social sciences

MDPI

Article

Topic Modeling and Characterization of Hate Speech against Immigrants on Twitter around the Emergence of a Far-Right Party in Spain

Carlos Arcila Calderón *, Gonzalo de la Vega and David Blanco Herrero

Department of Sociology and Communication, University of Salamanca, Campus Miguel de Unamuno, Paseo Francisco Tomás y Valiente, s/n, 37007 Salamanca, Spain; gdelaveg@outlook.com (G.d.l.V.); david.blanco.herrero@usal.es (D.B.H.)

* Correspondence: carcila@usal.es; Tel.: +34-626-256-855

Received: 18 September 2020; Accepted: 21 October 2020; Published: 23 October 2020

Abstract: In this paper, we sought to model and characterize hate speech against immigrants on Twitter in Spain around the appearance of the far-right party Vox. More than 240,000 tweets that included the term 'Vox' between November 2018 and April 2019 were automatically collected and analyzed. Only 1% of the sample included hate speech expressions. Within this subsample of 1977 messages, we found offenses (56%), incitements to hate (42%), and violent speech (2%). The most frequent terms used were classified into five categories: Spain, Immigration, Government, Islam, and Insults. The most common features were foul language, false or doubtful information, irony, distasteful expressions, humiliation or contempt, physical or psychological threats, and incitement to violence. Using unsupervised topic modeling, we found that the four underlying topics (control of illegal immigration, economic assistance for immigrants, consequences of illegal immigration, and Spain as an arrival point for African immigrants and Islamist terrorism) were similar to those in the discourse of Vox. We conclude that the hate speech against immigrants produced *around* Vox, and not necessarily *by* Vox, followed the general patterns of this type of speech detected in previous works, including Islamophobia, offensive language more often than violent language, and the refusal to offer public assistance to these collectives.

Keywords: Twitter; hate speech; topic modelling; natural language processing; VOX

1. Introduction

The use of social media is growing among societies independently from conditions, such as age, gender, or origin, and this has made it easier for people around the globe to share any kind of message, including audio-visual content, breaking the monopole of mass media in producing and spreading content. Among these social media platforms, Twitter has become a public space for political conversation (Moyá and Herrera 2015), allowing contact between politicians and citizens and becoming an essential player in the construction of the public agenda (McCombs and Shawn 1972).

Beyond the democratic and freeing effects of these media, this possibility for every citizen to express any opinion or feeling, making it public and accessible for almost every other person, has some risks associated, as offense and polemics can spread and reach a greater public. One of the clearest examples of this is the spread of online hate speech, as social media has allowed a faster and broader spread, which has led to greater visibility and, therefore, a greater impact and magnified effects. Through social media, a message that has not been verified in its production can be replicated and shared by any kind of account (Cueva 2012), which can be dangerous and harmful. Hate speech is especially dangerous as a trigger of potential hate crime (Müller and Schwarz 2018) and also as a

crime itself. However, there exist different approaches to defining hate speech, differentiating it from offensive language (Davidson et al. 2017), or generally speaking about "dangerous speech." This concept was established by Susan Benesch, who proposed that the efforts to reduce hate speech can be less effective due to the lack of clarity in its definition (Benesch 2014).

The increase in online hate speech (Bartlett et al. 2014) has taken place in a global context in which migratory movements are growing, as well as anti-immigration discourse, which makes hate speech against immigrants predominant. In the case of Spain, although the arrival of far-right parties to the institutions took place later than in other European or Western countries, the political party Vox is now the third force in the National Parliament and plays a relevant role in many regional and local governments, after entering in a regional parliament for the first time in the Andalusian Elections, held in December 2018.

In this context, the main goal of this paper was to use computational methods detect and analyze the dimensions of hate speech toward immigrants on Twitter within the frame of the Spanish socio-political scenario after the appearance of a strong far-right and anti-immigration party. The work seeks to fill the empirical gap existing regarding the features of hate speech against immigrants in the Spanish setting, to discover what characteristics define this kind of hateful speech in order to contribute to its identification and to the definition of a still unclear concept. This is intended to aid current efforts, such as the European project Preventing Hate against Refugees and Migrants (PHARM) or the Stop-Hate project, developed at the University of Salamanca of Spain, to identify and detect hate speech online.

For this research, we automatically collected data from social media using Twitter's application programming interface (API), and, using natural language processing techniques and topic modeling, we extracted valuable information from a large volume of unstructured data and tested the use of these two novel techniques in the field.

2. Theoretical Framework

The present work used, as a basis, studies that have already attempted to model or automatically detect hate speech online using big data or machine learning techniques, such as the study of Mondal et al. (2017), which used sentence structure to automatically detect hate speech in social media, or that of Schmidt and Wiegand (2017), which used natural language processing for the detection of hate speech. More focused on immigration and in Spanish, but without a focus on discovering hate speech, Gallego et al. (2017) used a semi-automatic coding method with a dictionary to analyze 862,999 tweets that included the word "refugee" in Twitter messages to study the discourse regarding women and refugees around the Crisis of Refugees of the Mediterranean.

The work of Ben-David and Matamoros-Fernandez (2016) monitored the activity of seven far-right pages in Facebook between 2009 and 2013 to analyze the presence of hate speech; they compared the frequency of certain words and their simultaneous occurrence to find patterns from which they could model underlying topics. This work showed how hate and discrimination on Facebook was being introduced within the legitimate boundaries of the Spanish political discourse. This study finished its analysis in 2013, prior to the appearance of Vox (at the end of that year) and long before its discourse became relevant in the Andalusian elections in December 2018. Our work follows a similar goal, but with a difference—that Vox is not a marginal party any longer, but one of the main actors in the Spanish political scene.

Another difference is that the focus is not on Facebook groups, but on Twitter content, not only due to the significance of this social medium for political communication (D'heer and Verdegem 2014) but also because it has been proven a relevant and fruitful line of study for hateful content. In this vein, we highlight the efforts of Burnap and Williams (2015), who developed a model to detect violent and hateful content on Twitter with the goal of monitoring the reaction of the public to specific events that could be potentially controversial phenomena. In this interaction between online and offline events, a very relevant project is Umati, led by Susan Benesch, who showed how the surge of online hate speech was influenced by real events.

Following these observations, other researchers explored social media to discover hate speech and its interactions with real events. Olteanu et al. (2018) characterized messages after extremist events along four dimensions (stance, target, severity, and framing) to detect hate speech, observing how some jihadist terrorist attacks that took place in Western countries had an impact and influenced hate speech towards Arabic and Muslim collectives, systematically increasing the number of messages promoting hate speech and violence towards these groups.

Evolvi (2018) also approached the spread of Islamophobia on Twitter in connection to the Brexit process with a qualitative analysis of Islamophobic tweets collected after the Brexit referendum in 2016. In the opposite direction, the aforementioned Müller and Schwarz (2018) observed how online hate speech in social media influenced and could even help in predicting real events of violence or hate crimes against refugees and migrants by modeling together anti-refugee attacks and the frequency of anti-refugee messaging on social media based on the Facebook page of the far-right party *Alternative für Deutschland* in Germany.

2.1. Defining Hate Speech

To approach this topic, it is important to consider freedom of speech, as the clash of its limits with xenophobic, extremist, or racist discourses goes beyond the law and becomes a discussion of political philosophy (Alcácer 2015). However, here, we will use a communicational perspective, understanding that speech has a social projection, as it aims at one or more audiences, whether they are broad or small, and it can be legally restricted if it harms or limits the freedom of others (Cueva 2012). Recommendation No. 15 of the European Commission against Racism and Intolerance (2016) defines hate speech as promoting hate, humiliation, or underestimation in any form against a person or a group, 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 Framework Decision 2008/913/JHA of 28 November 2008 of the European Council (2008) defines hate speech as "publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, color, religion, descent, or national or ethnic origin."

More specifically, online hate speech has been defined or characterized in several previous works. We highlight Miró's (2016) taxonomy of hateful content online, including any violent expression, which will be one of the bases for our model. Waseem and Hovy (2016) also defined a set of features to reliably decide whether a message shows hate speech or not, as it tends to be complex for different people to agree on homogeneous criteria in this matter. Other relevant work includes that of Warner and Hirschberg (2012), which addressed numerous problems when determining whether a message should be considered hate speech, as the use of specific words or expressions might not necessarily mean an expression of hate.

Davidson et al. (2017) observed that the combination of offensive language and hate speech might lead to errors in the distinction between both concepts. However, Mondal et al. (2017) considered that any post motivated entirely or partially by the author's prejudice toward an aspect of a group should be seen as hate speech and, in order to overcome the previously mentioned problems, they designed a different system, based on the detection of the complete structure of sentences. Finally, we also followed other indicators, including obscene language or other distasteful expressions, because Schmidt and Wiegand (2017) defended that this type of language is central to the detection of hate messages when combined with other features. Although less relevant for our text, there have been approaches to study anti-immigration discourse on social media from the perspective of qualitative techniques, such as the one of Kreis (2017), guided by critical discourse studies.

In Spain, it was not until 2015 when the *Disposición final sexta* of the LO/1/2015 CP of the Gobierno de España (2015) adopted this European Framework Decision to the domestic legislation. It is article 510 of the Penal Code of Spain that punishes hate speech in any form for "racist, anti-Semitic or other reason referring to ideology, religion or beliefs, family situation, belonging to a race, ethnicity or nation, national origin, sex, sexual orientation or identity, gender, illness or disability." However, the

Law in this country does not protect a group as a general rule without further reason; instead, it demands specific conditions that lead a group to a situation of vulnerability. According to the Ministry of the Interior of Spain, there are eight motivations or prejudices that lead to the existence of vulnerable groups: racism/xenophobia, sexual orientation or identity, religious praxis or beliefs, disability, gender, antisemitism, and aporophobia. They all might interact with each other, aggravating some situations.

The object of this study was hate speech against immigrants, given that, according to the figures of the Statistic System of Criminality (SEC) of the Ministry of the Interior of Spain, racism and xenophobia are the reasons behind the largest number of cases of hate speech and hate crime in the last years. Online hate speech against immigrants in Spain has been already tackled by Valdez-Apolo et al. (2019), who showed that negative messages are predominant when talking about migrants and refugees and also observed that immigrants are usually framed as a threat.

Arcila-Calderón et al. (2020) studied the presence of rejection of immigrants in Twitter messages with a mixed manual and automated content analysis of tweets. Gualda and Rebollo (2016, p. 208) used a semi-automatic coding method with a dictionary to study the discourse regarding refugees in Twitter in different European nations, including Spain, and observed how messages can have xenophobic connotations and how "sometimes these discourses are supported by politicians, such as Donald Trump or other organizations in Europe". In a broader sense, this article will also complement those works that tackle the attitudes toward immigrants, such as Murray and Marx (2013) and Verkuyten et al. (2018).

Together with the works analyzing hate speech or rejection against immigrants, both in Spain and internationally, it is relevant to study the connection of this type of discourse with nationalism, as Peherson et al. (2011) did using a cross-sectional and longitudinal study. More specifically, in recent years, parallel to the rise of far-right populist parties, scholars have paid great attention to the role that anti-immigration nationalism has for these parties. In this field, Lubbers and Coenders (2017) studied how nationalism connects with radical right voting. In the Spanish setting, the analysis focused on the multiple reasons for the absence (until recently) of a populist radical right (PPR) party. Alonso and Rovira Kaltwasser (2015) mentioned the cleavage structure of the country and the strategy of competition of the mainstream right and the electoral system, and Casals (2000) added the lack of organization and the archaic political culture, far from the influence of European far-right parties. In a similar line, Morales et al. (2015) focused on the ambivalent approach to immigration by the main Spanish political parties.

These works are, however, now outdated: first, because Teruel (2017) observed that concern about hate speech and the conducts built upon prejudices and stereotypes in Spain have grown, and more specifically, because the arrival of the far-right political party Vox has altered the political scenario in Spain, bringing also the topic of immigration—and, particularly, anti-immigration—into a more visible position of the political agenda as stated by Castromil et al. (2020) after analyzing the political program, the use of Twitter, and the political debates of different parties. Arango et al. (2019) also defend that the Spanish exception within the European context ended with the arrival of Vox to the Andalusian Regional Parliament in December 2018 and to the national one in April 2019, making anti-immigration a more relevant matter of the public discourse.

Other studies have investigated the individual-level determinants of vote choice that explain the rise of this party (Turnbull-Dugarte 2019; Turnbull-Dugarte et al. 2020), and Ferreira (2019) conducted a qualitative content analysis of the political programs and discourses based on the causal chain method, confirming anti-immigration nationalism as a differential aspect of this party. However, there exist no studies similar to the one of Ben-David and Matamoros-Fernandez (2016), focusing not only on the party, but on the discussion around it, since the arrival of Vox to the highest democratic institutions of Spain. That is why we proposed to answer the following research question:

RQ1: What are the features of hate speech toward immigrants around the emergence of a far-right party, such as Vox?

The interest of this question is not only the study of hate speech in a particular context but also, given that the anti-immigration discourse surrounding Vox is one of the most defining sources of

hate speech against immigrants in Spain, to model the topics underlying this type of discourse. That is why we proposed to answer:

> **RQ2:** What are the underlying topics of hate speech toward immigrants around the emergence of a far-right party, such as Vox?

Both questions attempt to go further than the observation of the amount or the visibility of hate speech, focusing on the features and characteristics and also attempting to comprehend what topics are addressed when this discourse is used. This is a key aspect to understanding what is behind this discourse and how to address it, complementing some preliminary efforts in this sense, such as the study of Arcila-Calderón et al. (2020), regarding what negatives aspects were associated with the rejection of migrants and refugees.

In order to answer those two questions, we will use, as a reference, the taxonomy of hate speech and violent communication online, built by Miró (2016) in his monitoring of hate speech in the frame of the jihadist attack toward the French magazine *Charlie Hebdo*, as well as the parameters that differentiate hate speech from offensive or vulgar language of Davidson et al. (2017). More specifically, we sought to use computational methods to discover the topics behind this discourse, as well as the most frequent and relevant terms in Spanish that allow for the detection of hateful content in digital media, creating a database that can be used in future projects.

3. Method

3.1. Sample and Procedure

Using Python's library *Tweepy*, we accessed Twitter's application programming interface (API) to collect tweets related to Vox, taking advantage of this interface (Ong et al. 2015) to obtain the unstructured data for the study. Specifically, we collected tweets both from API Rest and from API Streaming. The first collects tweets using one or more keywords or hashtags from the historic flow of messages of the last ten days, whereas the second collects all messages produced in real time with one or more keywords or hashtags.

We retrieved all tweets in Spanish (lang=es) and excluding retweets (exclude="retweets") that contained the word 'Vox' in the track—that is, any field of a tweet, including the name of the account that produces it, the text of the tweet, the links shared, etc.—from 25 November 2018 until 28 April 2019. The initial date of collection was close to the regional elections in the region of Andalucía (2 December), in which Vox obtained their first seats in a regional parliament of Spain. The collection period also included 15 February 2019, when the announcement of new elections in Spain by Pedro Sánchez, President of the Government, took place. The final day of collection was the day of the National Parliament Elections. In total, 244,095 messages were collected for a period of six months.

The tweets were collected in JSON format, which allows running filters, such as date, language, geographical location, name of the user, etc. However, only the text of the tweet was analyzed, given that the analysis was intended to study the features and topics of the message, not a time or geographical distribution. The messages were produced by official accounts of the party, or by media or citizens naming the word 'Vox' in their content, generating tweets of multiple and diverse topics. As explained in the next paragraphs, we later filtered this enormous number of messages by manually removing those not containing hate speech toward immigrants in one of the three ways defined by Miró (2016), so that we could use a subsample for manual and computational analysis.

The whole procedure, which will be detailed in the next section paying attention to each step, was as follows: a sample of 244,095 tweets that included the term 'Vox' somewhere in the track of the tweet was automatically collected, and then a manual classification allowed us to obtain a subsample of 1977 tweets that included expressions of hate against immigrants. That subsample was afterward classified in three groups following Miro's classification (2016), and a manual exploratory analysis was conducted to observe the features of language in those tweets. Then, two computational methods were used to investigate the features and topics of the discourse of the subsample of hateful contents: first, natural language processing was used to identify the most frequent terms in each of the three

groups in which hate speech was classified; and, second, latent Dirichlet allocation (LDA) topic modeling was used to discover what topics underlay the whole subsample of hateful messages.

3.2. Measures

3.2.1. Hate Speech towards Immigrants

Based on the contemporary discussion regarding hateful content online as explained in Section 2.1, we considered any message in Twitter that directly or indirectly damaged the image of individuals or groups based on their condition of immigrant, refugee, asylum seeker, or displaced as *hate speech toward immigrants* (Miró 2016; Waseem and Hovy 2016; Warner and Hirschberg 2012; Davidson et al. 2017; Mondal et al. 2017; Schmidt and Wiegand 2017).

Although these works offer guidelines for the detection of hate speech, Schmidt and Wiegand (2017) presented their concerns regarding the problems of reliability and the difficulty of consensus due to the lack of unanimity in the definition of hate speech. With this in mind, in the present study, the following criteria were established to determine whether a message contained hate speech against immigrants:

They had to be messages showing contempt or hate toward the collective of immigrants and, in particular, those expressions using pejorative terms against immigrants, as well as those demanding or justifying a restriction of the rights of immigrants. Messages that were considered offensive or hurtful against feelings or beliefs of the collective were also included, together with those containing insults or grave offenses against a particular person or group of the immigrant collective. It was also considered hate speech when there was an association of individual victims or the whole collective with crimes or illicit behaviors when this association was intentionally false or not concerned with the truth of the accusation. Finally, the direct or indirect promotion of physical violence against one well-known member or the whole immigrant collective, as well as expressions of defense, justification, trivialization, or glorification of that violence.

To obtain the inter-coder reliability of this variable, two independent judges were trained to analyze a random sub-sample of 24,225 messages (~10% of the total sample). According to the degree of agreement between both coders, we used Krippendorff's alpha to test the reliability, as this is the most recommended measure (Hayes and Krippendorff 2007). We obtained a value of 0.88, which is over the acceptable minimum of 0.7 (Neuendorf 2002).

3.2.2. Types of Hate Speech against Immigrants

Hateful messages were classified according to the types proposed by Miró (2016): (a) direct incitement or glorification of violence; (b) incitement to discrimination, hate, or restriction of rights; and (c) offenses against feelings. These types allowed a classification of hate speech at three levels of danger. According to the pyramidal shape (Figure 1) it was expected that a majority of hate messages would belong to the category of offenses against feelings, and the smallest proportion would be those directly inciting or glorifying violence. We also conducted an inter-coder reliability test, obtaining a Krippendorff's alpha of 0.78, which was adequate for the study.

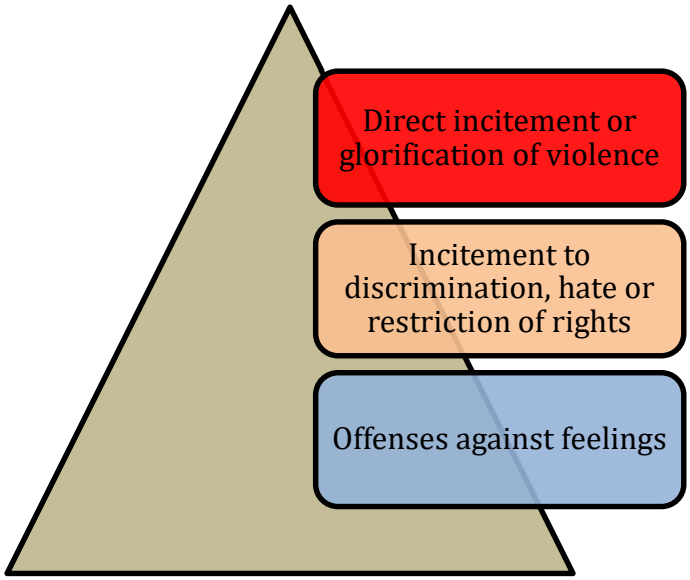

Figure 1. Theoretical types of hate speech from Miró (2016).

3.2.3. Frequency Distribution

We applied basic natural language processing (NLP) techniques to obtain the frequency distribution in hateful tweets against immigrants. NLP is a branch of computational sciences that is combined with applied linguistics and attempts to make a machine process and "comprehend" what a text in a particular language means. Essentially, NLP seeks to convert a text in a set of structured data that describe its meaning and the topics it mentions (Collobert et al. 2011). NLP-based technologies are growing in presence and play a relevant role in the current multi-linguistic societies (Bird et al. 2009). The programming language Python offers a broad library that includes components for graphic programming, numeric processing, and web connectivity. For the present study, it was essential to previously install the *Numpy* library, which adds stronger support for vectors and matrixes, as well as the *Natural Language Toolkit* (*NLTK*), which defines an infrastructure that allows the development of NLP programs (Bird et al. 2009).

This type of linguistic analysis based on the distribution of the terms of a text was used as previous step for the identification of underlying topics after a filtering process and counting of the most frequent terms, both in the sub-sample of messages containing hate speech against immigrants and in the different categories that are part of it. Knowing the most frequent words offers valuable information that will be useful for interpreting the results of the topic modeling.

The first step for properly conducting NLP techniques was the identification of tokens, the basic units, typically simple words or sentences, in which a text can be deconstructed for its following analysis. A token cannot be deconstructed into smaller parts; thus, in computational methods, a token is considered an atom (Webster and Kit 1992). For this process, called tokenization, we used the NLTK library in Python, together with the module that tokenizes the text at the level of *words*,[1] and we indicated the location of the text that should be analyzed.

The following step was the removal of Stop Words, that is, words that do not give relevant information and that are very common, such as articles or prepositions. It is vital to also remove punctuation, accents, and web links to avoid the repetition of terms and to obtain homogenous final results. Finally, we were able to observe the most repeated terms and their distribution and decide on the number of topics that we want to obtain.

[1] Other approaches such as TFIDF or N-gram for text representation were not considered in this study.

3.2.4. Topic Modeling (Latent Dirichlet Allocation—LDA)

To detect underlying topics in hateful tweets against immigrants, we applied unsupervised machine learning in the form of topic modeling using the latent Dirichlet allocation algorithm (Blei et al. 2003). LDA is the most commonly used algorithm for topic modeling (Grimmer and Stewart 2013) and is frequently used to identify the topics in a set of documents (Ramage et al. 2009), allowing the automatic modeling of a large amount of data and to visualize this data as a combination of topics (Canini et al. 2009). According to Keller et al. (2020), this technique "is a form of automated content analysis that infers latent thematic structures called topics within documents in a 'bottom-up' approach." This approach allows the inference of topics from texts—in this case, tweets—without prior knowledge or an extensive manual annotation. The topics are detected by discovering patterns in the presence of clusters of co-occurring words across documents (Jacobi et al. 2015).

This method tends to be used in larger texts, such as articles from newspapers (Keller et al. 2020) or abstracts of journals (Zou 2018), but there are some arguments that push us to employ it in shorter messages, such as tweets: the longer extension of tweets since 2017 of 280 characters instead of 140, the relevance of Twitter in the construction of public discourse in the present—particularly around populist and radical parties, and the interest to test this technique in this medium in Spain, discovering whether it can be applied in larger studies. In this case, the application of the model to the sub-sample will allow us to dig into the connection of the terms that build hate speech against immigrants in Spanish and, this way, obtain groups of words that can be used to deduct the topic behind it.

For this task, beside NLTK, it was also necessary to import the following libraries of Python's version 3.7: *pandas* (data analysis), *seaborn* (visualization), *gensim* (topic modeling), and *pyLDAvis* (visualization of topics). After importing all the requested libraries and modules and selecting the text we want to model, the first step was to remove punctuation signs and double spaces and convert all text into lowercase. The first model conducted here offers a *naïve model* that does not discriminate Stop Words; a list of these words can be also imported and applied so that we can achieve a more adequate modeling.

For this, it is advisable to use coherence measures of the topics; by calibrating the level of semantic similarity among words with a high score inside of a topic (Stevens et al. 2012), a more precise model can be achieved. For that goal, the *Umass coherence index* of the text we want to model must be calculated based on the number of topics and the number of terms inside of each; the further from 0 the obtained value is, the higher the coherence level is. For example, the lower coherence level that *naïve models* have is explained because of the presence of Stop Words, irrelevant terms that introduce noise in the text, reducing the coherence of the topics.

Finally, the *pyLDAvis* library will allow us to print a map for visually exploring the final result of the modeling in a quick and simple way. This library also contains a tool that adjusts the level of λ (lambda) to increase or decrease the frequency ratio of a selected topic.

4. Results

4.1. Distribution of the Sample and Sub-Sample

The total amount of collected messages was 244,095, of which 1977 were classified as hate speech against immigrants in the phase of manual tagging according to the previously mentioned rules (Figure 2 shows the proportion of messages that built the sub-sample inside of the total sample). The sub-sample built with those 1977 tweets in which hate speech was detected was divided into three categories depending the level of danger of the hateful discourse included in the text (Figure 3 shows the distribution of the sub-sample in the three previously specified categories of hate speech). The biggest group of this three, with 1026 tweets (56% of the total) was for the least serious type of hate, the type that included offenses against the sensibility of others; the second group, with messages that incite discrimination, hate, or the restriction of rights, had 757 messages (42%); and the most dangerous type of hate expression, the type promoting violence, was present in 42 messages—2% of the total.

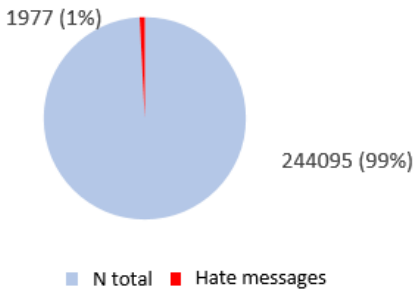

Figure 2. Proportion of hate messages inside of the total sample.

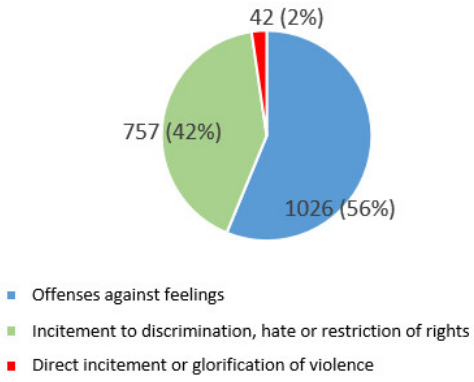

Figure 3. Types of detected hate speech.

4.2. Features of Hate Speech

Using exploratory analysis, we identified features of the different dimensions of hate speech, such as foul language, false or doubtful information, irony, distasteful expression, humiliation or contempt, physical or psychological threats, or incitement to violence.

- *Foul language*: Dishonest or obscene words are used. As previously mentioned, the presence of this type of language in a message does not necessarily mean that there is hate against immigrants, and it is the co-occurrence—the proximity between the two terms—that determines the presence of hate. For example, when the obscene word is applied to the collective of immigrants, the presence of hate speech is more common, as in: "Putos inmigrantes. La gente que quieren que se mueran de hambre. esos merecen la pena. Viva VOX."[2]
- *Incitement to violence*: This type of message invites others to conduct violent acts against a specific person or collective. This dimension is linked to physical and psychological threats (see next point), but it is based on an abstract call rather than a direct threat. In the next example, we can see how the emitter calls for the expulsion of immigrants in a violent way using despising terms in a threatening way: "A día de hoy solo Vox, pide acabar con la inmigración ilegal. Españoles hay que votarlos para limpiar España de estos salvajes. Y los que no los voteis, disfrutar de lo votado."[3]

[2] Published 30 November 2018 at 19:22. In English: "Fucking immigrants. People want them to starve to death. They deserve the punishment. Long live VOX."

[3] Published 25 November 2018 at 21:45. In English: "To this day only VOX demands to end illegal immigration. Spaniards, we have to vote to clean Spain from these savages. And those who do not vote for them, enjoy your vote."

- *Physical or psychological threat*: These messages go against the physical and psychological integrity of the victims (Miró 2016), and, unlike the previous group, the threat is more immediate and leads more directly to the completion of the violent act. It must be highlighted that violence might not be the end but rather a means, as in the following example: "A ver si sale vox y echamos a todas estas putas ratas de el pais."[4]
- *Humiliation and contempt*: Underestimation of a person or collective and rejection of them based on their inherent condition. For Schmidt and Wiegand (2017), this dimension is sometimes given by the context, and so it might be hard to detect. See, for example: "Pues a mí me han convencido los de #vox, por fin gente como #shakira, #Messi, #Griezmann, #benzema … Dejarán de quitarle el trabajo a nuestros hijos españoles! A su casa!! #VOXalNatural #Politica #EleccionesYa."[5]
- *Distasteful expressions*: Eschatological, vulgar, or disgusting expressions are used. This type of expressions can vary depending the geographical location of the emitter and the addition of a negative charge to the message. In the next example, it can be seen how these expressions highlight the hate against a specific group of immigrants: "Fuera los Moros!, … tomar por culo su religion! a si de claro!, … que se vaya la coño norte de Africa!"[6]
- *Irony*: This is the hardest to detect as the hate is expressed in a subtler way. In the next example, we can see how sarcasm is used to criticize and to say the opposite to the literal meaning of the words: "Pero los crucifijos fuera de las escuelas… y @vox_es son muy malos. Los siguientes hombres de paz van a ser los del ISIS… no?"[7]
- *False or doubtful information*: These messages include unconfirmed generalizations, stereotypes, or false affirmations regarding a collective. In the context of hate speech content, it is common that these messages attempt to create social alarm regarding something that attacks the internal culture or beliefs with external impositions. For example: "Exacto. Sin embargo, nos están destruyendo nuestras creencias, nuestras tradiciones e imponiéndonos islamismo radical y "culturas" ajenas a nosotros y que faltan el respeto."[8]

4.3. Frequency Distribution

We obtained the most common words used in hateful comments against immigrants in Spanish in order to characterize this kind of speech. After adding all the Stop Words[9] and removing the terms that share the lexeme, we obtained the following list of the 20 most representative terms of content containing hate speech against immigrants (n = 1977):

('inmigrantes', 540), ('españa', 383), ('pais', 264), ('ilegales', 251), ('inmigracion', 237), ('españoles', 160), ('mujeres', 134), ('musulmanes', 92), ('europa', 84), ('partido', 81), ('moros', 73), ('islam', 73), ('ayudas', 71), ('extranjeros', 69), ('votar', 64), ('gobierno', 60), ('pp', 55), ('expulsar', 55), ('trabajo', 51), and ('negro', 49).[10]

---

[4] Published 15 December 2018 at 23:29. In English: "Let's hope Vox wins and we remove all these fucking rats from the country."

[5] Published 12 December 2018 at 22:27. In English: "I have been convinced by Vox, finally people like #shakira, #Messi, #Griezmann, #benzema… will stop taking the jobs from our Spanish children! To their house! #Voxasitis #Politics #ElectionsNow."

[6] Published 11 December 2018 at 21:08. In English: "Out with the moors! Fuck off with their religion! Clear as day! Fuck off to the fucking North of Africa!"

[7] Published 13 December 2018 at 23:08. In English: "But the crucifixes out of the schools… and @vox_es are very bad. The next men of peace will be the ones of ISIS, right?"

[8] Published 25 November 2018 at 21:05. In English: "Exactly. However, they are destroying our beliefs, our traditions and forcing us into a radical Islamism and "cultures" that are alien to us and that are disrespectful."

[9] Even when some terms might be not special or meaningful for the analysis, we did not include in the Stop Word list reference terms such as "Inmigrantes" (immigrants) or "España" (Spain). We consider that far from being redundant they might offer better results in the co-occurrence analysis.

[10] In English: "immigrants, spain, country, illegals, immigration, spanish [masculine plural], women, muslims, Europe, party, moors, islam, benefits, foreigners, vote, government, pp [People's Party of Spain], expel, work, black."

We conducted the same approach for each of the sub-categories. The 20 most frequent terms in the group of *Offenses against the feelings* (n = 1026) were:

('inmigrantes', 307), ('españa', 181), ('inmigracion', 180), ('ilegales', 170), ('pais', 111), ('españoles', 70), ('partido', 48), ('pp', 41), ('musulmanes', 40), ('gobierno', 35), ('andalucia', 34), ('mujeres', 34), ('negro', 34), ('europa', 34), ('expulsar', 34), ('extranjeros', 32), ('votar', 31), ('islam', 31), ('programa', 29), ('melilla', 29).[11]

The representative words in the group *Incitement of discrimination, hate, or restriction of rights* (n = 757) were:

('inmigrantes', 179), ('españa', 173), ('pais', 130), ('españoles', 77), ('ilegales', 77), ('mujeres', 70) ('musulmanes', 64), ('ayudas', 55), ('moros', 44), ('europa', 44), ('inmigracion', 39), ('islam', 39), ('religion', 28), ('extranjeros', 26), ('musulmana', 24), ('mierda', 21), ('ley', 21), ('derechos', 21), ('cultura', 20), ('machistas', 20).[12]

In the case of *Direct incitement or glorification of violence* (n = 42) we obtained:

('españa', 11), ('pais', 9), ('inmigrantes', 6), ('putos', 6), ('mierda', 6), ('culo', 6), ('moros', 5), ('coño', 5), ('musulmanes', 5), ('puto', 4), ('puta', 4), ('niñas', 3), ('hijos', 3), ('basura', 3), ('violadores', 3), ('inmigracion', 2), ('españoles', 2), ('limpiar', 2), ('delincuentes', 2), ('gentuza', 2).[13]

To obtain a better understanding of this analysis, we manually grouped all the terms detected into five topics selected for the study (see Table 1).

Table 1. Manual grouping of the terms by topic.

Spain	Immigration	Government	Islam	Insults
España	inmigrantes	programa	religión	mierda
españoles	inmigrante	ayudas	cultura	putos
español	Ilegales	trabajo	musulmana	puto
país	Ilegal	votar	musulmán	negro
Andalucía	extranjeros	problema	musulmanes	basura
Melilla	expulsar	ley	moros	gentuza
	países	Europa	mujeres	violadores
		derechos	mujer	delincuentes
		partido		coño
		pp		culo
		psoe		machistas

4.4. Topic Modeling

After obtaining the distribution of frequencies for the general hateful tweets and for each of the specific three categories, we conducted topic modeling to automatically detect the underlying topics in this kind of speech. To determine an adequate number of topics, we measured the level of coherence—the farther from 0, the better—and we compared several models with 15 words for each topic and decided that the adequate number of topics was five, as this number offered a coherence value of −7.8415, the farthest from 0, as, from 6 topics onward, it started decreasing. After recursively removing the Stop Words, we detected and labeled the next topics:

Topic 1: Lack of control of illegal immigration by the State. This refers to an alleged negligence from the Spanish government and the public institutions to control illegal immigration, especially the lack of strong measures to stop it from the traditional parties (Figure 4):

---

[11] In English: "immigrants, spain, immigration, illegals, country, spanish [masculine plural], party, pp, muslims, government, andalusia, women, black, europe, expel, foreigners, vote, islam, program, melilla."

[12] In English: "immigrants, spain, country, spanish [masculine plural], illegals, women, muslims, benefits, moors, europe, immigration, islam, religion, foreigners, muslim, shit, law, rights, culture, male chauvinists."

[13] In English: "spain, country, immigrants, fucking [masculine plural], shit, ass, moors, cunt, muslims, fucking [masculine singular], whore/fucking [feminine singular], girls, sons, garbage, rapists, immigration, spanish [masculine plural], clean, offenders, riffraff."

1 ('0.021*"inmigrantes" + 0.011*"españa" + 0.008*"inmigracion" + 0.007*"musulmanes" + 0.006*"pais" + 0.005*"mujeres" + 0.005*"españoles" + 0.004*"partido" + 0.004*"ilegales" + 0.004*"europa" + 0.003*"paises" + 0.003*"inmigrante" + 0.003*"psoe" + 0.003*"pp" + 0.003*"ilegal"').[14]

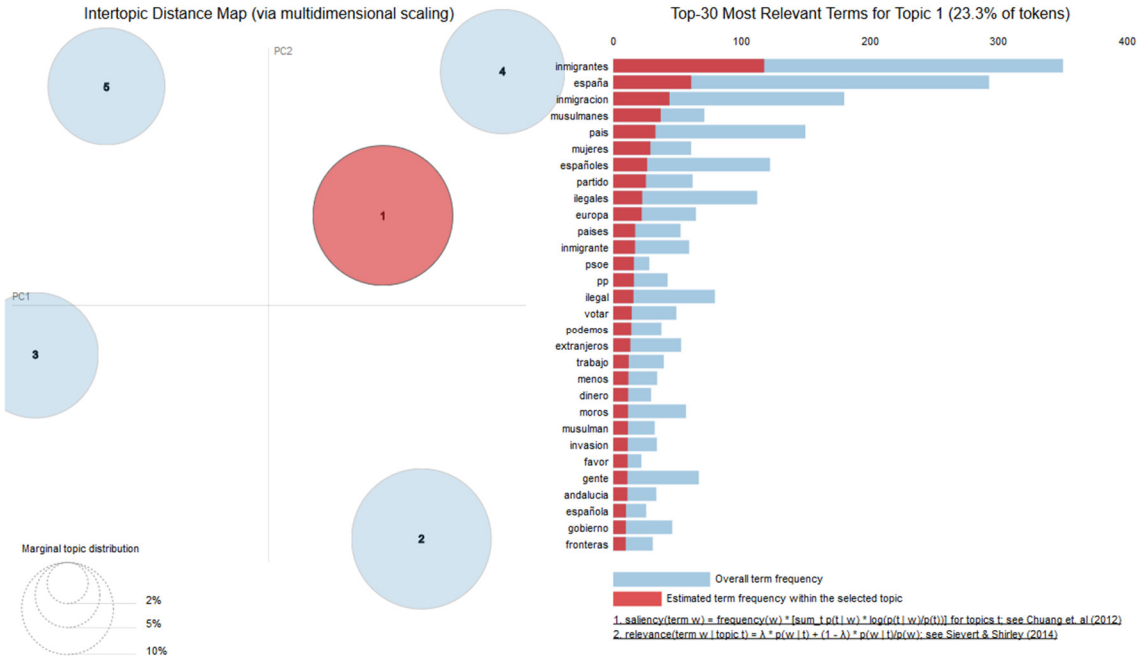

Figure 4. Interactive map of topic 1, λ = 1.

Topic 2: Economic assistance to immigrants. This approaches the alleged frauds of immigrants when obtaining economic support from public institutions, as well as their advantages compared to Spanish citizens (Figure 5):

2 ('0.020*"españa" + 0.010*"pais" + 0.009*"inmigrantes" + 0.009*"inmigracion" + 0.007*"españoles" + 0.006*"ilegal" + 0.005*"moros" + 0.004*"ayudas" + 0.003*"millones" + 0.003*"mujer" + 0.003*"ilegales" + 0.003*"musulmana" + 0.003*"expulsar" + 0.003*"inmigrante" + 0.003*"paises"').[15]

[14] In English: "immigrants, spain, immigration, muslims, country, women, spanish [masculine plural], party, illegals, europe, countries, immigrant, psoe [Spanish Socialist Party], pp, illegal."

[15] In English: "Spain, country, immigrants, immigration, spanish [masculine plural], illegal, moors, benefits, millions, woman, illegals, muslim [feminine singular], expel, immigrant, countries."

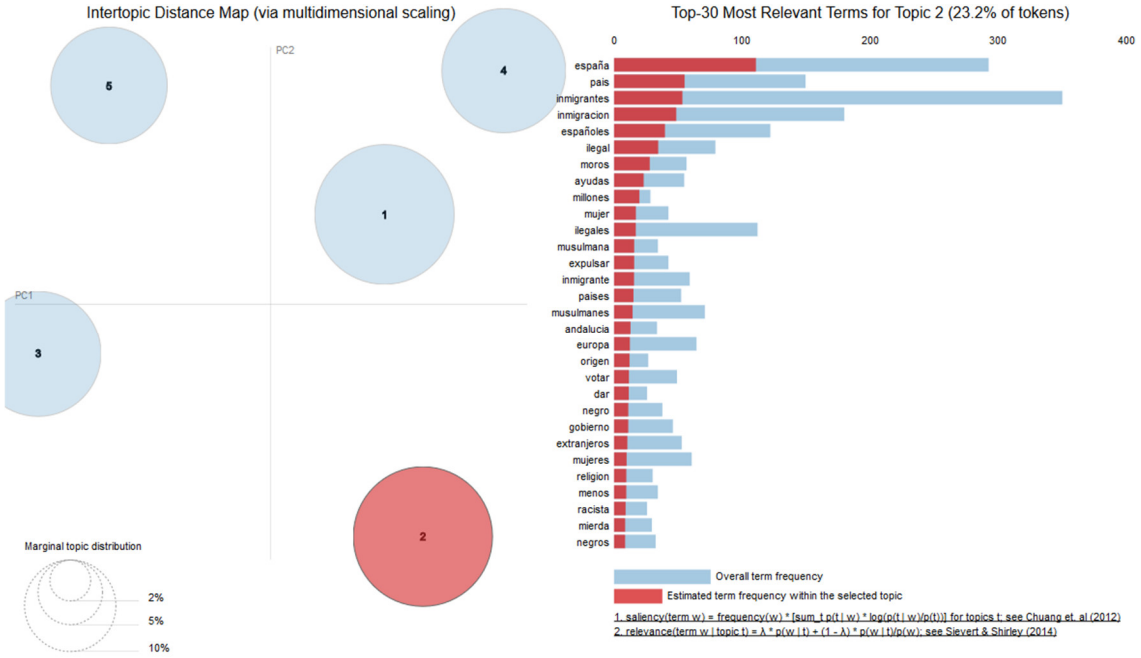

Figure 5. Interactive map of topic 2, λ = 1.

Topic 3: Consequences of illegal immigration. The focus is on the negative consequences that illegal immigration has for the Spanish population (Figure 6):

3 ('0.012*"inmigrantes" + 0.008*"españa" + 0.006*"ilegales" + 0.006*"españoles" + 0.005*"gente" + 0.005*"pais" + 0.004*"inmigracion" + 0.004*"islam" + 0.003*"inmigrante" + 0.003*"europa" + 0.003*"extranjeros" + 0.003*"mujer" + 0.003*"marroquies" + 0.003*"problema" + 0.003*"ayudas"').[16]

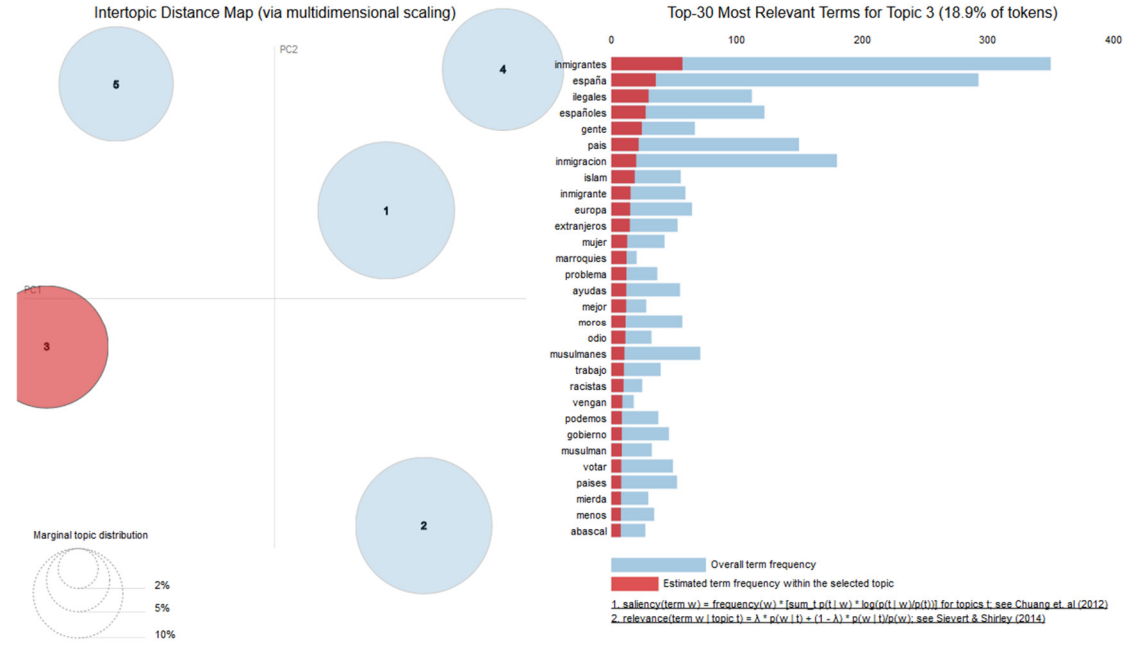

Figure 6. Interactive map of topic 3, λ = 1.

[16] In English: "immigrants, spain, illegals, spanish [masculine plural], people, country, immigration, islam, immigrant, Europe, foreigners, woman, Moroccan [masculine plural], problem, benefits."

Topic 4: Spain as the entrance of African immigrants to Europe. This focuses primarily on the arrival of African immigrants to the Southern border of Spain, especially in the city of Melilla in the Northern coast of Africa (Figure 7):

4 (1, '0.015*"inmigrantes" + 0.013*"españa" + 0.008*"ilegales" + 0.008*"inmigracion" + 0.005*"partido" + 0.004*"pais" + 0.004*"gente" + 0.003*"izquierda" + 0.003*"ilegal" + 0.003*"melilla" + 0.003*"fronteras" + 0.003*"europa" + 0.002*"español" + 0.002*"españoles" + 0.002*"invasion"').[17]

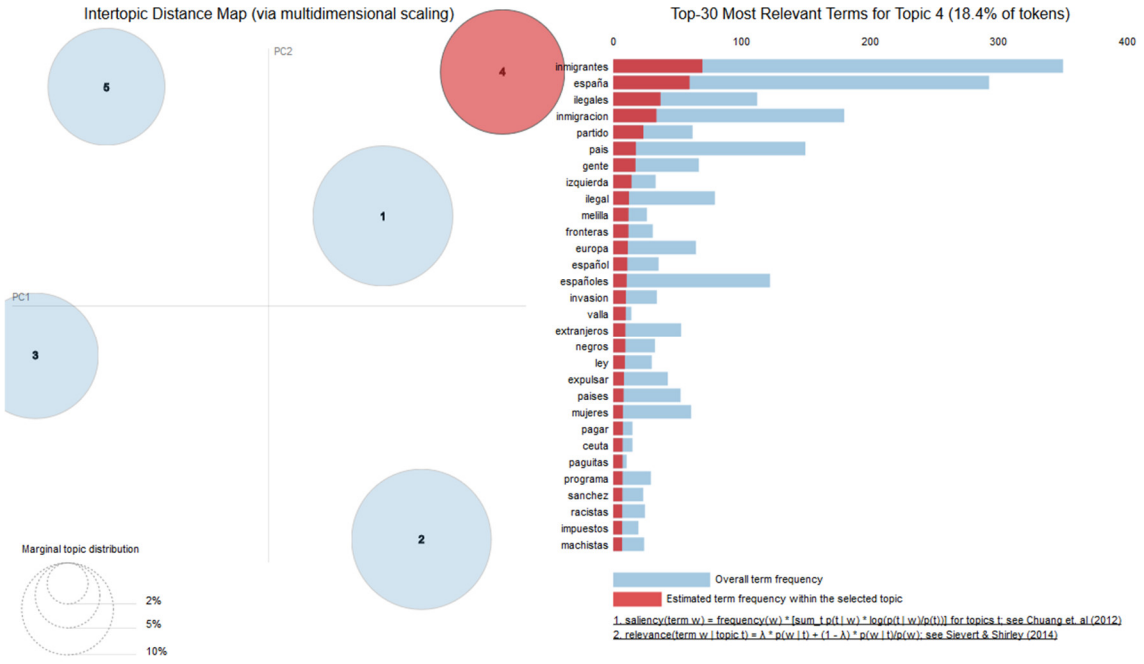

Figure 7. Interactive map of topic 4, λ = 1.

Topic 5: Islamist terrorism. This topic focuses on the association of terrorism with Islam or with Muslim immigrants or people from Muslim countries or backgrounds (Figure 8):

5 (0, '0.013*"inmigrantes" + 0.009*"inmigracion" + 0.007*"españa" + 0.006*"pais" + 0.005*"españoles" + 0.004*"islam" + 0.004*"miedo" + 0.003*"islamistas" + 0.003*"expulsion" + 0.002*"gobierno" + 0.002*"mujeres" + 0.002*"ilegal" + 0.002*"familia" + 0.002*"problema" + 0.002*"negro"').[18]

[17] In English: "immigrants, spain, illegals, immigration, party, country, people, left, illegal, melilla, borders, Europe, Spanish [masculine singular], spanish [masculine plural], invasion."

[18] In English: "immigrants, immigration, spain, country, spanish [masculine plural], islam, fear, islamists, expel, government, women, illegal, family, problem, black."

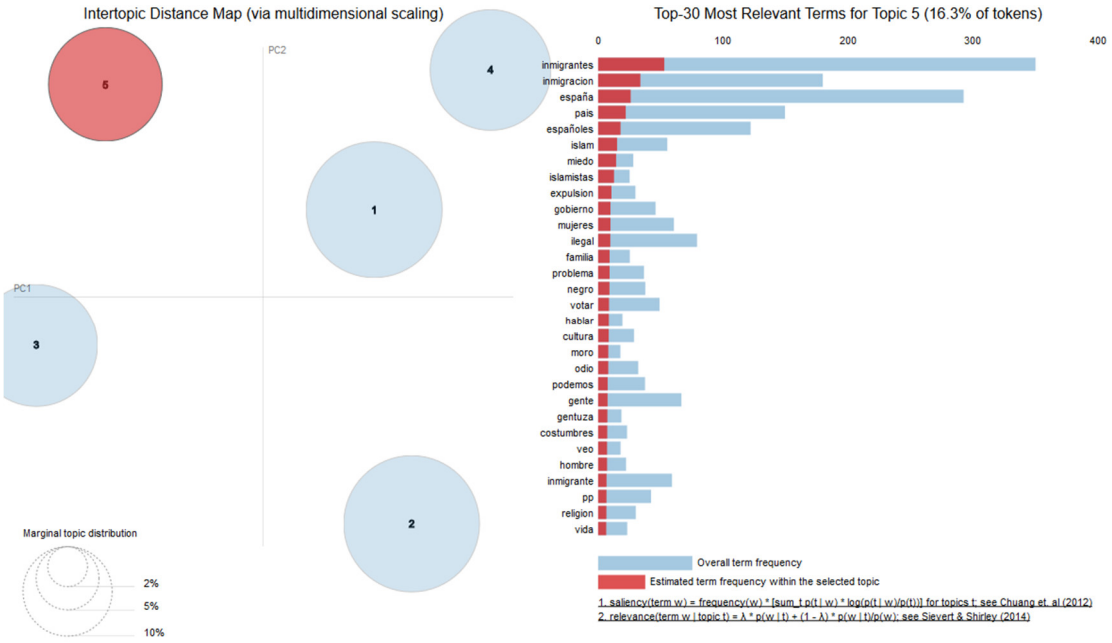

Figure 8. Interactive map of topic 5, λ = 1.

5. Discussion of Results and Conclusions

As expected, the distribution of the different types of hate speech against immigrants follows a pyramid shape, from the most common offenses against the feelings, with 1026 tweets and 56% of this subsample; through the incitement of discrimination, hate, and the restriction of rights, with 757 messages and 42% of the subsample; finally to the direct incitement or glorification of violence, with 42 messages and only 2% of the subsample. This agrees with the distribution observed in previous studies, such as Miró (2016), in which only 2% of the original sample collected using hashtags after the Charlie Hebdo attacks in Paris in 2015 included hate speech expressions, and from those, the distribution also followed a pyramid, from the least to the most harmful.

In total, we observed that 1% of the discussion surrounding Vox during the six months analyzed contained hate speech against immigrants. This does not indicate that 1% of the messages produced by Vox or its supporters included some form of hate speech against immigrants. First, because not all the conversation *around* Vox is produced solely *by* Vox and its followers, but also by the media and by citizens talking about the party; second, because there can be expressions of hate speech against immigrants that are not produced by Vox or its followers, but in which the party is mentioned for different reasons; and third, because Vox or its followers might produce hate speech against immigrants without mentioning the name of the party in the text of the tweet or in other fields of the track of the tweet. At the same time, those 244,095 messages collected with the term 'Vox' also included other topics discussed around Vox discourse—the economy, social protection, fight against criminality, etc.—without relation to immigration or without including hateful expressions.

Similarly, although in a very small amount, other messages might have included the term 'Vox' in another context; for example, typos using the word "voz,"[19] the Latin expression *vox populi,* or a brand of dictionaries also named Vox. As a conclusion of this aspect, we can affirm that around 1% of the conversation that mentioned Vox included hate speech against immigrants, but not that 1% of the discourse produced by the party, its leaders, and its followers does; future studies will be needed to determine if that proportion is bigger or smaller.

This approach of the study also showed the need to continue researching different forms of hate speech around this far-right party, as a stronger relevance of other topics, such as Catalonia's independency or the management of the COVID-19 crisis by the Government, could be expected, which might modify the proportion of hate speech against immigrants in Vox's discussion on Twitter.

[19] In English: "voice."

Thus, the volume of hate speech against a particular group may have increased or decreased, particularly, because it has been observed that online hate speech can be intensified on social media after high-impact events, such as a terrorist attack or news with the presence of disadvantaged or denigrated groups of the population (Awan and Zempi 2015; Awan 2014). In this line, a complementary study of the spikes and troughs in hate speech prevalence coinciding with the events presented along the period of study would have been interesting, and this will be developed in future works; however, given the focus of the study in the features and topics of this discourse rather than in the amount or evolution, it was not included at this time.

The features and topics of this type of hateful discourse tend to be stable over time, which suggests that the answers of our research questions will remain valid for a long period of time, even as hate speech against immigrants might gain or lose presence. Therefore, in general terms, the study confirmed what previous studies have already pointed out, that anti-immigration nationalist discourses are closely associated to this far-right party, as observed by Ferreira (2019) and that the features of hateful contents go from (the more or less harmless) offensive language, as studied by Davidson et al. (2017), to public and direct incitements to violence that already constitute a crime by themselves.

The second research question regarding the topics underlying these discourses was answered using a computational approach. The results of the distribution of frequencies and the five topics that were modeled demonstrated the presence of the main elements that build and explain hate speech against immigrants in the discourse around Vox on Twitter. These topics, which include the expulsion of illegal immigrants from Spain, the removal of public benefits for immigrants, and the "invasion" coming from the North of Africa, relate closely to the anti-immigration discourse of Vox, and are consistent with the broader analysis of hate speech or rejection toward immigration on Twitter in Spain.

This study, combining NLP and LDA topic modeling, offers a complete analysis of the discourse of hatred against immigrants, going further than the studies focused on just one technique, such as that of Schmidt and Wiegand (2017), which focused on NLP. Due to this approach, we observed that Muslim immigrants were frequently the victims of hate speech, as the presence of a topic focused on them supports, which agrees with the Islamophobic condition of this party (Gould 2019). The consideration of this religion as barbaric or misogynist is a common way to justify the rejection and hate toward this collective.

We also commonly discovered distasteful or obscene expressions in much of the content, as these expressions were used to stigmatize or hurt the feelings of immigrants. We also found expressions demanding discrimination or the restriction of rights, such as the reduction or removal of public support for these collectives. This connection of immigrants with their cost to public money was also found in the study that Arcila-Calderón et al. (2020) conducted regarding the reasons for the rejection of migrants and refugees in Twitter messages in Spain. Additionally, the use of lies or untrue generalizations or stereotypes regarding the public benefits that immigrants receive was common, something that previous works also observed, connecting this party with disinformation campaigns (Hernández Conde and Fernández García 2019).

As a limitation of the study, the poor grammar of some of the messages made the study of frequencies and the topic modeling less accurate. The analysis of only one social media platform, Twitter in this case, although justified and common to many previous studies, makes it impossible to generalize the observations to the offline construction of hate speech against immigrants and the public discussion around Vox outside of Twitter; however, the relevance of this medium and the discoveries about the topics and terms that define hate speech against immigrants makes it useful for designing techniques to analyze and combat this form of hateful discourse.

One of the main contributions of the article was precisely to dig into the features and topics behind hate speech, complementing previous works in the Spanish context (Miró 2016; Gallego et al. 2017) and to apply novel computational methods that have not been broadly applied toward this goal. Regarding the use of these techniques, particularly topic modeling, even when this method is frequently used in larger texts, the co-occurrence of words offers better results in larger contents. The

topics detected in tweets offered good exploratory results to discover the characteristics of hate speech against immigrants taking place in the Twitter content surrounding Vox.

Despite the use of these two novel techniques, other methods that could have been complementary, such as an n-gram analysis to study the interrelation and overlaps of the most frequent terms, are not present in this article. This technique, already proven useful by previous works (Burnap and Williams 2015), will be applied in the future so that a more detailed effort can be conducted.

Author Contributions: C.A.C. and G.d.l.V. designed and planned the study. G.d.l.V. collected and analyzed the data. D.B.H. produced the theoretical background and organized and adapted the contents to the form of a manuscript. C.A.C. coordinated the whole process. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the European Union's Right Equality and Citizenship Programme (2014-2020). REC-RRAC-RACI-AG-2019 Grant Agreement 875217.

Conflicts of Interest: The authors report no conflict of interest.

References

Alcácer, Rafael. 2015. Víctimas y disidentes. El "discurso del odio" en EE.UU. y Europa. *Revista Española De Derecho Constitucional* 35: 45–86.

Alonso, Sonia, and Cristóbal Rovira Kaltwasser. 2015. Spain: No country for the populist radical right? *South European Society and Politics* 20: 21-45. doi:10.1080/13608746.2014.985448.

Arango, Joaquín Ramón Mahía, David Moya, and Elena Sánchez-Montijano. 2019. *Inmigración, Elecciones Y Comportamiento Político. Anuario CIDOB De La Inmigración*. Edited by En Joaquín Arango, Ramón Mahía, David Moya and Elena Sánchez-Montijano. Barcelona: CIDOB, pp. 16–30. doi:10.24241/AnuarioCIDOBInmi.2019.16.

Arcila-Calderón, Carlos, David Blanco-Herrero, and María Belén Valdez-Apolo. 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 (REIS)* 172: 21–40. doi:10.5477/cis/reis.172.19.

Awan, Imran. 2014. Islamophobia on Twitter: A Typology of Online Hate against Muslims on Social Media. *Policy & Internet* 6: 133–50. doi:10.1002/1944-2866.POI364.

Awan, Imran, and Irene Zempi 2015. The affinity between online and offline anti-Muslim hate crime: Dynamics and impacts. *Aggression and Violent Behaviour* 27: 1–8. doi:10.1016/j.avb.2016.02.001.

Bartlett, Jamie, Jeremy Reffin, Noelle Rumball, and Sarah Williamson. 2014. *Anti-Social Media*. London: DEMOS.

Ben-David, Anat, and Ariadna Matamoros-Fernandez. 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: 1167–93.

Benesch, Susan. 2014. Countering Dangerous Speech: New Ideas for Genocide Prevention. Working Paper, US Holocaust Memorial Museum, Washington, DC, USA.

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. Sebastopol: O'Reilly Media.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3: 993–1022.

Burnap, Pete, and Matthew L. Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet* 7: 223–42. doi:10.1002/poi3.85.

Canini, Kevin, Lei Shi, and Thomas Griffiths. 2009. Online Inference of Topics with Latent Dirichlet Allocation. In *Proceedings of the Artificial Intelligence and Statistics*. Clearwater Beach: JMLR, pp. 65–72.

Casals, Xavier. 2000. La ultraderecha española: una presencia ausente (1975–1999). *Historia y política: Ideas, procesos y movimientos sociales* 3: 147–74.

Castromil, Antón R., Raquel Rodríguez-Díaz, and Paula Garrigós. 2020. La agenda política en las elecciones de abril de 2019 en España: programas electorales, visibilidad en Twitter y debates electorales. *El Profesional de la Información* 29: e290217. doi:10.3145/epi.2020.mar.17.

Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12: 2493–537.

Cueva, Ricardo. 2012. El «discurso del odio» y su prohibición; hate speech and its ban; hate speech and its ban. *DOXA. Cuadernos de Filosofía del Derecho* 35: 437–55. doi:10.14198/DOXA2012.35.18.

Davidson, Thomas, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media.* Palo Alto: AAAI, pp. 512–15.

D'heer, Evelien, and Verdegem, Pieter. 2014. Conversations about the elections on Twitter: Towards a structural understanding of Twitter's relation with the political and the media field. *European Journal of Communication* 29: 720–34. doi:10.1177/0267323114544866

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

European Council. 2008. Framework Decision 2008/913/JHA of 28 November 2008 on Combating Certain Forms and Expressions of Racism and Xenophobia by Means of Criminal law. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32008F0913 (accessed on 21 October 2020)

Evolvi, Giulia. 2018. Hate in a tweet: Exploring Internet-Based Islamophobic Discourses. *Religions* 9: 307. doi:10.3390/rel9100307.

Ferreira, Carles. 2019. Vox como representante de la derecha radical en España: un estudio sobre su ideología. *Revista Española de Ciencia Política* 51: 73–98. doi:10.21308/recp.51.03.

Gallego, Mar, Estrella Gualda, and Carolina Rebollo. 2017. Women and Refugees in Twitter: Rhetorics on Abuse, Vulnerability and Violence from a Gender Perspective. *Journal of Mediterranean Knowledge* 2: 37–58.

Gobierno de España. 2015. Ley Orgánica 1/2015, de 30 de marzo, por la que se modifica el Código Penal. *Boletín Oficial del Estado* 77: 27061–176.

Gould, Robert 2019. Vox España and Alternative für Deutschland: Propagating the Crisis of National Identity. *Genealogy* 3: 64. doi:10.3390/genealogy3040064.

Grimmer, Justin, and Brandon M. Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* 21: 267–97. doi:10.1093/pan/mps028.

Gualda, Estrella, and Carolina Rebollo. 2016. The Refugee Crisis on Twitter: A Diversity of Discourses At A European Crossroads. *Journal of Spatial and Organizational Dynamics* 4: 199–212.

Hayes, Andrew F., and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures* 1: 77–89. doi:10.1080/19312450709336664.

Hernández Conde, Macarena, and Manuel Fernández García. 2019. Partidos emergentes de la ultraderecha: ¿fake news, fake outsiders? Vox y la web Caso Aislado en las elecciones andaluzas de 2018. *Teknokultura* 16: 33–53. doi:10.5209/TEKN.63113.

Jacobi, Carina, Wouter van Atteveldt, and Kasper Welbers. 2015. Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism* 4: 89–106. doi:10.1080/21670811.2015.1093271.

Keller, Tobias R., Valerie Hase, Jagadish Thaker, Daniela Mahl, and Mike S. Schäfer. 2020. News Media Coverage of Climate Change in India 1997–2016: Using Automated Content Analysis to Assess Themes and Topics. *Environmental Communication* 14: 219-235. doi:10.1080/17524032.2019.1643383

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

Lubbers, Marcel, and Marcel Coenders. 2017. Nationalistic attitudes and voting for the radical right in Europe. *European Union Politics* 18: 98–118. doi:10.1177/1465116516678932

McCombs, Maxwell E., and Donald L. Shawn. 1972. The agenda-setting function of the mass media. *Public Opinion Quarterly* 36: 176–87. doi:10.1086/267990.

Miró, Fernando. 2016. Taxonomía de la comunicación violenta y el discurso del odio en internet. *IDP: Revista De Internet, Derecho Y Política* 22: 82–107.

Mondal, Mainack, Leandro Araújo Silva, and Fabrício Benevenuto. 2017. A Measurement Study of Hate Speech in Social Media. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*. New York: ACM, pp. 85–94.

Morales, Laura, Sergi Pardos-Prado, and Virginia Ros. 2015. Issue emergence and the dynamics of electoral competition around immigration in Spain. *Acta Politica* 50: 461–85. doi:10.1057/ap.2014.33

Moyá, Miguel, and Susana Herrera. 2015. Cómo puede contribuir twitter a una comunicación política más avanzada. *Arbor* 191: a257. doi:10.3989/arbor.2015.774n4012.

Müller, Karsten, and Carlo Schwarz. 2018. Fanning the Flames of Hate: Social Media and Hate Crime. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3082972 (accessed on 10 October 2020).

Murray, Kate E. and David A. Marx. 2013. Attitudes toward unauthorized immigrants, authorized immigrants, and refugees. *Cultural Diversity and Ethnic Minority Psychology* 19: 332–41. doi:10.1037/a0030812.

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

Ong, Shyue Ping, Shreyas Cholia, Anubhav Jain, Miriam Brafman, Dand Gunter, Gerbrand Ceder, and Kristin A. Persson. 2015. The materials application programming interface (API): A simple, flexible and efficient API for materials data based on Representational state transfer (REST) principles. *Computational Materials Science* 97: 209–15. doi:10.1016/j.commatsci.2014.10.037.

Olteanu, Alexandra, Carlos Castillo, Jeremy Boy, and Kush R. Varshney. 2018. The effect of extremist violence on hateful speech online. In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media*. Palo Alto: AAAI, pp. 221–30.

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

Ramage, Daniel, David Hall, Ramesh Nallapati, and Christopher D. Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1*. Stroudsburg: ACL, pp. 248–56.

Schmidt, Anna, and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. New York: ACM, pp. 1–10.

Stevens, Keith, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Stroudsburg: ACL, pp. 952–61.

Teruel, Germán M. 2017. El discurso del odio como límite a la libertad de expresión en el marco del convenio europeo. *Revista De Derecho Constitucional Europeo* 27: 81–108

Turnbull-Dugarte, Stuart J. 2019. Explaining the end of Spanish exceptionalism and electoral support for Vox. *Research & Politics* 6. doi:10.1177/2053168019851680.

Turnbull-Dugarte, Stuart J., José Rama, and Andrés Santana. 2020. The Baskerville's dog suddenly started barking: Voting for VOX in the 2019 Spanish general elections. *Political Research Exchange* 2. doi:10.1080/2474736X.2020.1781543.

Valdez-Apolo, María Belén, Carlos Arcila-Calderón, and Javier J. Amores. 2019. El discurso del odio hacia migrantes y refugiados a través del tono y los marcos de los mensajes en Twitter. *RAEIC, Revista de la Asociación Española de Investigación de la Comunicación* 6: 361–84. doi:10.24137/raeic.6.12.2RAEIC.

Verkuyten, Maykel, Kieran Mepham, and Mathijs Kros. 2018. Public attitudes towards support for migrants: the importance of perceived voluntary and involuntary migration. *Ethnic and Racial Studies* 41: 901–18. doi:10.1080/01419870.2017.1367021.

Warner, William, and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*. Stroudsburg: ACL, pp. 19–26.

Waseem, Zeerak, and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*. Stroudsburg: ACL, pp. 88–93.

Webster, Jonathan J., and Chunyu Kit. 1992. Tokenization as the initial phase in NLP. In *COLING 1992: The 15th International Conference on Computational Linguistics*. Nantes: ACL, pp. 1106–10.

Zou, Chen. 2018. Analyzing research trends on drug safety using topic modeling. *Expert Opinion on Drug Safety* 17: 629–36. doi:10.1080/14740338.2018.1458838.