia, el rechazo al pobre: un desafío para la democracia*. Madrid: Paidós.

Crandall, Christian S.; Miller, Jason M. and White, Mark H. (2018). "Changing norms following the 2016 US presidential election: The Trump effect on prejudice". *Social Psychological and Personality Science*, 9(2): 186-192. doi: 10.1177/1948550617750735

Davidson, Thomas (2017). "Automated Hate Speech Detection and the Problem of Offensive Language". En: *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)*. Available at: http://sdl.soc.cornell.edu/img/publication_pdf/hatespeechdetection.pdf, access December 17, 2019.

Díez Nicolás, Juan (2009). "Construcción de un índice de Xenofobia-Racismo". *Revista del Ministerio de Trabajo e Inmigración*, 80: 21-38. Available at: http://www.mitramiss.gob.es/es/publica/pub_electronicas/destacadas/revista/numeros/80/est01.pdf, access August 20, 2019.

European Commission against Racism and Intolerance (2016). *ECRI General Policy Recommendation N.° 15 on Combating Hate Speech*. Strasbourg: European Council.

Fox, Jesse; Cruz, Carlos and Lee, Ji Young (2015). "Perpetuating online sexism offline: Anonymity, interactivity, and the effects of sexist hashtags on social media". *Computers in Human Behavior*, 52: 436-442. doi: 10.1016/j.chb.2015.06.024

Gaffney, Devin and Puschmann, Cornelius (2014). "Data collection on Twitter". In: Bruns, A.; Weller, K.; Burgess, J.; Mahrt, M. and Puschmann, C. (eds.). *Twitter and Society*. New York: Peter Lang.

Gallego, Mar; Gualda, Estrella and Rebollo, Carolina (2017). "Women and Refugees in Twitter: Rhetorics on Abuse, Vulnerability and Violence from a Gender Perspective". *Journal of Mediterranean Knowledge*, 2(1): 37-58. doi: 10.26409/2017JMK2.1.03

Gorham, Bradley W. (2006). "News media's relationship with stereotyping: The linguistic intergroup bias in response to crime news". *Journal of communication*, 56(2): 289-308. doi: 10.1111/j.1460-2466.2006.00020.x

Gualda, Estrella and Rebollo, Carolina (2016). "The Refugee Crisis on Twitter: A Diversity of Discourses at A European Crossroads". *Journal of Spatial and Organizational Dynamics*, 4(3): 199-212. Available at: https://www.jsod-cieo.net/journal/index.php/jsod/article/view/72, access August 20, 2019.

Gualda, Estrella; Borrero, Juan Diego and Cañada, José Carpio (2015). "La 'Spanish Revolution' en Twitter (2): Redes de hashtags y actores individuales y colectivos respecto a los desahucios en España. *Revista hispana para el análisis de redes sociales, REDES*, 26(1): 1-22. doi: 10.5565/rev/redes.535

Kalyanam, Janani; Quezada, Mauricio; Poblete, Barbara and Lanckriet, Gerts (2016). "Prediction and Characterization of High-Activity Events in Social Media Triggered by Real-World News". *PloS one*, 11(12): e0166694. doi: 10.1371/journal.pone.0166694

Kelleher, John D.; MacNamee, Brian and D'Arcy, Aoife (2015). *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*. London: MIT Press.

Kreis, Ramona (2017). "#refugeesnotwelcome: Anti-refugee discourse on Twitter". *Discourse & Communication*, 11(5): 498-514. doi: 10.1177/1750481317714121

Maass, Anne; Salvi, Daniela; Arcuri, Luciano and Semin, Gün R. ((1989). "Language use in intergroup contexts: The linguistic intergroup bias". *Journal of personality and social psychology*, 57(6): 981-993. doi: 10.1037/0022-3514.57.6.981

Muller, Karsten and Schwarz, Carlo (2018). "Fanning the Flames of Hate: Social Media and Hate Crime". *SSRN Electronic Journal*. doi: 10.2139/ssrn.3082972

Murray, Kate E. and Marx, David A. (2013). "Attitudes toward unauthorized immigrants, authorized immigrants, and refugees". *Cultural Diversity and Ethnic Minority Psychology,* 19(3): 332-341. doi: 10.1037/a0030812

Neuendorf, Kimberly A. (2002). *The content analysis guidebook*. Thousand Oaks, California: Sage.

O'Rourke, Kevin H. and Sinnott Richard (2006). "The determinants of individual attitudes towards immigration". *European Journal of Political Economy*, 22(4): 838-861. doi: 10.1016/j.ejpoleco.2005.10.005

Oxford Dictionaries (2016). *Word of the year 2016 is...* Available at: https://en.oxforddictionaries.com/word-of-the-year/word-of-the-year-2016, access August 26, 2019.

Park, Sung-Yeon (2012). "Mediated intergroup contact: concept explication, synthesis, and application". *Mass Communication and Society*, 15(1): 136-159. doi: 10.1080/15205436.2011.558804

Peherson, Samuel; Brown, Rupert and Zagefka, Hanna (2011). "When does national identification lead to the rejection of immigrants? Cross-sectional and longitudinal evidence for the role of essentialist in-group definitions". *British Journal of Social Psychology,* 48(1): 61-76. doi: 10.1348/014466608X288827

Rebollo, Carolina and Gualda, Estrella (2017). "La situación internacional de las personas refugiadas y su imagen en Twitter. Un reto para la intervención desde el trabajo social". *Documentos de trabajo social*, 59: 190-207. Available at: http://www.trabajosocialmalaga.org/archivos/revista_dts/59_8.pdf, access August 28, 2019.

Schäfer, Claudia and Schadauer, Andreas (2019). "Online Fake News, Hateful Posts Against Refugees, and a Surge in Xenophobia and Hate Crimes in Austria". In: Dell'Orto, G. and Wetzstein, I. (eds.). *Refugee News, Refugee Politics: Journalism, Public Opinion and Policymaking in Europe*. Oxford: Routledge.

United Nations (1951). *Convención sobre el Estatuto de los Refugiados*. Available at: https://eacnur.org/files/convencion_de_ginebra_de_1951_sobre_el_estatuto_de_los_refugiados.pdf, access August 20, 2019.

Verkuyten, Maykel (2014). *Identity and Cultural Diversity: What Social Psychology Can Teach Us*. Hove: Routledge.

Verkuyten, Maykel and Brug, Peary (2004). "Multiculturalism and group status: The role of ethnic identification, group essentialism and protestant ethic". *European Journal of Social Psychology*, 34(6): 647-661. doi: 10.1002/ejsp.222

Verkuyten, Maykel; Mepham, Kieran and Kros, Mathijs (2018). "Public attitudes towards support for migrants: the importance of perceived voluntary and involuntary migration". *Ethnic and Racial Studies*, 41(5): 901-918. doi: 10.1080/01419870.2017.1367021

Whitley Jr., Bernard E. and Kite, Mary E. (2016). *Psychology of prejudice and discrimination*. New York: Routledge.

Wike, Richard; Stokes, Bruke and Simmons, Katie (2016). *Europeans Fear Wave of Refugees Will Mean More Terrorism, Fewer Jobs*. Available at: https://immigrazione.it/docs/2016/Pew-Research-Center-July-11-2016.pdf, access August 28, 2019.

RECEPTION: March 13, 2019
REVIEW: July 10, 2019
ACCEPTANCE: February 11, 2020

doi:10.5477/cis/reis.172.21

Rechazo y discurso de odio en Twitter: análisis de contenido de los tuits sobre migrantes y refugiados en español

Rejection and Hate Speech in Twitter: Content Analysis of Tweets about Migrants and Refugees in Spanish

Carlos Arcila Calderón, David Blanco-Herrero y María Belén Valdez Apolo

Palabras clave

Análisis de sentimientos
- *Big data*
- Discurso de odio
- Inmigración
- Refugiados
- Twitter

Resumen

Se analiza el rechazo verbal al extranjero como potencial detector de discurso de odio a través de dos análisis de contenido de tuits en español recogidos con la API de Twitter: el primero, manual, a 1.469 tuits; el segundo, automático, analiza otros 337.116 tuits utilizando técnicas de *big data*. El rechazo fue predominante en el primer análisis y minoritario en el segundo, mostrando la fluctuación que experimentan estas expresiones en función del contexto mediático. En ambos casos el rechazo hacia los migrantes fue significativamente mayor que hacia los refugiados, como se había observado ya en contextos internacionales. El trabajo también generó un corpus de entrenamiento sobre rechazo al extranjero y observó los aspectos negativos asociados más frecuentemente al rechazo, así como la relación entre este y la condición informativa u opinativa del tuit.

Key words

Sentiment Analysis
- Big Data
- Hate Speech
- Immigration
- Refugees
- Twitter

Abstract

We use Twitter to study the verbal rejection towards migrants and refugees as a potential hate speech predictor with two content analysis of tweets in Spanish collected with Twitter's API: the first analysis, manual, with 1,469 tweets; the second, automatic, uses big data techniques to study 337,116 new tweets. In the first one rejection was predominant over acceptance and neutrality. Rejection was smaller in the second one, showing how fluctuant these expressions are depending the media context. In both cases rejection toward migrants was significantly bigger than over refugees, as it had already been observed in international contexts. This work also created a training corpus about immigrant rejection, valid for future studies, andobserved the negative aspects most frequently associated to rejection of immigrants, as well as the relationship existing between this and the fact of tweets being information or opinion.

Cómo citar

Arcila Calderón, Carlos; Blanco-Herrero, David y Valdez Apolo, María Belén (2020). «Rechazo y discurso de odio en Twitter: análisis de contenido de los tuits sobre migrantes y refugiados en español». *Revista Española de Investigaciones Sociológicas,* 172:21-40. (http://dx.doi.org/10.5477/cis/reis.172.21)

La versión en inglés de este artículo puede consultarse en http://reis.cis.es

Carlos Arcila Calderón: Universidad de Salamanca | carcila@usal.es

David Blanco-Herrero: Universidad de Salamanca | david.blanco.herrero@usal.es

María Belén Valdez Apolo: Universidad del Azuay (Ecuador)| mariabelenvaldezapolo@gmail.com

Introducción[1]

En un contexto marcado por grandes migraciones globales, en los últimos años se ha visto cómo diversas opciones políticas abiertamente antinmigración y de corte xenófobo se han instalado en países de todo el mundo. Simultáneamente, vivimos un auge del odio hacia la otredad en plataformas digitales (Muller y Schwarz, 2018), lo que contribuye a un aumento del discurso de odio en redes sociales (Bartlett *et al.*, 2014) y, con ello, al potencial aumento de los ataques a extranjeros.

En esta línea, lo más frecuente ha sido medir las actitudes hacia los inmigrantes utilizando herramientas como la encuesta, método que puede resultar problemático por el sesgo de *deseabilidad social*, pues una persona difícilmente se autoevaluará y clasificará sus expresiones como racistas o xenófobas (Cea D'Ancona, 2009). No es objetivo de esta investigación cuestionar la fiabilidad de la encuesta como método para medir el rechazo al extranjero, sino brindar información complementaria que permita contemplar el fenómeno en todas sus dimensiones, destacando que el análisis de la opinión pública expresada en redes sociales permite visibilizar opiniones dominantes.

Partiendo de la idea de que el rechazo es una categoría construida socialmente (Berger y Luckman, 1966), y teniendo en cuenta la capacidad de las redes sociales de evidenciar en la esfera pública las actitudes de rechazo de los individuos hacia el exogrupo, este artículo busca determinar la presencia de expresiones de rechazo verbal hacia migrantes y refugiados en las redes sociales como potencial base de otros tipos de rechazo de mayor magnitud. Asimismo, se pretende evidenciar si los tuits en español asociados a migrantes tienen una connotación más negativa que los asociados a refugiados. Para ello, se llevará a cabo un análisis de contenidos manual y otro automatizado. A su vez, con una vocación exploratoria y complementaria, se busca conocer cuáles son las principales problemáticas que se asocian al rechazo a estos colectivos, además de descubrir si el tipo de mensaje (informativo/opinión) está relacionado con la expresión de rechazo.

Con estos objetivos, el presente trabajo amplía el conocimiento existente sobre el rechazo hacia las personas migrantes y refugiadas en los medios sociales. De manera concreta, este artículo analiza el contenido de las redes sociales para estudiar la presencia de expresiones de rechazo verbal hacia inmigrantes en el entorno hispanohablante, complementando y actualizando trabajos pasados, que han analizado la aproximación a la crisis de los refugiados en distintos países (Gualda y Rebollo, 2016) o que han medido aspectos más concretos del rechazo al inmigrante, como los estereotipos relacionados con el género (Gallego *et al.*, 2017) o el discurso de odio (Ben-David y Matamoros-Fernández, 2016), pero que no han abordado la presencia de rechazo verbal como elemento central.

Al mismo tiempo, la relevancia del discurso público en las redes sociales convierte a este tipo de análisis en una herramienta importante tanto para captar el pulso de la sociedad en torno a ciertos temas como para actuar como herramientas predictoras de futuros comportamientos (Kalyanam *et al.*, 2016). Por esto, la implementación del análisis del rechazo al extranjero en redes sociales o de trabajos que tomen los medios sociales como fuente de datos tiene un gran potencial, especialmente en temas sensibles y guiados por la deseabilidad social como este. Así, el prin-

[1] Los autores agradecen el apoyo y los recursos prestados por los proyectos Preventing Hate Against Refugees and Migrants (PHARM), financiado por la Unión Europea en el marco del programa Rights, Equality and Citizenship (REC-RRAC-RACI-AG-2019 (GA N.º 875217), y al proyecto Desarrollo y evaluación de un detector del discurso de odio en línea en español (STOP-HATE), financiado por la Fundación General de la Universidad de Salamanca en el marco del Plan TCUE 2018-2020 (PC-TCUE18-20_016).

cipal aporte metodológico de este estudio es la creación de un corpus con ejemplos de muestras de aceptación/neutralidad y rechazo de extranjeros que puede servir para entrenar modelos de descarga y análisis automatizado en español.

Contextualización del estudio

Nuevas formas de medición del rechazo al migrante o refugiado

A pesar de tratarse de la herramienta más habitual, investigadores como Cea D'Ancona (2009) y Díez Nicolás (2009) han profundizado en la validez de la encuesta como metodología para medir las actitudes hacia migrantes y refugiados por el ya citado sesgo de *deseabilidad social*. Al mismo tiempo, el estudio basado en las redes sociales va ganando fuerza, ya que es aquí donde discurre una gran parte del discurso público de las sociedades. Como observaron Schäfer y Schadauer (2019), las noticias falsas y la desinformación que se propagan en la red están a menudo detrás de contenidos que fomentan el rechazo hacia migrantes y refugiados. Y dado que el fenómeno de las *fake news* es inseparable de las redes sociales (Bakir y McStay, 2018), el estudio de estas es de especial interés. Por ello, Twitter es una de las plataformas que más popularidad ha ganado en la investigación académica. Centrados en las actitudes hacia migrantes y refugiados, Chaudhry (2015) ha demostrado la capacidad de rastrear el racismo *online* usando esta plataforma. Trabajos más recientes han intentado establecer correlaciones entre el discurso de odio en redes sociales y acontecimientos violentos, como Muller y Schwarz (2018), que investigan el vínculo entre las redes sociales y los crímenes de odio usando datos de Facebook y Twitter. En general, la utilización de herramientas digitales como Twitter para descargar y procesar grandes volúmenes de datos y analizar las actitudes hacia refugiados y migrantes va ganando peso, como demuestra el estudio con 862.999 tuits de Gallego *et al.* (2017), que incorpora una perspectiva de género al estudio de representaciones de los refugiados. Rebollo y Gualda (2017) realizaron también un estudio similar con una muestra de 151.294 tuits en español y, aunque con objetivos distintos, este método fue seguido también por Gualda *et al.* (2015). En el ámbito internacional, el presente trabajo sigue los pasos de investigaciones como la modelización del discurso de odio *online* en Twitter realizada por Burnap y Williams (2015).

En esta investigación hemos seguido estos estudios y hemos optado por Twitter, ya que, aunque no es una plataforma representativa de todos los ciudadanos, la fácil viralización de contenidos, su popularidad y la rapidez de la comunicación resultan de gran interés para su monitorización y análisis. Además, esta plataforma ofrece un registro abierto de sentimientos y opiniones acerca de asuntos de todo tipo, lo que incluye discursos de odio u otras muestras de rechazo que son expresados libremente y sin las barreras que muchas veces están presentes en espacios *offline*.

Discurso de odio hacia migrantes y refugiados

La discusión en torno al prejuicio y el rechazo hacia el exogrupo es extensa en las ciencias sociales (Brewer, 1999; Peherson *et al.*, 2011). En el campo que nos interesa, Bourhis y Dayan (2004) indican que una identidad nacional fuerte se relaciona con actitudes negativas hacia los inmigrantes, algo que también defienden Verkuyten y Brug (2004). No obstante, todavía existen importantes limitaciones en el estudio de la intolerancia hacia el diferente (Billig, 2002). Concretamente, Brown (2000) señala que, desde la perspectiva de la identidad social, el rechazo al otro puede abarcar desde

el rechazo verbal hasta el genocidio. En la misma línea, el Cuestionario de Discriminación Étnica Percibida (PEDQ) de Contrada *et al.* (2001) identifica el rechazo verbal como la forma más básica de discriminación. El presente trabajo busca abordar el problema desde su base, investigando la expresión verbal de las formas más generales de rechazo hacia migrantes y refugiados, sirviendo de punto de partida a otras investigaciones que se adentren en el estudio de otras formas más específicas de rechazo.

Desde el punto de vista teórico, existe una intensa conexión entre el uso del lenguaje y la transmisión del prejuicio hacia el *otro*, el denominado exogrupo (Maass *et al.*, 1989). Sin embargo, y contrario al paradigma de la hipótesis de la teoría del sesgo lingüístico intergrupal (Whitley y Kite, 2016; Gorham, 2006), en la transmisión del prejuicio a través de los medios sociales la evidencia empírica más reciente muestra que las descripciones negativas (como el *rechazo*) hacia el exogrupo dejan de ser vagas o abstractas para convertirse en específicas, observables y medibles cuando están amparadas en discursos oficiales (Crandall *et al.*, 2018) o en redes con capacidad de anonimato (Fox *et al.*, 2015).

Al mismo tiempo, se ha observado que el rechazo a inmigrantes, promoviendo su expulsión o prohibiendo su entrada, por figuras de elevado alcance o líderes de opinión también repercute en el potencial aumento del discurso de odio (Gualda y Rebollo, 2016). Estas descripciones negativas y muestras de rechazo son, precisamente, el principal soporte de las narrativas de odio hacia públicos altamente estereotipados y vulnerables.

El discurso del odio implica la promoción de mensajes que alienten el rechazo, el menosprecio, la humillación, el acoso, el descrédito y la estigmatización de individuos o colectivos sociales basados en atributos que van desde la nacionalidad hasta la orientación sexual. La Comisión Europea contra el Racismo y la Intolerancia (ECRI), mediante su Recomendación General n.º 15 (2016), especifica que este discurso puede venir motivado por razones de raza, color, ascendencia, origen nacional o étnico, edad, discapacidad, lengua, religión o creencias, sexo, género, identidad de género, orientación sexual y otras características o condición personales. El Consejo de Europa, en su Recomendación n.º 97 (1997), añade que debe tratarse de una expresión que «propague, incite, promueva o justifique el odio racial, la xenofobia, el antisemitismo y otras formas de odio basadas en la intolerancia».

En un contexto en el que los medios digitales y las redes sociales permiten que la creación y difusión de estos contenidos sea mayor y más rápida, la relevancia del discurso de odio deriva, sobre todo, de su papel como desencadenante de crímenes de odio (Muller y Schwarz, 2018). La investigación de Muller y Schwarz sugiere que existe una relación significativa entre el discurso de odio *online* y los ataques reales y que «la exposición a la prominencia de contenido antirrefugiados y de extrema derecha es un poderoso predictor de ataques violentos hacia refugiados» (2018: 24). Con esta base, autores como Kreis (2017) han analizado el discurso de odio en Europa hacia migrantes y refugiados en Twitter, algo que también hizo Chaudhry (2015) en Canadá.

Con esto, consideramos que conviene determinar hasta qué punto las expresiones de *rechazo* hacia el colectivo vulnerable de los extranjeros son predominantes en los medios sociales en español, así como analizar las problemáticas y aspectos negativos asociados a ellos y el tipo de mensajes en los que ese rechazo se manifiesta, para, desde esa base, poder articular estrategias más efectivas que frenen tanto los casos más directos de discurso de odio como el rechazo en general. Así surgen las siguientes preguntas de investigación:

PI1. ¿Cuál es la opinión dominante en términos de *aceptación* o *rechazo* hacia migrantes y refugiados en mensajes en español en Twitter?

PI2. ¿Cuáles son los aspectos negativos con los que se asocian las expresiones de *rechazo* hacia migrantes y refugiados en mensajes en español en Twitter?

PI3. ¿En qué tipo de mensajes (informativos o de opinión) es más frecuente encontrar *rechazo* a migrantes y refugiados?

Diferencias entre migrantes y refugiados

Aunque en el lenguaje cotidiano mucha gente puede utilizar los términos de manera indistinta y en los grandes desplazamientos nos encontramos generalmente con personas que cumplen con ambos perfiles, conviene diferenciar entre «migrante(s)» y «refugiado(s)»[2]. La Convención de Ginebra sobre el Estatuto de los Refugiados de 1951 indica que «refugiado» es la persona que

> […] debido a fundados temores de ser perseguida por motivos de raza, religión, nacionalidad, pertenencia a un determinado grupo social u opiniones políticas, se encuentre fuera del país de su nacionalidad y no pueda o, a causa de dichos temores, no quiera acogerse a la protección de su país; o que careciendo de nacionalidad y hallándose, a consecuencia de tales acontecimientos fuera del país donde antes tuviera su residencia habitual, no pueda o, a causa de dichos temores, no quiera regresar a él.

Los migrantes, por su parte, eligen trasladarse no a causa de una amenaza directa de persecución o muerte, sino en busca de una mejora en la calidad de sus vidas, principalmente por motivos sociales o económicos —que, no obstante, pueden resultar igualmente acuciantes, aunque no tengan la condición de refugiados—.

La importancia de esta diferencia radica, por tanto, en la protección internacional urgente y el asilo que demandan los refugiados, que, según acuerdos internacionales como la Convención de Ginebra, deben ser facilitados por los organismos nacionales de un país, ayudado por organizaciones supranacionales como ACNUR. A pesar de que conceder el estatus de refugiado a una persona puede obedecer a legislaciones o criterios que no siempre se corresponden con la realidad de cada individuo, la diferencia entre ambos grupos también se refleja en la aproximación hacia el fenómeno migratorio por parte de las sociedades receptoras, que tienden a mostrar un mayor apoyo a quienes perciben que han migrado involuntariamente —como los refugiados— que quienes lo han hecho voluntariamente —el caso de los migrantes— (Verkuyten, 2014). O'Rourke y Sinnott (2006) y Murray y Marx (2013) también apoyan esta distinción, pues, en términos generales, las personas tienden a ser menos hostiles hacia los refugiados que hacia los migrantes, sean «legales» o «ilegales» (Murray y Marx, 2013).

Con esto, es muy probable que la teoría del sesgo lingüístico intergrupal permita explicar parte de las diferencias en las formas en que se transmite el prejuicio cuando se comparan públicos con cargas de empatía distinta producidas por la cobertura mediática positiva (Park, 2012). Es decir, al ser la empatía mayor —como en el caso de los refugiados debido al tratamiento mediático de victimización—, es probable que la expresión del rechazo sea más vaga o abstracta, mientras que en los casos de menor empatía —como en el caso de los migran-

[2] En este estudio se incluye el término «inmigrante(s)» junto al término «migrante(s)», pues en ambos casos se trata de personas que no poseen la condición de refugiado y que, por lo tanto, son percibidos por el público general como migrantes voluntarios.

Más allá de la inclusión de este término en el grupo de los «migrantes» a ojos de la Hipótesis 1, a lo largo de estas páginas se utilizarán los términos «inmigrante» y «extranjero» para hacer referencia a migrantes y refugiados de forma conjunta, pues ambos colectivos, más allá de su estatus legal, son inmigrantes y extranjeros en el país de acogida.

tes cuyo tratamiento mediático es de carga negativa para los países— el rechazo será más evidente y manifiesto.

Esto ha sido probado en países como Estados Unidos o Países Bajos, observando que la sociedad de acogida tiende a considerar que los refugiados no tienen alternativa —y, por lo tanto, se les considera víctimas inocentes— y son menos rechazados y más apoyados que los migrantes, pues se entiende que estos se desplazan voluntariamente (Verkuyten *et al.*, 2018; O'Rourke y Sinnott, 2006). Estas investigaciones nos permiten asumir que también encontraremos esta actitud en el contexto hispanohablante:

H1. El rechazo en Twitter es más frecuente hacia migrantes que hacia refugiados.

Método

Este estudio cuantitativo tiene alcance descriptivo y correlacional, y se basa en el análisis de contenido y en la clasificación automatizada de textos basada en el aprendizaje automático supervisado (*supervised machine learning*). Esta técnica se aplicó a mensajes de Twitter, de manera que cada tuit conformaba una unidad de análisis. Por su parte, la clasificación automatizada de textos es una técnica de *big data* que usa algoritmos de clasificación para generar modelos predictivos basados en un conjunto de ejemplos previamente etiquetados con diferentes técnicas, incluido el propio análisis de contenido. El trabajo tiene dos etapas: la primera, de análisis de contenido manual, sirvió para responder a las tres preguntas de investigación y a la hipótesis, además de para elaborar el modelo de clasificación de textos utilizado en la segunda etapa de análisis automatizado a gran escala. En esta segunda fase se da respuesta, con un volumen de datos mucho mayor, a la PI1 y a la H1, que constituyen los dos elementos principales del estudio, permitiendo complementar y comparar los resultados de ambas fases. Dado su carácter exploratorio y complementario, las Preguntas de Investigación 2 y 3 son respondidas únicamente en la primera etapa.

Análisis manual de contenidos

Muestra y procedimiento

La primera descarga de los tuits se realizó en el entorno integrado Pycharm, a través de la herramienta Autocop (Arcila-Calderón *et al.*, 2017), conectada a la Application Programming Interface (API) de Twitter, que permite descargar tuits en tiempo real (*streaming*) o del historial (*rest*). En este caso se utilizó el API *streaming*, que descarga todos los tuits publicados en la red en cualquier lugar del mundo en el idioma seleccionado[3] y que contengan una palabra clave determinada durante todo el tiempo que la herramienta esté activa. Durante los meses de abril y mayo de 2018 se descargaron de manera aleatoria 4.000 tuits en español con el único requisito de que incluyeran alguna de las palabras clave: «refugiado», «refugiados», «migrante», «migrantes», «inmigrante» e «inmigrantes».

Inicialmente se filtraron los 4.000 tuits, eliminando aquellos que utilizaran las palabras clave en otro contexto que el de la migración de personas, los tuits repetidos, los que no tuviesen sentido lógico, aquellos cuyo signi-

[3] La herramienta detecta el idioma declarado en el JSON —el lenguaje de marcado del tuit— y descarga aquellos contenidos que cumplan este requisito y que incorporen las palabras clave introducidas tanto en el cuerpo del tuit como en los elementos que incorpora (imágenes, enlaces, etc.), por lo que es posible que un usuario que tenga su cuenta configurada en español introduzca uno de los términos seleccionados en un mensaje redactado en otro idioma o dialecto. Estos mensajes, que también fueron descargados por la herramienta, fueron eliminados en la posterior fase de limpieza.

ficado dependiera de un hipervínculo o una imagen, los redactados en otros idiomas y aquellos que solo contuvieran emoticonos o menciones a otros usuarios, obteniendo una muestra final de 1.469 tuits.

Medidas

Los tuits fueron clasificados manualmente por dos codificadores entrenados siguiendo las siguientes medidas:

a) *Sentido o presencia de expresiones de rechazo*: esta clasificación, la principal del estudio, exige la comprensión del sentido del tuit para conocer la actitud hacia los migrantes o refugiados, especialmente el *rechazo*, variable principal del estudio. Se codificó en las categorías *rechazo, aceptación*, o, si no se asumía ninguna postura, *neutral*, de manera que la *aceptación* y la *neutralidad* implicaban ausencia de rechazo. Aun cuando un tuit es informativo, este puede despertar aceptación o rechazo en función de su contenido si, por ejemplo, comparte declaraciones en uno u otro sentido. Se codificaron como *neutral* aquellos tuits en los que no se puede detectar si la opinión o la información promueve o expresa la aceptación del migrante o refugiado, o el rechazo. Conviene destacar que hay tuits que expresan solidaridad o compasión, sin embargo, si no se asume una actitud de defensa, acogida o se exigen derechos y acciones, no fueron considerados como *aceptación*, si no como *neutral*, puesto que la compasión no implica necesariamente la aceptación del inmigrante, sino su victimización, ya que la compasión no impide por sí misma que se considere, al mismo tiempo, al inmigrante como una carga para el Estado. En *aceptación*, por lo tanto, se incluyeron los tuits que manifestaron acogida, bienvenida, integración o defensa. Se identifica como *rechazo* los tuits que reflejen la no aceptación de migrantes o refugiados, o la asociación de estos con aspectos negativos, como vincularlos con la delincuencia, con una carga económica, con una invasión o avalancha, con el empobrecimiento, etc. También es rechazo cuando se utiliza el término refugiado, migrante o inmigrante de manera despectiva o como un insulto. Siguiendo la línea del Cuestionario de Discriminación Étnica Percibida (Contrada *et al.*, 2001), se trata de contenidos que expresen rechazo a través de comentarios ofensivos contra una persona o grupo, o a través de la utilización de nombres peyorativos o descalificativos, pero sin exigir la presencia de discurso de odio ni de lenguaje ofensivo (Davidson *et al.*, 2017).

b) *Asociaciones negativas que justifican el rechazo*. Esta categoría indaga en el o los motivos asociados con el rechazo a los migrantes o refugiados. Se midió la presencia o ausencia de seis indicadores de rechazo, que se construyeron *ad hoc* para este estudio gracias a los ítems utilizados por Díez Nicolás (2009), Wike *et al.* (2016) y Cea D'Ancona (2009) en sus estudios, y que fue contrastada con las ideas que Gualda y Rebollo (2016) y Rebollo y Gualda (2017) observaron que las poblaciones autóctonas asocian con los inmigrantes. Estos indicadores constituyen, por lo tanto, una combinación y síntesis de los anteriores trabajos, agrupando en seis categorías las posibles expresiones de rechazo a los inmigrantes. El objetivo es descubrir, de estas seis grandes asociaciones, cuáles son más habituales y, por lo tanto, en qué aspectos se debe incidir para reducir ese rechazo. Se intentó primar el indicador que fuera predominante en cada tuit; sin embargo, ante la imposibilidad de seleccionar uno solo, en algunos textos se seleccionó más de un argumento, lo que generó que el porcentaje acu-

mulado de indicadores fuera superior al 100% de tuits que expresan rechazo. Los indicadores utilizados son los siguientes:

- *Carga económica* (1). Si indica que los extranjeros suponen un esfuerzo económico para el Estado o sus ciudadanos, quitando beneficios sociales o puestos de trabajo que corresponderían a los nacionales de ese país; si expresa desacuerdo con que se concedan ayudas a extranjeros, y también si considera que los extranjeros están en posición de ventaja en comparación con los nacionales a la hora de recibir apoyo estatal.
- *Amenaza a la seguridad* (2). Si se considera que los inmigrantes son violentos, responsables de la inseguridad o que representan un peligro de cualquier clase, especialmente terrorismo.
- *Amenaza a la identidad* (3). Se muestra que los inmigrantes amenazan la cultura del país imponiendo las de sus países de origen y se teme que la inmigración acabe provocando que el país de destino pierda su identidad, o cuando se muestra contrariedad por la «imposición» de las prácticas religiosas o creencias de los inmigrantes. También cuando se habla de multiculturalismo de una manera negativa.
- *Amenaza de invasión* (4). Se detecta por la presencia de palabras como «manada», «ola», «miles», «millones», «invasión», «avalancha», haciendo referencia a la gran cantidad de migrantes y refugiados. Se considera que hay «muchos» o «demasiados» migrantes, y que estos deberían ser expulsados o que se deberían fortalecer las fronteras, pero no por causas concretas, sino por el miedo a ser invadido.
- *Rechazo manifiesto* (5). Cuando las palabras expresan rechazo u hostilidad explícita, sin especificar motivos y/o se utiliza las palabras «inmigrante» o «refugiado» como un insulto, de manera despectiva o de forma negativa, con expresiones como «maldito refugiado», «inmigrante tenía que ser», etc. Esta categoría, en la que los inmigrantes son rechazados por su condición de tales, es la que más semejanza guarda con las expresiones que incluyen discurso de odio y es, por lo tanto, la más peligrosa.
- *Prejuicio social* (6). Cuando se destaca que la presencia de los inmigrantes daña el «ambiente» de la ciudad o país, se menciona el estatus de pobreza o de clase social del refugiado, su educación, aptitudes, etc. En línea con los planteamientos de Cortina (2017), no se desprecia al inmigrante por su condición de tal, sino por su clase social.

c) *Tipo*: categoría formal que distingue entre mensajes informativos o de opinión. Como la anterior, son variables exploratorias que buscan ampliar el análisis principal; esta categoría tiene especial relevancia por la necesidad de distinguir entre hechos y opiniones en la ola de posverdad actual, según la que se concede mayor peso a las opiniones que a los hechos en la construcción de la realidad (Oxford Dictionaries, 2016). Si se reconoce que es de tipo noticioso, objetivo, difunde invitaciones o convocatorias, presenta datos o estadísticas, se codificó como *Informativo*. Si es de tipo personal, con un carácter subjetivo, con más adjetivación, se codificó como *Opinión*. Se consideró también opinión una expresión de una cuenta de una organización u colectivo si en esta se expresa con una postura personal o subjetiva, o

cuando se hacen valoraciones sobre algún tipo de información.

Para garantizar la fiabilidad de las medidas se realizó una prueba de intercodificador con una muestra aleatoria de 150 mensajes (~10% de la muestra total). Se utilizaron los estadísticos Kappa de Cohen y Alpha de Krippendorff (medidas de 0 a 1, donde 1 refleja el máximo acuerdo). Como se aprecia en la tabla 1, los valores de ambas pruebas son cercanos o superiores a 0,7, lo que demuestra una fiabilidad adecuada. Solo la presencia de prejuicios por la clase social como justificación del rechazo al migrante o refugiado obtuvo valores algo menores pero, al tratarse de una variable menos clara de apreciar y por acercarse al 0,6 que Neuendorf (2002) demanda en las investigaciones exploratorias, se mantuvo en el estudio.

TABLA 1. *Fiabilidad de las medidas*

Variable	Kappa de Cohen	Alpha de Krippendorff
Sentido o presencia de rechazo	0,778	0,777
Argumento: carga económica	0,754	0,753
Argumento: amenaza a la seguridad	0,784	0,784
Argumento: amenaza de invasión	0,680	0,681
Argumento: amenaza a la identidad	0,688	0,688
Argumento: rechazo manifiesto	0,687	0,687
Argumento: prejuicio social	0,580	0,580
Tipo	0,766	0,766
Media	**0,715**	**0,715**

Fuente: Elaboración propia.

Análisis computacional a gran escala

En una segunda etapa, se utilizaron métodos computacionales para ampliar la muestra original y escalar el estudio inicial con un enfoque de *big data*, que permitiera responder a la PI1 y a la H1 con mayor propiedad. Para ello, se utilizaron los 1.469 mensajes clasificados en la primera etapa como ejemplos para generar un modelo predictivo con técnicas de aprendizaje automático supervisado que permitieran estimar la probabilidad de cada nuevo tuit de pertenecer a la clase *rechazo* (el 45%, según la codificación manual explicada en los resultados) o *aceptación/neutral* (55%). Esta división se realizó para centrar el análisis en el *rechazo*, como oposición a la neutralidad o a la aceptación, pues es esa categoría la que nos interesa medir, ya que podría dar lugar a un potencial discurso de odio e, incluso, la que estaría detrás de sentimientos xenófobos o racistas.

Con esta intención, se utilizaron las librerías NLTK y SciKit-Learn de Python para generar modelos de clasificación binaria con la presencia de expresiones de rechazo como categoría de referencia, utilizando seis algoritmos habitualmente aplicados a la clasificación de textos —Naive Bayes original, Naive Bayes para modelos multimodales, Naive Bayes para modelos multivariados Bernoulli, Regresión logística, Regresión logística con gradiente descendente estocástico y Máquinas de vectores

soporte con estimador SVC—. También se aplicaron técnicas de procesamiento de lenguaje natural (*natural language processing*, NLP) para extraer las características del conjunto de mensajes etiquetados.

Con estas herramientas, se procedió en primer lugar a limpiar los caracteres extraños, como emoticonos, símbolos no lingüísticos o grafías de otros alfabetos, y se convirtió todo el texto a letras minúsculas. Luego se entrenó un modelo en castellano para el etiquetado de partes de discurso (*parts of speech*, POS) basado en el corpus de ejemplo *es-cast3lb* del módulo *cess-esp* contenido en la librería NLTK, que permitió seleccionar solo las palabras de los tuits que fueran adjetivos, verbos o sustantivos. Las 5.000 palabras más repetidas de estos tipos se «tokenizaron» y se convirtieron en características cuantitativas (vectores) de los ejemplos para poder generar los modelos predictivos.

Los 1.469 mensajes se dividieron aleatoriamente en dos grupos: 70% para el corpus de entrenamiento y 30% para el corpus de prueba. Se generaron clasificadores optimizados para cada uno de los seis algoritmos mencionados y se implementaron sobre el corpus de entrenamiento con el fin de generar seis modelos de clasificación. Con esto, se generó un clasificador basado en el voto de cada uno de los seis modelos generados con un indicador de confianza basado en el grado de acuerdo de los modelos para cada predicción. Así, el clasificador elige la categoría —*aceptación/neutral* o *rechazo*— que la mayoría de los modelos haya predicho —si hay empate, lo hace aleatoriamente—, añadiendo un indicador de confianza basado en la proporción de dicho acuerdo (número de votos para la clase mayoritaria/Número de votos posibles), lo que permitió establecer un umbral de confianza superior a 0,8 (80%) para cada predicción.

Cada uno de los seis clasificadores, además del basado en la votación de los otros modelos, fue evaluado utilizando el corpus de prueba para comparar las etiquetas originales con las clasificaciones producidas por los modelos creados. Se utilizaron las métricas de evaluación clásicas en aprendizaje automático supervisado (Kelleher *et al.*, 2015): la exactitud (*accuracy*), la precisión (*precision*), el recuerdo (*recall*) y la media armónica (*F-score*). En la tabla 2 se puede apreciar cómo todos los valores estuvieron significativamente por encima de la línea base de 55%; en especial, el clasificador basado en la votación de los modelos obtuvo una exactitud del 76,19%, lo que se traduce en un adecuado poder predictivo de los modelos.

TABLA 2. *Métricas de evaluación de los modelos*

Algoritmos	Accuracy %	Precision %	Recall %	F-score %
Naive Bayes original	75,36	78,10	76,92	77,50
Naive Bayes para modelos multimodales	74,64	79,88	72,23	75,86
Naive Bayes para modelos multivariados Bernoulli	72,15	68,80	90,62	78,22
Regresión logística	74,53	73,56	84,05	78,46
Regresión lineal con gradiente descendente estocástico	71,84	73,18	77,30	75,18
Máquinas de vectores soporte (estimador SVC)	75,36	76,11	80,68	78,32
Clasificador basado en la votación de los modelos	**76,19**	**73,18**	**77,30**	**75,18**

Fuente: Elaboración propia.

Los resultados de cada modelo fueron almacenados en formato *pickle* con el fin de escalar la investigación inicial y así analizar una muestra de gran tamaño con métodos computacionales. Específicamente, como en la descarga para el análisis manual, se utilizó la API *streaming* de Twitter y, enlazándola con los modelos de clasificación de textos entrenados a partir de los mensajes etiquetados manualmente, se recogieron automáticamente todos los tuits en español producidos entre el 19 y el 29 de julio de 2019 que incluyeran las mismas palabras clave de la recolección inicial («refugiado», «refugiados», «migrante», «migrantes», «inmigrante» e «inmigrantes»). Durante este periodo se descargaron y clasificaron en tiempo real un total de 337.116 mensajes.

Finalmente, utilizando el clasificador basado en la votación de los seis modelos generados a partir de los mensajes etiquetados manualmente, y estableciendo un nivel de confianza del 0,8 para cada predicción, se procedió al análisis. De los 337.116 mensajes recogidos, 187.305 fueron clasificados como *rechazo* o como *aceptación/neutral* con el mínimo de confianza establecido por el estudio.

Resultados

Análisis manual

Respondiendo de manera preliminar a la PI1 sobre la presencia de rechazo en Twitter en español hacia el colectivo de migrantes y refugiados (tabla 3), encontramos que, de los 1.469 tuits codificados, en el porcentaje más alto de mensajes (45%) se encontraron expresiones de *rechazo* en cualquiera de sus dimensiones. En un 16,7% de los tuits se expresó *aceptación* hacia el colectivo, mientras que un 38,3% de los tuits fueron mensajes *neutros*.

TABLA 3. *Clasificación manual del sentido de los tuits*

		Frecuencia	Porcentaje	Porcentaje válido
Sentido	Aceptación	245	16,7	16,7
	Rechazo	**660**	**44,9**	**45,0**
	Neutral	561	38,2	38,3
	Total	1.466	99,8	100,0
Perdidos		3	0,2	
Total		1.469	100,0	

Fuente: Elaboración propia.

Con respecto a la segunda pregunta de investigación (PI2) sobre las problemáticas o aspectos negativos asociados al rechazo hacia los migrantes o refugiados (tabla 4), se obtuvieron los siguientes resultados: en el 35,6% de los tuits en los que se mostraba rechazo, este se manifestaba de forma explícita de manera hostil; en el 30,2% se debía a que los migrantes o refugiados eran asociados con una amenaza para la seguridad; en el 26,1% el rechazo se debía a que estos colecti-

vos eran percibidos como una carga económica; la amenaza de invasión se mencionó en un 17% de los tuits; un 7,4% de los textos mostraban un prejuicio social; y el 4,8% de quienes mostraron rechazo lo hicieron por sentir amenazada su identidad.

Cabe destacar, de nuevo, la posibilidad de que algunos tuits incluyeran más de un argumento o asociación para rechazar a migrantes y refugiados, motivo por el que el total asciende a 121,1%; lo más frecuente es que se mencionaran dos (57,1%) o tres (22,4%) aspectos negativos en el mismo texto. En el 14% de los tuits de rechazo no fue posible identificar de forma clara la existencia de ninguna de las problemáticas predeterminadas.

TABLA 4. *Codificación de las asociaciones que justifican el rechazo*

		Frecuencia	Porcentaje	Porcentaje válido
Justificación	Carga económica	172	26,1	26,2
	Amenaza a la seguridad	199	30,2	30,2
	Amenaza de invasión	112	17,0	17,0
	Amenaza a la identidad	32	4,8	4,8
	Rechazo manifiesto	**235**	**35,6**	**35,7**
	Prejuicio social	49	7,4	7,4
	Total	799	121,1	121,3

Fuente: Elaboración propia.

Con respecto al *tipo* de mensaje, los resultados mostraron que en el 75,7% de los tuits analizados se mostraba una *opinión* o posición acerca del tema, frente a un 24,3% de tuits que fueron expresados de manera *informativa* o noticiosa. Respondiendo a la PI3, encontramos una clara asociación entre el *tipo* de tuit y el *sentido* del mismo [$\chi^2(2, 1.464) = 208{,}972$, $p < 0{,}001$]: así, los tuits *informativos* tienen mayor probabilidad de ser clasificados como tuits *neutros*; mientras que los tuits de *rechazo* tienden a pertenecer a publicaciones de *opinión*. Los residuos tipificados señalan que existe una mayor probabilidad de que el mensaje sea de rechazo cuando se trata de un tuit informativo ($13{,}6 > 3{,}29$) y que la probabilidad de que sean neutros ($13{,}1 > 3{,}29$) es significativamente mayor cuando se trata de tuits informativos, algo coherente con el predominio de los mensajes provenientes de medios de comunicación que, en su mayoría, muestran un sentido neutral. Con esto, la asociación entre el tipo de tuit y el sentido es significativa y débil: $|\Phi| = 0{,}378$, $p < 0{,}001$. De hecho, si estudiamos únicamente los 1.108 tuits considerados de *opinión*, asumiendo que una gran parte de los tuits informativos proceden de medios de comunicación y que el discurso de odio se nutre más de opiniones que de datos —enmarcado en la ola de posverdad contemporánea—, encontramos que 610 textos, esto es, un 55%, son de rechazo.

TABLA 5. *Sentido de aceptación o rechazo según el tipo de tuit*

			Tipo de tuit	
			Opinión	**Informativo**
Sentido	Aceptación	Recuento	178,0	66,0
		% del total	16,1	18,5
	Rechazo	Recuento	**610,0**	49,0
		% del total	**55,0**	13,8
	Neutral	Recuento	320,0	**241,0**
		% del total	28,9	**67,7**

Fuente: Elaboración propia.

En cuanto a la H1, se pudo observar una mayor presencia de rechazo cuando se trataba de migrantes que de refugiados, mientras la aceptación o la ausencia de un sentimiento explícito fueron mayores hacia el colectivo de refugiados (tabla 6). Observamos que 474 mensajes (un 32,3% de la muestra) hacían alusión a refugiados —de manera singular o plural—, mientras 994 (el 67,7%) se referían a migrantes —tanto en singular como en plural—. Las pruebas estadísticas mostraron que las diferencias entre el sentido del tuit y el tipo de inmigrante al que hace referencia fueron significativas [$\chi^2(2,\ 1.465) = 145{,}815,\ p < 0{,}001$], siendo más probable que los tuits de rechazo estuvieran dirigidos a migrantes que a refugiados. Así, los residuos tipificados señalan que existe una mayor probabilidad de que el mensaje incluya expresiones de rechazo cuando se menciona a los migrantes ($12{,}1 > 3{,}29$) y que la probabilidad de que sean neutros ($5{,}2 > 3{,}29$) o de aceptación ($8{,}4 > 3{,}29$) es significativamente mayor cuando se trata de refugiados. Por lo tanto, la asociación que existe entre la condición del extranjero y el volumen de rechazo es significativa y débil: $|\Phi| = 0{,}315$, $p < 0{,}001$.

Podemos añadir que la percepción de los refugiados como una carga económica [$t(171{,}135) = -2{,}977$, $p < 0{,}01$, $d = 0{,}46$] y como una amenaza a la seguridad [$t(157{,}788) = -2{,}186$, $p < 0{,}05$, $d = -0{,}35$] es significativamente menor que cuando se trata de migrantes. De hecho, los migrantes son asociados con una carga económica en el 28% de los tuits de rechazo, mientras que solo un 16% de los tuits de rechazo hacia refugiados se refiere a esta condición. Por su parte, el 32% de los textos con expresiones de rechazo hacia migrantes incluye el componente de riesgo para la seguridad, por el 22% de los tuits centrados en refugiados.

TABLA 6. *Sentido de aceptación o rechazo hacia los distintos tipos de inmigrantes*

			Tipo de inmigrante	
			Refugiado	**Migrante**
Sentido	Aceptación	Recuento	114,0	131,0
		% del total	24,1	13,2
	Rechazo	Recuento	106,0	554,0
		% del total	22,4	55,9
	Neutral	Recuento	254,0	306,0
		% del total	53,6	30,9

Fuente: Elaboración propia.

Análisis computacional

En esta segunda etapa se descargaron y analizaron 337.116 tuits producidos entre el 19 y el 29 de julio de 2019 y que hacían referencia a inmigrantes/migrantes o refugiados; 187.305 de ellos fueron clasificados con un nivel de confianza superior al 80%. Esta etapa pretendía ampliar el análisis inicial y poner a prueba en otro periodo temporal la principal pregunta de investigación e hipótesis, esto es, se quería conocer con una muestra de gran tamaño cuál es el volumen de tuits que expresan rechazo hacia inmigrantes/migrantes y refugiados en Twitter en español y comprobar si dicho rechazo era mayor hacia los migrantes que hacia los refugiados.

Como vemos en la tabla 7, y respondiendo a la PI1, el porcentaje de rechazo hacia migrantes y refugiados se situó en un 9,19% de los mensajes en los que se mencionaba a estos colectivos. Se trata de una cifra mucho menor a la encontrada a pequeña escala quince meses antes, lo que refleja las fluctuaciones sobre este tipo de mensajes en función de eventos detonadores o de climas de opinión. Adicionalmente, debemos tener en cuenta que, al modelar las expresiones de rechazo frente a las otras categorías y controlar fundamentalmente el error tipo I (falsos positivos), el sesgo del algoritmo se inclina a clasificar como *rechazo* solo aquellos mensajes de los que esté completamente seguro. Al analizar la clasificación del total de mensajes (N = 337.116), sin tomar en cuenta o no la confianza o acuerdo entre los modelos, observamos que el porcentaje de rechazo asciende al 26,68%, lo que demuestra, precisamente, que, al elegir *rechazo* como categoría de referencia, el clasificador solo incluye en dicha categoría los mensajes de los que esté más seguro, mientras que en la otra categoría entrarían todos los demás.

TABLA 7. *Clasificación de los mensajes a gran escala*

		Frecuencia	Porcentaje
Sentido	Aceptación / Neutral	170.084	90,81
	Rechazo	17.221	9,19
	Total	187.305	100,0
Perdidos		0	0
Total		187.305	100,0

Fuente: Elaboración propia.

Seguidamente llevamos a cabo dos pruebas estadísticas clásicas para comprobar si existía alguna asociación entre la expresión verbal de rechazo y el tipo de colectivo, respondiendo así a la H1. Para ello, añadimos de forma automática dos variables a cada mensaje, que reflejaban, por un lado, si el tuit contenía (1) o no (0) las palabras «migrante», «migrantes», «inmigrante» e «inmigrantes»; y, por otro, si incluía (1) o no (0) las palabras «refugiado» y «refugiados». Estas variables categóricas se cruzaron por medio de una tabla de contingencia con la variable sentido del tuit, es decir, si expresaba rechazo o aceptación/neutralidad. Las pruebas estadísticas revelaron que existe una asociación significativa entre el rechazo y la mención a migrantes [$\chi^2(1,\ 187.305) = 9.828{,}634,\ p < 0{,}001$]; y

el rechazo y la mención a refugiados [$\chi^2(1, 187.305) = 3.138{,}518$, $p < 0{,}001$]. Al analizar la primera asociación, los residuos tipificados señalan que existe una mayor probabilidad de que el mensaje sea de rechazo cuando se menciona a los migrantes ($140{,}8 > 3{,}29$); mientras que de la siguiente asociación se desprende que dicha probabilidad es menor cuando se menciona a refugiados ($-56{,}0 < -3{,}29$). En ambos casos podríamos hablar de una asociación significativa pero débil: $|\Phi| = 0{,}325$, $p < 0{,}001$ y $|\Phi| = -0{,}129$, $p < 0{,}001$, respectivamente. Ambas pruebas respaldan nuestra hipótesis de investigación con datos a gran escala y durante un periodo temporal diferente al primer estudio.

Discusión y conclusiones

La investigación ha mostrado una presencia notable, aunque fluctuante, de tuits que muestran rechazo hacia migrantes y refugiados. Conviene anotar que el porcentaje de rechazo encontrado en los mensajes estudiados no indica que en 2018 hubiera un 45% de personas de habla hispana racistas o xenófobas ni que en 2019 ese volumen hubiera caído al 9,19%, pero sí sugiere dos conclusiones clave: por un lado, la enorme variación que se puede observar en las expresiones de rechazo o aceptación de inmigrantes en las redes sociales en función de los últimos fenómenos mediáticos; y, por otro lado, la presencia de rechazo en redes sociales que en algunos casos puede estar basado en sentimientos y/o actitudes racistas o xenófobos y que en ocasiones es manifestada a través de discursos de odio, especialmente cuando se ampara en discursos oficiales (Crandall *et al.*, 2018) o en redes anónimas (Fox *et al.*, 2015), por lo que conviene seguir analizando esta materia.

Se ha apreciado, tanto en el estudio manual de 2018 como en el automatizado de 2019, que el rechazo mostrado hacia los migrantes es significativamente superior que hacia los refugiados, que son aceptados o retratados de una forma neutra de forma significativamente más frecuente. Esto coincide con lo apuntado por O'Rourke y Sinnott (2006), por Verkuyten *et al.* (2018) o por Verkuyten (2014), que observaron cómo se percibe que los refugiados no tienen opción y son menos rechazados y más apoyados que los migrantes, pues se considera que se trasladan voluntariamente o que no huyen de algo tan amenazante como una guerra; de esta forma se apoyan también las teorías sobre el sesgo lingüístico intergrupal, por las que la distinta carga de empatía recibida por determinados públicos —mayor empatía en la cobertura mediática de refugiados que de migrantes—, implica menor rechazo (Park, 2012). No obstante, muchos de los estudios que se llevan a cabo actualmente para medir las actitudes hacia la inmigración no distinguen entre estos dos grupos, por lo que las diferencias encontradas en este estudio sugieren la importancia de continuar analizando esta diferencia y a cada grupo y sus características particulares por separado.

Al mismo tiempo, y con vocación exploratoria, se observó que lo habitual es que quienes rechazan a extranjeros lo hagan en la mayoría de dimensiones —como demuestra el hecho de que en varios tuits coincidieran más de una—. De estas dimensiones, las más frecuentes fueron la hostilidad manifiesta hacia el colectivo, la amenaza que se percibe que estas personas suponen para la seguridad y la consideración de estos colectivos como una carga económica. Se entiende que, al ser estas las asociaciones más visibles y las que más se expresan en redes sociales, las actuaciones más útiles, si se quiere contrarrestar este rechazo, deberían ir orientadas en esta dirección. Dado que esta parte de la investigación tiene un carácter preliminar, también debemos hacer hincapié en la necesidad de seguir ampliando el conoci-

miento sobre las causas que hay detrás del rechazo al extranjero para poder enfrentarlo de la manera más adecuada, tanto en los estadios iniciales que se analizan en este texto como en las formas de rechazo más dañinas que se manifiestan a través de discursos o crímenes de odio.

Por último, es importante destacar que nuestros resultados pueden servir para explorar nuevos mecanismos de detección del discurso de odio analizando el rechazo verbal en la red, especialmente en el entorno hispanoparlante. Este trabajo apunta hacia Twitter y, en consecuencia, hacia otros medios sociales, como fuentes de información valiosas para el análisis de la opinión pública y de las actitudes ciudadanas, en los casos en los que el estudio a través de encuestas resulta en gran medida limitado. En esta línea, una de las mayores aportaciones de este trabajo es la generación de un corpus de ejemplos de aceptación y rechazo hacia migrantes y refugiados[4] con el que se pueda entrenar modelos con técnicas de aprendizaje automático supervisado, para permitir la detección automática y a gran escala de dichos discursos.

Limitaciones e investigación futura

Este estudio, aunque utiliza información extraída de Twitter, todavía utiliza una muestra limitada. Tanto la finitud del número de tuits como su extracción en dos contextos temporales concretos —entre abril y mayo de 2018 y en julio de 2019— impiden una extrapolación absoluta, pues, como se ha podido observar, las muestras de rechazo y de aceptación fluctúan notablemente en función de fenómenos puntuales que influyen en la opinión pública. No obstante, junto a la lectura de la situación en estos dos momentos, el trabajo también incorpora un análisis preliminar de los aspectos negativos asociados con el rechazo y de la relación entre el tipo de tuit y la expresión de rechazos.

Al mismo tiempo, dado que no se geolocalizaron los contenidos, es imposible realizar una lectura más detallada por países, algo que, por otra parte, no es sencilla en los análisis que utilizan Twitter como fuente, dado que muy poca gente hace pública su localización —según Gaffney y Puschman (2014) tan solo un 1% del tráfico en Twitter se geoetiqueta— y gran cantidad de estos datos están protegidos por el medio. Así, los datos obtenidos son preliminares y generales, pero ofrecen una herramienta para que futuros trabajos puedan comparar el rechazo al extranjero en distintos contextos hispanohablantes. Por otro lado, también se excluyeron imágenes, hipervínculos y tuits compuestos únicamente por emoticonos, limitando el estudio al análisis textual.

La utilización de medios sociales en el análisis científico implica algunas limitaciones por las dificultades técnicas de las interfaces. Un ejemplo de estas dificultades, que también menciona Chaudhry (2015), es el API de Twitter, que a las cuentas gratuitas de desarrollador, como la utilizada en el estudio, solo le ofrece acceso al 1% de todos los tuits publicados descargados desde el *stream* o limita los accesos al *rest* a los últimos siete días. No obstante, la cantidad de datos es enormemente mayor de lo que cualquier otro modelo analógico de recolección de datos permitiría. Asimismo, dada la composición de usuarios de redes sociales, y de Twitter en concreto, es imposible generalizar las conclusiones de los estudios que utilicen estas plataformas como fuente de datos, ya que ciertos grupos sociodemográficos, especialmente los grupos de mayor edad, apenas están representados.

[4] Dicho corpus está disponible con acceso abierto en el enlace: https://github.com/carlosarcila/rejection

Bibliografía

Arcila-Calderón, Carlos; Ortega-Mohedano, Félix; Jiménez-Amores, Javier y Trullenque, Sofía (2017). «Análisis supervisado de sentimientos políticos en español: clasificación en tiempo real de tuits basada en aprendizaje automático». *El profesional de la información,* 26(5): 973-982. doi: 10.3145/epi.2017.

Bakir, Vian y McStay, Andrew (2018). «Fake News and the Economy of Emotions». *Digital Journalism*, 6(2): 154-175. doi: 10.1080/21670811.2017.1345645

Bartlett, Jamie; Reffin, Jeremy; Rumbale, Noelle y Williamson, Sarah (2014). *Anti-social media*. London: Demos.

Ben-David, Anat y Matamoros-Fernández, Ariadna (2016). «Hate Speech and Covert Discrimination on Social Media: Monitoring the Facebook Pages of Extreme-right Political parties in Spain». *International Journal of Communication,* 10: 1167-1193. Disponible en: https://ijoc.org/index.php/ijoc/article/view/3697/1585, acceso el 17 de diciembre de 2019.

Berger, Peter L. y Luckmann, Thomas (1966). *The Social Construction of Reality*. New York: Random House.

Billig, Michael (2002). «Henri Tajfel's "Cognitive Aspects of Prejudice" and the Psychology of Bigotry». *British Journal of Social Psychology,* 41(2): 171-188. doi: 10.1348/014466602760060165

Bourhis, Richard V. y Dayan, Joelle (2004). «Acculturation Orientations towards Israeli Arabs and Jewish Immigrants in Israel». *International Journal of Psychology,* 39(2): 118-131. doi: 10.1080/00207590344000358

Brewer, Marilynn B. (1999). «The Psychology of Prejudice: Ingroup Love and Outgroup Hate?». *Journal of Social Issues*, 55(3): 429-444. doi: 10.1111/0022-4537.00126

Brown, Rupert (2000). «Social Identity Theory: Past Achievements, Current Problems and Future Challenges». *European Journal of Social Psychology,* 30(6): 745-778. doi: 10.1002/1099-0992(200011/12)30:6<745::AID-EJSP24>3.0.CO;2-O

Burnap, Pete y Williams, Matthew L. (2015). «Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making». *Policy & Internet*, 7(2): 223-242. doi: 10.1002/poi3.85

Cea D'Ancona, María Ángeles (2009). «La compleja detección del racismo y la xenofobia a través de encuesta. Un paso adelante en su medición». *Revista Española de Investigaciones Sociológicas (REIS),* 125: 13-45. Disponible en: http://reis.cis.es/REIS/PDF/REIS_125_011231144723167.pdf, acceso el 28 de agosto de 2019.

Chaudhry, Irfan (2015). «Hashtagging Hate: Using Twitter to Track Racism Online». *First Monday*, 20(2). doi: 10.5210/fm.v20i2.5450

Contrada, Richard J.; Gary, Melvin L.; Coups, Elliot; Egeth, Jill D.; Sewell, Andrea; Ewell, Kevin; Goyal, Tanya M. y Chasse, Valerie (2001). «Measures of Ethnicity-Related Stress: Psychometric Properties, Ethnic Group Differences, and Associations with Well-being». *Journal of Applied Social Psychology,* 31: 1775-1820. doi:10.1111/j.1559-1816.2001.tb00205.x

Cortina, Adela (2017). *Aporofobia, el rechazo al pobre: un desafío para la democracia*. Madrid: Paidós.

Crandall, Christian S.; Miller, Jason M. y White, Mark H. (2018). «Changing Norms Following the 2016 US Presidential Election: The Trump Effect on Prejudice». *Social Psychological and Personality Science*, 9(2): 186-192. doi: 10.1177/1948550617750735

Davidson, Thomas; Warmsley, Dana; Macy, Michael y Weber, Ingmar (2017). «Automated Hate Speech Detection and the Problem of Offensive Language». En: *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)*. Disponible en: http://sdl.soc.cornell.edu/img/publication_pdf/hatespeechdetection.pdf, acceso el 17 de diciembre de 2019.

Díez Nicolás, Juan (2009). «Construcción de un índice de Xenofobia-Racismo». *Revista del Ministerio de Trabajo e Inmigración*, 80: 21-38. Disponible en: http://www.mitramiss.gob.es/es/publica/pub_electronicas/destacadas/revista/numeros/80/est01.pdf, acceso el 20 de agosto de 2019.

European Commission against Racism and Intolerance (2016). *ECRI General Policy Recommendation N.º 15 on Combating Hate Speech*. Strasbourg: European Council.

Fox, Jesse; Cruz, Carlos y Lee, Ji Young (2015). «Perpetuating Online Sexism Offline: Anonymity, Interactivity, and the Effects of Sexist Hashtags on Social Media». *Computers in Human Behavior*, 52: 436-442. doi: 10.1016/j.chb.2015.06.024

Gaffney, Devin y Puschmann, Cornelius (2014). «Data collection on Twitter». En: Bruns, A.; Weller, K.; Burgess, J.; Mahrt, M. y Puschmann, C. (eds.). *Twitter and Society*. New York: Peter Lang.

Gallego, Mar; Gualda, Estrella y Rebollo, Carolina (2017). «Women and Refugees in Twitter: Rhetorics on Abuse, Vulnerability and Violence from a Gender Perspective». *Journal of Mediterranean Knowledge*, 2(1): 37-58. doi: 10.26409/2017JMK2.1.03

Gorham, Bradley W. (2006). «News Media's Relationship with Stereotyping: The Linguistic Intergroup Bias in Response to Crime News». *Journal of Communication*, 56(2): 289-308. doi: 10.1111/j.1460-2466.2006.00020.x

Gualda, Estrella y Rebollo, Carolina (2016). «The Refugee Crisis on Twitter: A Diversity of Discourses at a European Crossroads». *Journal of Spatial and Organizational Dynamics*, 4(3): 199-212. Disponible en: https://www.jsod-cieo.net/journal/index.php/jsod/article/view/72, acceso el 20 de agosto de 2019.

Gualda, Estrella; Borrero, Juan Diego y Cañada, José Carpio (2015). «La "Spanish Revolution" en Twitter (2): Redes de hashtags y actores individuales y colectivos respecto a los desahucios en España». *Revista Hispana para el Análisis de Redes Sociales, REDES*, 26(1): 1-22. doi: 10.5565/rev/redes.535

Kalyanam, Janani; Quezada, Mauricio; Poblete, Barbara y Lanckriet, Gerts (2016). «Prediction and Characterization of High-ActivityEvents in Social Media Triggered by Real-World News». *PloS one*, 11(12): e0166694. doi: 10.1371/journal.pone.0166694

Kelleher, John D.; MacNamee, Brian y D'Arcy, Aoife (2015). *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. London: MIT Press.

Kreis, Ramona (2017). «#refugeesnotwelcome: Anti-refugee Discourse on Twitter». *Discourse & Communication*, 11(5): 498-514. doi: 10.1177/1750481317714121

Maass, Anne; Salvi, Daniela; Arcuri, Luciano y Semin, Gün R. (1989). «Language Use in Intergroup Contexts: The Linguistic Intergroup Bias». *Journal of Personality and Social Psychology*, 57(6): 981-993. doi: 10.1037/0022-3514.57.6.981

Muller, Karsten y Schwarz, Carlo (2018). «Fanning the Flames of Hate: Social Media and Hate Crime». *SSRN*. doi: 10.2139/ssrn.3082972

Murray, Kate E. y Marx, David A. (2013). «Attitudes toward Unauthorized Immigrants, Authorized Immigrants, and Refugees». *Cultural Diversity and Ethnic Minority Psychology,* 19(3): 332-341. doi: 10.1037/a0030812

Naciones Unidas (1951). *Convención sobre el Estatuto de los Refugiados.* Disponible en: https://eacnur.org/files/convencion_de_ginebra_de_1951_sobre_el_estatuto_de_los_refugiados.pdf, acceso el 20 de agosto de 2019.

Neuendorf, Kimberly A. (2002). *The Content Analysis Guidebook*. Thousand Oaks, California: Sage.

O'Rourke, Kevin H. y Sinnott, Richard (2006). «The Determinants of Individual Attitudes towards Immigration». *European Journal of Political Economy*, 22(4): 838-861. doi: 10.1016/j.ejpoleco.2005.10.005

Oxford Dictionaries (2016). *Word of the Year 2016 is…* Disponible en: https://en.oxforddictionaries.com/word-of-the-year/word-of-the-year-2016, acceso el 26 de agosto de 2019.

Park, Sung-Yeon (2012). «Mediated Intergroup contact: Concept Explication, Synthesis, and Application». *Mass Communication and Society*, 15(1): 136-159. doi: 10.1080/15205436.2011.558804

Peherson, Samuel; Brown, Rupert y Zagefka, Hanna (2011). «When Does National Identification Lead to the Rejection of Immigrants? Cross-sectional and Longitudinal Evidence for the Role of Essentialist in Group Definitions». *British Journal of Social Psychology,* 48(1): 61-76. doi: 10.1348/014466608X288827

Rebollo, Carolina y Gualda, Estrella (2017). «La situación internacional de las personas refugiadas y su imagen en Twitter. Un reto para la intervención desde el trabajo social». *Documentos de Trabajo Social*, 59: 190-207. Disponible en: https://dialnet.unirioja.es/servlet/articulo?codigo=6588971, acceso el 28 de agosto de 2019.

Schäfer, Claudia y Schadauer, Andreas (2019). «Online Fake News, Hateful Posts Against Refugees, and a Surge in Xenophobia and Hate Crimes in Austria». En: Dell'Orto, G. y Wetzstein, I. (eds.). *Refugee News, Refugee Politics: Journalism, Public Opinion and Policymaking in Europe*. Oxford: Routledge.

Verkuyten, Maykel (2014). *Identity and Cultural Diversity: What Social Psychology Can Teach Us*. Hove: Routledge.

Verkuyten, Maykel y Brug, Peary (2004). «Multiculturalism and Group Status: The Role of Ethnic

Identification, Group Essentialism and Protestant Ethic». *European Journal of Social Psychology,* 34(6): 647-661. doi: 10.1002/ejsp.222

Verkuyten, Maykel; Mepham, Kieran y Kros, Mathijs (2018). «Public Attitudes towards Support for Migrants: The Importance of Perceived Voluntary and Involuntary Migration». *Ethnic and Racial Studies*, 41(5): 901-918. doi: 10.1080/01419870.2017.1367021

Whitley Jr., Bernard E. y Kite, Mary E. (2016). *Psychology of Prejudice and Discrimination*. New York: Routledge.

Wike, Richard; Stokes, Bruke y Simmons, Katie (2016). *Europeans Fear Wave of Refugees Will Mean More Terrorism, Fewer Jobs*. Disponible en: https://immigrazione.it/docs/2016/Pew-Research-Center-July-11-2016.pdf, acceso el 28 de agosto de 2019.

RECEPCIÓN: 13/03/2019
REVISIÓN: 10/07/2019
APROBACIÓN: 11/02/2020