CUADERNOS.INFO Nº 49
Versión electrónica: ISSN 0719-367x
http://www.cuadernos.info
https://doi.org/10.7764/cdi.49.27817

Received: 11-26-2020 / Accepted: 04-19-2021

Detecting ideological hatred on Twitter. Development and evaluation of a political ideology hate speech detector in tweets in Spanish

Detectando el odio ideológico en Twitter. Desarrollo y evaluación de un detector de discurso de odio por ideología política en tuits en español

Detectando o ódio ideológico no Twitter. Desenvolvimento e avaliação de um detector de discurso de ódio por ideologia política no Twitter em espanhol

Javier J. Amores, Universidad de Salamanca, Salamanca, España (javieramores@usal.es)

David Blanco-Herrero, Universidad de Salamanca, Salamanca, España (david.blanco.herrero@usal.es)

Patricia Sánchez-Holgado, Universidad de Salamanca, Salamanca, España (patriciasanc@usal.es)

Maximiliano Frías-Vázquez, Universidad de Salamanca, Salamanca, España (maxfrias@usal.es)

ABSTRACT | Hate speech spread through social media such as Twitter deserves special attention, as its increase may be related to the rise in hate crimes. Of the 11 categories of discrimination contemplated by the Spanish Ministry of Internal Affairs, the second in which the most hate crimes are registered per year is political ideology. However, this category falls outside of most action plans to study and combat hate crimes. The same occurs in the case of academic works since most focus on analyzing and detecting hate in English and at a general level. The few authors who have targeted their studies to a single type of hate to improve accuracy, have focused on racism, xenophobia, or gender discrimination, but never on political ideology. Furthermore, the detection prototypes developed so far have not used databases generated manually by various coders. This paper aims to overcome these limitations, developing and evaluating an automatic hate speech detector on Twitter in Spanish for reasons of ideological discrimination, using supervised machine learning techniques. For this, we developed a total of eight predictive models from

HOW TO CITE

Amores, J. J., Blanco-Herrero, D., Sánchez-Holgado, P. & Frías-Vázquez, M. (2021). Detectando el odio ideológico en Twitter. Desarrollo y evaluación de un detector de discurso de odio por ideología política en tuits en español. *Cuadernos.info, (49)*, 98-124. https://doi.org/10.7764/cdi.49.27817

an ad-hoc generated training corpus, and making use of shallow modelling, but also deep learning, which has allowed to improve the final performance of the prototype. In addition, the development of the corpus allowed us to observe that 16.2% of the sample, collected in autumn 2019 and manually analyzed, included some type of ideological hatred.

KEYWORDS: hate speech; online hate; Twitter; political ideology; deep learning; machine learning; supervised classification.

RESUMEN | *El discurso de odio propagado a través de redes sociales como Twitter merece atención especial, ya que su incremento puede relacionarse con el aumento de crímenes de odio. De las 11 categorías de discriminación que contempla el Ministerio de Interior de España, la segunda en la que más delitos de odio se registran al año es la ideología. Sin embargo, esta categoría queda fuera de la mayor parte de los planes de acción para estudiar y combatir los delitos de odio. Lo mismo ocurre con los trabajos académicos, que se centran mayoritariamente en el odio en inglés y a nivel general. Los que estudian un único tipo de odio se han enfocado en el racismo, la xenofobia o la discriminación de género, pero nunca en la ideología política. Asimismo, los prototipos de detección desarrollados hasta ahora no usan bases de datos generadas manualmente por varios codificadores. Esta investigación busca superar estas limitaciones, desarrollando y evaluando un detector automático de discurso de odio por motivos ideológicos en Twitter en español a partir de técnicas de aprendizaje automático supervisado. Para ello, se ha desarrollado un total de ocho modelos predictivos a partir de un corpus de entrenamiento generado ad-hoc, y haciendo uso de modelado superficial y de aprendizaje profundo, lo que permite mejorar el rendimiento final del prototipo. El desarrollo del corpus permitió observar, además, que un 16,2% de la muestra, recogida en el otoño de 2019, incluyó algún tipo de odio ideológico.*

PALABRAS CLAVE: *discurso de odio; odio en línea; Twitter; ideología política; aprendizaje profundo; aprendizaje automático; clasificación supervisada.*

RESUMO | O discurso de ódio que se espalha pelas redes sociais como o Twitter merece atenção especial, pois seu aumento pode estar relacionado ao aumento dos crimes de ódio. Das onze categorias de discriminação contempladas pelo Ministério do Interior da Espanha, a segunda em que mais crimes de ódio são registrados por ano é a ideologia política. No entanto, esta categoria está fora da maioria dos planos de ação para estudar e combater os crimes de ódio. O mesmo acontece com os trabalhos acadêmicos, já que a maioria concentra-se em analisar e detectar o ódio em inglês e em um nível geral, e os poucos autores que limitaram seus estudos a um único tipo de ódio concentraram-se no racismo, xenofobia ou discriminação de gênero, mas nunca na ideologia política. Além disso, os protótipos de detecção desenvolvidos

até o momento não usaram bancos de dados gerados manualmente por vários codificadores. A presente pesquisa visa superar essas limitações, desenvolvendo e avaliando um detector automático de discurso de ódio no Twitter em espanhol por motivos de discriminação ideológica, baseado em técnicas de aprendizagem automática supervisionada. Para isso, foram desenvolvidos um total de 8 modelos preditivos a partir de um corpus de treinamento gerado ad-hoc, e fazendo uso de modelagem superficial, mas também de aprendizagem profunda, que tem permitido melhorar o desempenho final do protótipo. O processo de elaboração do corpus também nos permitiu observar que 16,2% da amostra, coletada no outono de 2019, incluía algum tipo de ódio ideológico.

PALAVRAS-CHAVE: discurso de ódio; ódio online; Twitter; ideologia política; aprendizagem profunda; aprendizagem de máquina; classificação supervisionada.

INTRODUCTION AND JUSTIFICATION

Hate speech deserves special academic attention due to its social implications, since it can be an important precursor to more serious crimes, which have increased in recent years (Organization for Security and Cooperation in Europe (OSCE), 2020). In this regard, Müller and Schwarz (2020) suggest that there is a correlation between the increase in online hate speech and hate crimes committed in certain regions and contexts; thus, studying this type of messages spread on social platforms is essential to prevent and counter its effects.

Social platforms seem to be the ideal environment to spread hate, especially on Twitter, due to its role in shaping public opinion, thanks to its use volume, 16% of the Spanish population according to the Reuters Institute Digital News Report (Newman et al., 2019), and the presence of politicians and journalists (Rodríguez & Ureña, 2011). This ability of social media to set the public agenda is also based on the interest they receive from traditional media (Bane, 2019).

In recent years, messages that express hatred, rejection, intolerance, or discrimination towards certain vulnerable groups have not stopped increasing in this social platform, spread by all types of users. Even during the recent health crisis, hate speech on Twitter has continued to increase, towards the sick, the elderly, migrants, foreigners, but also towards political leaders, who have been taking their discourses to the extreme. This increase in hate is observed in the latest *Online Hate and Harassment* reports from the Anti-Defamation League (2020, 2021), which reflect an exponential increase in all forms of cyber-hate in most social media since 2018. Recent studies have shown a negative trend in the representation of migrants and refugees transmitted by the main media of the Mediterranean countries (Amores et al., 2020) and Western Europe (Amores et al., 2019), which can be directly linked to the increase in hatred due to racism or xenophobia and, indirectly, with ideology.

This situation is especially noticeable in Spain, one of the countries in which the recent crises have had more visible effects, which may have fueled extremist and radical discourses (Ferreira, 2019). In addition to the increase in online and offline hate speech for ideological reasons, Spain does not have an independent national strategy aimed at preventing this type of crime. Although in September 2018 the government signed an institutional cooperation agreement with the General Council of the Judiciary and the State Attorney General's Office to fight intolerance (Ministerio de Empleo, Migraciones y Seguridad Social, 2018), it does not define an articulated action plan to combat the increase in hate crimes and, specifically, hate speech, which usually constitutes the root of other crimes. This makes implementing new methods that help automatically identifying and

monitoring online hate speech on a large scale to prevent and combat it even more necessary, while countering hate on the streets, and with it, hate crimes. In a digital environment free of surveillance and regulation, there is growing concern about the potential victims of this online hate, evidenced by the latest Raxen Report (Movimiento contra la Intolerancia, 2019). In this regard, it is difficult to calculate precisely the hate volume in a social platform, given its temporary fluctuation based on media events (Arcila Calderón et al., 2020); however, there is a growing trend in the volume of hate crimes, both in Spain and in the rest of Europe (OSCE, 2020), and given the connection between online hate and these crimes (Müller & Schwarz, 2020), the relevance of studying cyber-hate is unquestionable.

Political ideology is the second category of discrimination by volume of hate crimes after racism and xenophobia, according to the Report on the Evolution of Hate Crimes in Spain (Ministerio de Interior, 2020). Thus, the objective of this work is to develop and evaluate an automatic detector of ideological-based hate speech disseminated in Spanish through Twitter, using computational methods. To do this, we start from a sample previously selected with a filter of thematic keywords manually identified throughout the 2019 fall.

DEFINING ONLINE HATE SPEECH

Hate speech is not an exclusive concern of current societies but has traditionally existed as a radical way of expressing rejection and intolerance of otherness (Krippendorf, 2010). As early as 1997, Calvert pointed out this type of discourse as a problem to be analyzed, understood, and fought with communication approaches, necessarily involving all the elements of communication transmission models (source, message, channel, and receiver). Nevertheless, this type of discourse is of particular concern today due to the rise of social media and the new profile of prosumers (Carmona, 2010), with followers spreading unregulated content, in addition to the demonstrated social and public opinion influence they have (Isasi & Juanatey, 2017). Therefore, they are considered as a possible crime to investigate, and the need, urgency, and difficulty of their detection and elimination are debated (Jubany & Roiha, 2018; Tamarit Sumalla, 2018).

Before facing any strategy for detecting hate speech in digital environments, it is convenient to try to define it. In this regard, although there is still no single and standardized conceptualization of hate speech due to the term's breadth and subjectivity, several authors have proposed a definition and taxonomy, discussing the types and levels of hate speech that currently occur in society, based on whether they could be considered a crime or conceived within the margins of freedom

of expression. In this vein, Benesch (2014) moves away from hate terminology to propose the term dangerous speech, with which she refers to those speeches that have a considerable probability of triggering violence episodes. Leader Maynard and Benesch (2016) argue that both this discourse and the dangerous ideology that promotes it constitute a real risk of ending up turning into crimes and attacks, so any expression of hate must be monitored and combated given its dangerousness. Gagliardone and colleagues (2015) understand as hate speech all kinds of expressions that directly incite the commission of discrimination or violence acts for reasons of racial hatred, xenophobia, sexual orientation, or other forms of intolerance; they widen the term to those expressions that promote prejudice, considering that they can indirectly contribute to generate a hostility climate that can lead to discriminatory acts or violent attacks. According to these authors, the use of the term hate speech has become generalized to refer to a heterogeneous conglomerate of manifestations that range from threats to individuals or groups to cases in which some people simply express their anger against the authorities in a more or less offensive way. However, the conflict lies in the blurred limits of what is conceptualized as hate speech, which can often conflict with the fundamental right to freedom of expression, and how to discern what part of this complex amalgamation of speeches can constitute a crime. Something that, as Arroyo (2017) explains, usually resides in the mere theoretical and jurisprudential interpretation of the criminal code, which in the Spanish case only refers to this type of crime in article 510 of Organic Law 10/1995, regarding hate speech.

Miró Llinares (2016) tries to resolve this conflict by offering, in addition to a broad definition, a taxonomy that allows differentiating between the type of hate speech that could constitute a crime because it is more explicit, direct, or instigates physical violence, and the more subtle one that, although represents an offense and expresses rejection towards certain individuals or vulnerable groups, can be framed within the margins of freedom of expression. Nevertheless, when monitoring and combating cyber-hate, it is convenient to consider all the levels at which it can be represented and propagated, since, due to a cumulative effect, all of them can contribute in the same way when it comes to generating dehumanization, stigmatization and, ultimately, episodes of violence towards any type of otherness (Isasi & Juanatey, 2017).

On the other hand, at an institutional level, the European Union has tried to define the limits of freedom of expression, increasingly narrowing the conceptualization of hate speech, although without much practical success since it does not have a clear reflection in the jurisprudence of the different member countries. For the Council of Europe, through its Recommendation No. R (97)20

of the Committee of Ministers on hate speech (Council of Europe, 1997), this speech is understood as the promotion of messages that imply rejection, contempt, humiliation, harassment, disrepute, and stigmatization of individuals or social groups based on specific attributes. Thus, for a speech to be considered a hate crime, it must propagate, incite, promote, or justify racial hatred, xenophobia, anti-Semitism, and other forms of intolerance-based hate. In this vein, the European Commission against Racism and Intolerance, through its General Recommendation No. 15 on Combating Hate Speech (2016), specifies that hate can be motivated by reasons of race, color, descent, national or ethnic origin, ideology, age, disability, language, religion, sex, gender, gender identity, sexual orientation, and other personal characteristics or conditions. The Spanish Ministry of Internal Affairs (Ministerio de Interior de España, 2020), in its latest Evaluation Report on hate crimes in the country, includes 11 discrimination categories into which crimes committed against vulnerable audiences can be classified: (1) racism/xenophobia, (2) political ideology, (3) sexual orientation and gender identity, (4) religious beliefs or practices, (5) disability, (6) gender reasons, (7) antisemitism, (8) aporophobia, (9) anti-Gypsyism, (10) generational discrimination, and (11) discrimination due to illness. The first three are the ones that motivate the greatest number of hate crimes in Spain each year, according to the figures collected by the latest reports from the Ministry; the second –political ideology– is the type of discrimination that has increased the most in recent years, especially in digital spaces. However, this usually falls outside the margins of social, institutional, and academic interest when studying and analyzing hate speech. Based on these premises, this work focuses on detecting hate speech specifically motivated by political ideology. It also aims to cover all levels of typified hate, trying to broaden the detection of hate speech "spread on Twitter in Spanish", since it is expected that the most explicit –the one that could be considered a crime– will not have a large presence in the Spanish context.

DETECTING ONLINE POLITICAL IDEOLOGY-BASED HATE SPEECH

In recent years, numerous authors have studied these discourses from different perspectives. Chetty and Alathur (2018) analyze it from the jurisprudential basis, concluding that the appropriate political measures, as well as the actions of the social platforms themselves, are essential to effectively counter hate speech. Others, such as ElSherief and colleagues (2018), study it by using a data-based linguistic and psycholinguistic perspective, offering a framework of understanding from which to identify the hate spread on social platforms. With a more automated and massive detection approach, Mondal and colleagues (2017) propose a system to measure and monitor hate speech propagated on Twitter and Whisper based

on certain expressions and keywords and focusing attention on recognizing the main targets of massive hate. Malmasi and Zampieri (2017) propose a method for detecting hate spread on social media based on natural language processing and supervised classification techniques.

These works focus on online hate speech as a problem to be detected and combated, and treat it from a generic and international point of view, i.e., trying to identify hate speech propagated in English, motivated by all kinds of reasons, aimed at all kinds of audiences, at any time and place. This approach is very ambitious and can present problems of internal validity, especially in large-scale strategies. Even Salminen and colleagues (2020) recent prototypes, one of the most innovative and advanced because they use deep learning and include detection in various online sources, fall on this generic approach. This is a limitation, because the resulting models fail to be as effective, reliable and, paradoxically, generalizable, as those that are trained with real examples of a single type of hate and in a specific discriminatory category, thus differentiating concepts, characteristics, and linguistic nuances.

In this regard, it should be noted that on the international scene there are some examples of a cyber-hate detection strategy that considers levels, prejudice categories, or the vulnerable groups that are victims of this discourse. The work of Davidson and colleagues (2017) is one of them; they differentiate between direct hate messages and offensive messages. Another example is the one developed by Badjatiya and colleagues (2017), which identifies messages with racist or sexist content using deep modeling. Likewise, most of the works cited have a second limitation in common, i.e., that they have not used training corpus generated ad-hoc. To date, most of the existing prototypes base detection on previously developed lexicon dictionaries; when using corpus of examples to train classification algorithms, they usually use datasets already available from other authors and previous works, as occurs with the one developed by Salminen and colleagues (2020). This also influences the internal validity of the prototype and its final reliability. In the Spanish context, one of the few works that attempt to detect online hate speech in Spanish are those developed by Pereira Kohatsu (2017) and Pereira Kohatsu and colleagues (2019). Their prototype has the same limitations as most of the international ones previously exposed: although it developed an ad-hoc training corpus to generate the predictive models, it was elaborated by a single coder, which implies an internal validity problem due to its potential subjectivity.

Considering the above, this work focuses on developing a prototype capable of detecting political ideology-based hate speech propagated on Twitter in Spanish.

Until now, Arcila Calderón, Valdez Apolo, Blanco Herrero, and Amores are among the few authors who have focused their attention on analyzing and detecting the rejection manifested on Twitter for specific discrimination reasons, specifically racism and xenophobia. To do this, they first used manual analysis methods (Valdez-Apolo et al., 2019) to later develop a large-scale automatic detection technique based on supervised machine learning and using the corpus previously created manually (Arcila -Calderón et al., 2020). In this vein, this work's objective is to develop a more advanced detection strategy focused on hate based on political ideology. In this regard, it should be noted that the messages of a political nature transmitted through Twitter have been analyzed on numerous occasions, even in the Spanish context, but normally with the aim of studying the use made of this social platform by the parties or politicians (Marín Dueñas & Díaz Guerra, 2016; López-García, 2016), to analyze the contexts surrounding the campaigns and electoral days (LópezMeri, 2017; García-Ortega & Zugasti-Azagra, 2018), or to detect the orientation ideology and predict electoral results (Alonso González, 2017; Said-Hung et al., 2017). Arcila Calderón and colleagues (2017) previously developed a strategy for detecting political feelings on Twitter in Spanish based on supervised classification, which could also be applied to analyze political contexts, the support for different parties, and the prediction of election results. Nevertheless, no work has focused on the analysis and detection of hate speech for political reasons until now.

With these premises, we intend to solve and overcome, thanks to a series of differentiating elements, the limitations outlined. First, we will use supervised machine learning techniques to generate our own datasets, generated ad-hoc with manually classified examples, and with full inter-coder agreement, which serve as a training corpus for the resulting predictive models. Second, we will develop a specific training corpus of political ideology, to generate more reliable predictive models. In this regard, since the creation of the training corpus requires the manual classification of examples previously downloaded and filtered from the Twitter APIs, the following research question is posited: What frequency and percentage of political ideology-based hate tweets are detected through manual classification on a sample of previously filtered tweets? (RQ1).

The third innovative element is the use of deep learning to generate the predictive models. Specifically, this work uses recurrent neural networks, an algorithm that, a priori, should have significant advantages over traditional classification algorithms, offering better performance especially when applied to classifying texts. In this regard, the following questions are posited: Which machine learning algorithm presents the best performance to generate a predictive model capable of detecting political ideology-based hate speech on Twitter in

Spanish? (RQ2), and Does deep learning perform better than shallow learning to generate models capable of detecting political ideology-based hate speech spread on Twitter in Spanish? (RQ2A).

METHODOLOGY

To develop the political ideology-based hate speech detector on Twitter, we followed a large-scale detection strategy, based on intensive data computation under the supercomputing infrastructure of Castilla y León, Scayle, and using natural language processing techniques and supervised machine learning. The work was divided into three main phases, outlined in figure 1.

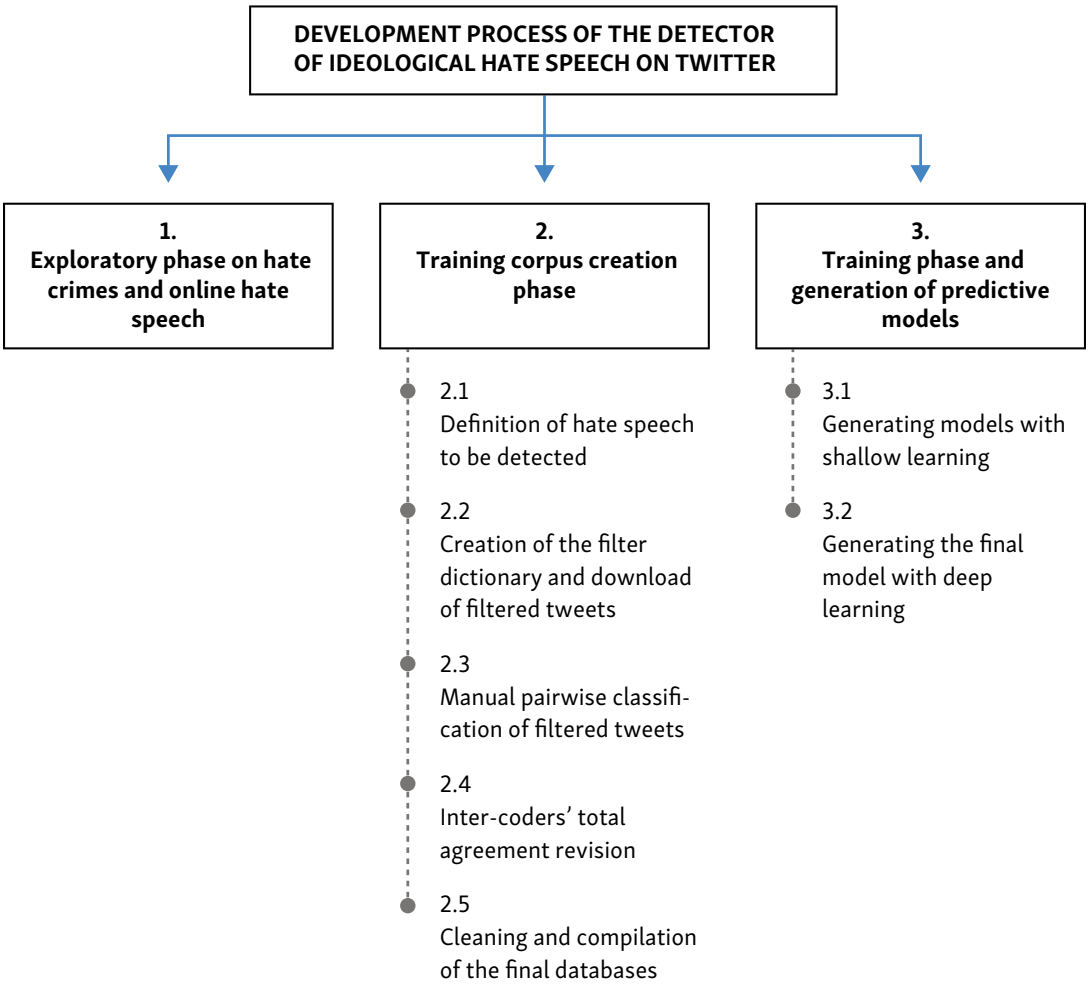

Figure 1. Methodological process conducted to develop the Automatic Detector of ideology-based hate speech on Twitter in Spanish

Source: Own elaboration.

Exploratory phase

In this first phase, we conducted an in-depth qualitative exploration of ideology-based hate speech spread on social media such as Twitter. In addition, as a theoretical approach, we carried out a literature review; accounts, profiles, and hashtags which publish a great number of hate messages due to political ideology were also identified. Exploring these potential hate sources on Twitter would serve to understand and limit the different ways in which ideology-based hate is expressed, the different contexts in which it spreads, as well as the most used terms and expressions. This later helped generating the linguistic filters that would allow tweets to be downloaded for manual classification.

Training corpus creation phase

In this phase, we created ad-hoc datasets from reliable examples of the type of hate to be detected that would serve as a corpus to train the predictive models that would finally allow hate messages to be detected automatically and massively. It is the longest and most laborious stage, which allows overcoming the limitations of the previously developed prototypes, which used dictionaries or general and pre-existing datasets. It is composed of a series of sub-stages, described below.

Definition and typology of hate speech to be detected

We first established criteria to limit the type of discourse to be detected to generate custom datasets. According to the possibilities identified in the qualitative exploration –and considering both the definitions provided by the different authors and institutions, as well as the European legal framework– we expanded the definition of hate speech, encompassing the different meanings and types offered from the academy, public institutions, and the Spanish criminal code, as well as the three levels of online hate provided by Miró Llinares (2016). Thus, to generate the datasets, we included all types of hate speech that can constitute a crime, but also the more subtle ones that, a priori, could be considered within the range of freedom of expression. This was determined because, in the previous phase, very little direct and explicit hate had been detected, and the intention was to be able to detect as many messages as possible with this type of content. In addition, in the validation process of the manual classification conducted following the bases of a content analysis, and in the models' subsequent training, the results would be refined, leaving the safest examples, filtering and rejecting the doubtful or ambiguous ones that did not have an intercoder agreement, which is why it was also interesting to cover the widest possible range of hate forms. Thus, given the scarcity of messages with illegal hate speech on Twitter within the Spanish framework, we decided to train the models to detect all levels of hateful content. What would be considered ideologically based hate speech was also defined, compiling all the derogatory terms, expressions, and objectives collected in the exploratory phase.

Creation of the filter dictionary and download of tweets

Once the levels and types of hate speech were defined, we generated a dictionary of words and their combinations that could serve as a filter to initially download potential hateful tweets based on political ideology. To do so, we used the Twitter accounts and hashtags propagating a greater amount of hate speech in Spain for ideological reasons. Subsequently, these messages were manually classified by referenced audiences and by inclusion of hate.

Second, based on these examples, we made a final selection of search terms, in a words' list format, roots or combinations of words that could be representative of ideological-based hate, following the distinction made by Kalampokis and colleagues (2013) to form the definitive dictionary that would serve as a filter for downloading. We then translated it into computational language (figure 2) to download the necessary number of tweets from the Twitter APIs. Thus, although a greater number of messages were downloaded, we finally collected a sample of 24,000 tweets, compiled in a dataset for later manual classification.

```
word = ['podemita', 'trifachito', 'falangito', 'rojo\ncomunista', 'extrema\nizquierda', 'extrema\nderecha', 'izmierda',
    'independentista\nfacha', 'independentista\nfascista', 'independentista\ngentuza', 'independentista\nbasura',
'independentista\nfacha','independentista\nimbecil', u'independentista\nimbécil', 'independentista\ngilipollass',
'independentista\nlacra', 'independentista\nescoria', 'independentista\asco', 'independentista\nmierda',
'independentista\nnazi', 'independentista\nputo', 'independentista\nputa', 'independentista\nmaldito',
'independentista\nmaldita', 'independentista\nsucio', 'independentista\nsucia',
    'independentismo\nfacha', 'independentismo\nfascista', 'independentismo\ngentuza', 'independentismo\nbasura',
'independentismo\nfacha','independentismo\nimbecil', u'independentismo\nimbécil', 'independentismo\ngilipollas',
'independentismo\nlacra', 'independentismo\nescoria', 'independentismo\asco', 'independentismo\nmierda',
'independentismo\nnazi', 'independentismo\nputo', 'independentismo\nputa', 'independentismo\nmaldito',
'independentismo\nmaldita', 'independentismo\nsucio', 'independentismo\nsucia',
    'socialista\nfacha', 'socialista\nfascista', 'socialista\ngentuza', 'socialista\nbasura',
'socialista\nfacha','socialista\nimbecil', 'socialista\ngilipollas', 'socialista\nlacra', 'socialista\nescoria', 'socialista\asco',
'socialista\nmierda', 'socialista\nnazi', 'socialista\nputo', 'socialista\nputa', 'socialista\nmaldito', 'socialista\nmaldita',
'socialista\nsucio', 'socialista\nsucia',
    'socialismo\nfacha', 'socialismo\nfascista', 'socialismo\ngentuza', 'socialismo\nbasura',
'socialismo\nfacha','socialismo\nimbecil', u'socialismo\nimbécil', 'socialismo\ngilipollass', 'socialismo\nlacra',
'socialismo\nescoria', 'socialismo\asco', 'socialismo\nmierda', 'socialismo\nnazi', 'socialismo\nputo', 'socialismo\nputa',
'socialismo\nmaldito', 'socialismo\nmaldita', 'socialismo\nsucio', 'socialismo\nsucia',
    'nacionalista\nfacha', 'nacionalista\nfascista', 'nacionalista\ngentuza', 'nacionalista\nbasura',
'nacionalista\nfacha','nacionalista\nimbecil', 'nacionalista\ngilipollas', 'nacionalista\nlacra', 'nacionalista\nescoria',
'nacionalista\asco', 'nacionalista\nmierda', 'nacionalista\nnazi', 'nacionalista\nputo', 'nacionalista\nputa',
'nacionalista\nmaldito', 'nacionalista\nmaldita', 'nacionalista\nsucio', 'nacionalista\nsucia',
    'nacionalismo\nfacha', 'nacionalismo\nfascista', 'nacionalismo\ngentuza', 'nacionalismo\nbasura',
'nacionalismo\nfacha','nacionalismo\nimbecil', 'nacionalismo\ngilipollas', 'nacionalismo\nlacra', 'nacionalismo\nescoria',
'nacionalismo\asco', 'nacionalismo\nmierda', 'nacionalismo\nnazi', 'nacionalismo\nputo', 'nacionalismo\nputa',
'nacionalismo\nmaldito', 'nacionalismo\nmaldita', 'nacionalismo\nsucio', 'nacionalismo\nsucia',
    'comunista\nfacha', 'comunista\nfascista', 'comunista\ngentuza', 'comunista\nbasura',
'comunista\nfacha','comunista\nimbecil', 'comunista\ngilipollas', 'comunista\nlacra', 'comunista\nescoria',
'comunista\asco', 'comunista\nmierda', 'comunista\nnazi', 'comunista\nputo', 'comunista\nputa', 'comunista\nmaldito',
'comunista\nmaldita', 'comunista\nsucio', 'comunista\nsucia',
    'comunismo\nfacha', 'comunismo\nfascista', 'comunismo\ngentuza', 'comunismo\nbasura',
'comunismo\nfacha','comunismo\nimbecil', 'comunismo\ngilipollas', 'comunismo\nlacra', 'comunismo\nescoria',
'comunismo\asco', 'comunismo\nmierda', 'comunismo\nnazi', 'comunismo\nputo', 'comunismo\nputa',
'comunismo\nmaldito', 'comunismo\nmaldita', 'comunismo\nsucio', 'comunismo\nsucia',
    'golpista\nfacha', 'golpista\nfascista', 'golpista\ngentuza', 'golpista\nbasura', 'golpista\nfacha','golpista\nimbecil',
'golpista\ngilipollas', 'golpista\nlacra', 'golpista\nescoria', 'golpista\asco', 'golpista\nmierda', 'golpista\nnazi',
'golpista\nputo', 'golpista\nputa', 'golpista\nmaldito', 'golpista\nmaldita', 'golpista\nsucio', 'golpista\nsucia',
    'golpismo\nfacha', 'golpismo\nfascista', 'golpismo\ngentuza', 'golpismo\nbasura',
'golpismo\nfacha','golpismo\nimbecil', 'golpismo\ngilipollas', 'golpismo\nlacra', 'golpismo\nescoria', 'golpismo\asco',
'golpismo\nmierda', 'golpismo\nnazi', 'golpismo\nputo', 'golpismo\nputa', 'golpismo\nmaldito', 'golpismo\nmaldita',
'golpismo\nsucio', 'golpismo\nsucia',
```

Figure 2. Fragment of the final script used for the filtered download of potential ideology-based hate tweets

Source: Own elaboration.

Manual pairwise classification

Subsequently, the messages were manually classified through the Doccano platform, which eased the task of labeling the texts between various coders (figure 3). Thus, all tweets were classified by one main coder and eight secondary ones (3000 tweets each). To subsequently cross-check the results and make the resulting messages more reliable, the secondary judges had to be people external to the the project, so we chose undergraduate and postgraduate students from the Universidad de Salamanca, who were trained prior to classification and who were given the examples manual as a code book. The messages were tagged in a binary way as hate and non-hate, and the main coder discarded messages unrelated to the topic.

Figure 3. Manual classification on the Doccano platform

Source: Own elaboration.

Checking inter-coder agreement

Once all the tweets were classified by two coders, we checked inter-coder reliability, keeping only those messages that were classified with the same label by both coder and discarding the rest. I.e., the inter-coder reliability would be $\alpha=1$. This step, in addition to guaranteeing the coding's quality, allows overcoming one of the main limitations of some prototypes such as that of Pereira-Kohatsu (2017), mentioned above.

Cleaning and compilation of the final datasets

Once this process was completed, the datasets were cleaned, with which the training corpus resulted in 16.2% reliable hateful tweets (N=3879) and 26.4% non-hateful (N=6334) (figure 4).

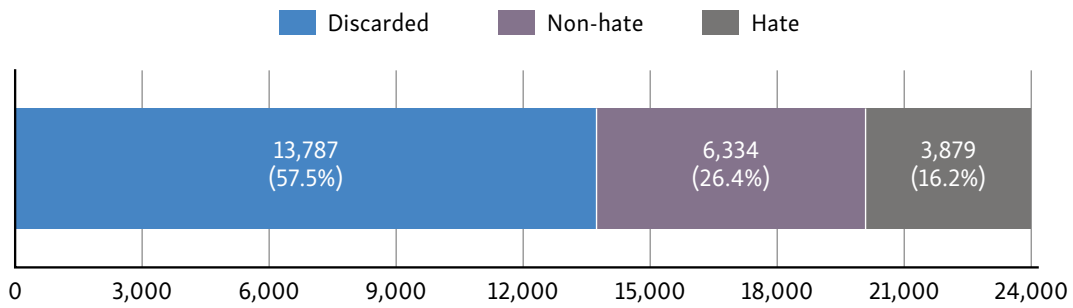

Figure 4. Frequencies and percentages of manual classification of political tweets

Source: Own elaboration.

Generating the predictive models

Once the training corpus validated, it was used to train and generate the predictive models that would finally allow detecting ideology-based hate speech on Twitter in Spanish automatically and on a large scale. A total of eight predictive models were generated: six using shallow learning algorithms, one generated from the votes of those previous models, and one using deep learning.

Shallow learning

The six predictive models that were generated using traditional classification algorithms were based on Bag of Words as a text representation, from which each word is taken as a vector. We used the NLTK and SciKit-Learn Python libraries to generate binary classification models and the following conventional shallow learning algorithms: Original Naïve Bayes, Multinomial Naïve Bayes, Bernoulli Naïve Bayes for multivariate models, Logistic Regression, Stochastic Gradient Descent Linear Regression, and Support Vector Machines. Natural language processing (NLP) techniques were also applied to extract features from the set of tagged messages. In the model training process, the most repeated words from the set of examples that made up the training corpus were tokenized and converted into quantitative characteristics or vectors with which the predictive models could work. In this modeling process, each of the corpus was randomly divided into two subgroups: 70% dedicated to training and 30% to testing and validating the models themselves. Thus, optimized classifiers were generated for each of the six algorithms mentioned and implemented on the training corpus to generate six predictive models capable of detecting hate speech in tweets in Spanish for political ideology reasons. Once these models were developed, we generated a model based on the vote of each of the previous six. This classifier chooses the category –hate/no hate– that most of the models predict (in the event of a tie, it does so randomly), adding a confidence indicator based on the proportion of said agreement (number

of votes for the majority class/number of possible votes), which made it possible to establish a confidence threshold greater than 80% (0.8) for each prediction. Each of the six classifiers, in addition to the one based on the voting of the other models, was evaluated using 30% of the training corpus destined for testing, to compare the manual classification of that sample with the predictions produced by the models.

Deep modeling

After modeling based on shallow learning, we developed a second strategy to classify the texts based on deep modeling, using embeddings as a form of text representation, and deep learning, specifically, recurrent neural networks (RNN). TensorFlow (v2) and Keras environment were used to create a sequential model with four layers:

- The first input layer converts each word into embeddings, dense vectors that represent the categorical value of any given word. The embeddings were trained using the 10,000 most common words from the created dataset, plus 1,000 words outside that vocabulary. Therefore, the embedding matrix included one row for each of these 11,000 words and one column for each of the six embedding dimensions (this hyperparameter was adjusted multiple times and performed best with size = 6).
- The second and third layers included are Gated Recurrent Unit hidden layers (GRU), with 128 neurons each. GRUs are simplified versions of traditional LSTMs, durable short-term memory cells used to create recurrent neural networks that allow predictions to be made about streams of data. Although both work fine for classifying text (converging quickly and detecting long-term dependencies), we chose to apply the first instead of the latter because the simplified version has similar performance and, being simpler, it offers faster execution.
- The last output layer is the one that allows detection, a dense layer with only one neuron that uses sigmoid activation to predict the probability that a message contains hate for each of the reasons present in the training data corpus.

To compile the deep model, we used standard loss with binary crossentropy and adam optimizer. Finally, the training corpus was adjusted for five periods and the part of the corpus intended for testing was used for validation in thirty steps. Since neural networks require a lot of computing capacity and there was a need to scale the processes from local to distributed, all the collection of examples, manual classification, and generation of the models were executed remotely and in parallel using the services of computation of the Supercomputing Center of Castilla y León.

RESULTS

Before reviewing the performance of the generated models, it is convenient to analyze the results of the manual classification conducted to generate the training corpus.

The first thing to point out is that there is a high percentage of the sample's tweets finally discarded, 57.7% (N=13,787), adding those that did not have an inter-coder agreement and those that were discarded in the classification. It is also observed that the percentage of ideology-based hateful tweets and validated with full agreement is low, despite being previously filtered messages. Specifically, and responding to RQ1, 16.2% of hateful tweets (N=3,879) were validated, compared to 26.4% of non-hateful messages (N=6,334). These figures show, first, that linguistic filter dictionaries, as complete and complex as they may be, as well as detection techniques based on expressions and keywords, do not serve as an effective method to identify online hate messages, something that was already assumed. Nevertheless, they served to limit and optimize the process, since without these linguistic filters the work to find examples of ideology-based hate in the general history of Twitter would have been arduous. Secondly, we can conclude from these data that the amount of hate (at least the explicit and the ideologically motivated) spread through Twitter is not as important as might be expected, although it may be very noisy and gimmicky. Table 1 shows a sample of example tweets from each of the resulting categories.

On the other hand, to evaluate the predictive models generated, we used three of the most used evaluation metrics in supervised machine learning: accuracy, the harmonic mean F-score –which offers a balanced metric and calculated from precision and recall–, and AUC-ROC –showing the performance of the classification models at all classification thresholds. Table 2 shows that all the values returned by these metrics were acceptable, in most cases above 0.70. Comparing the performance of each of the algorithms, the accuracy and AUC-ROC values were considerably higher in the model generated with recurrent neural networks, confirming the comparative advantage of deep learning applied to text classification. Thus, to answer RQ2, it can be concluded that, focusing specifically on shallow modelling, the traditional classification algorithm with the best performance in this case is logistic regression, followed by the detector based on the votes of shallow models and by Multinomial Naïve Bayes. However, in general terms, the deep model offers better performance than the models generated with shallow algorithms (see figure 5).

***Hate* tweets with agreement**
Anda a trabajar de una vez en tu vida, sucio comunista de mierda vende humos (Go to work for once in your life, you dirty smoke seller shitty communist)
Eres un cutre, puto facha asqueroso (You're seedy, fucking disgusting fascist)
La maldita izquierda y sus malditos delincuentes independentistas (Damn left and its damn pro-independence criminals)
Leña y más leña... ya esta bien de aguantar a #Guarros parasitos. #Asco #Izmierda (Firewood and more firewood... enough of putting up with #disgusting parasites. #disgust #leftshit)
Mandemos a esa escoria socialista al gulag, hay que aplastarlos a todos nido de rojos de mierda (Let's send that socialist scum to the gulag, we must crush them all, shitty communists)
Muerte a los separatas y a los podemitas terroristas y narcocomunistas (Death to the separatists, the terrorist, and narco-communist from Podemos)
Izquierdosos de mierda ojalá se mueran todos (Shitty leftists I hope they all die)
Es que os metía a todos en una cámara de gas hijos de la gran puta (I would put you all in a gas chamber sons of a bitch)
este tb es un hp. Un fascista de mierda el Rivera (This one too is a son of a bitch. Rivera is a shitty fascist)
INCREÍBLE como la perversa basura comunista pudre todo! Hay que acabar con ellos (INCREDIBLE how the perverse communist garbage rots everything! We have to end them)

***Non-hate* tweets with agreement**
La izquierda en lo suyo como ya es costumbre (The left-wing in its own thing, as usual)
Las pancartas son siempre de los mismos (The posters are always from the same people)
Si algunos ultras son de extrema izquierda, por qué nunca lo decís (If some ultras are extreme left, why do you never say so?)
okdiario Para ellos, un español es extrema derecha o no es nada (okdiario: For them, a Spanish is extreme right-winged or is nothing)
Estos comportamientos en Alemania estan castigado con cárcel! Vergüenza (These behaviors in Germany are punishable by jail! Shame)
Aquí un constitucionalista apoyando a los de la bandera del grajo, ya no se esconden (Here is a constitutionalist supporting those of the communist flag, they no longer hide)
No tan solo la extrema izquierda gana, la extrema derecha esta al acecho... (Not only the extreme left wins, the extreme right is on the prowl...)
el_pais Uy madre mía la que están liando las derechas ay Vox uy el trifachito (el_pais Oh my goodness the mess that the rights are doing... oh Vox oh the triple fascist)
A3Noticias Puto montaje, que mediocridad y poca seriedad, buscan en sí populismo! (A3Noticias Freaking montage, what mediocrity and little seriousness, they seek populism!)
Este es el nivel de respeto de Vox, es decir, ninguno (This is Vox's level of respect, that is, none)

Tweets without agreement (ambiguous messages that generated discrepancy)
Y esto es lo que pasa cuando se queman los contenendores, que la basura se acumula (And this is what happens when the containers burn, that the garbage accumulates)
Ojalá sigan cargándose el país estos comunistas que son lo mejor que nos ha pasado (I hope these communists continue to destroy the country, they are the best thing that has happened to us)
Gran gestión de los amigos zurdos (Great management of lefty friends)
Esos son los cómplices civiles de los crímenes de la derecha fascista (Those are the civilian accomplices of the crimes of the fascist right)
Vaya iagen: fachas de la estelada vs fachas de la bandera franquista. Qué absurdo y qué espanto (What an image: fascists of the estelada flag vs. fascist of the Franco flag. How absurd and how
No se porque todavía me sorprendo con la habilidad de la izquierda de volver todo caos (I don't know why I'm still amazed at the left's ability to turn all into chaos)
Vaya, me encanta cuando el podemita le ronea a Marhuender (Wow, I love it when the Podemos guy purrs at Marhuender)
Antifascistas apoyando a un movimiento nacionalista, insolidàrio y xenófobo. No sé si serán antifascistas o solo idiotas (Anti-fascists supporting a nationalist, unsupportive, and xenophobic movement. I don't know if they are antifascists or just idiots)
Claro que sí, que los islamoterroristas son niño de pecho al lado de la extrema izquierda liberal progres Globalistaa (Of course yes, Islamo-terrorists are babies next to the extreme liberal left Globalist)
Pues no se ya qué va a hacer Naranjito...buscar un espacio en la extrema izquierda (Well, I don't know what Naranjito is going to do...find a space on the extreme left)

Table 1. Example tweets of each of the resulting categories after coding

Source: Own elaboration.

Shallow learning	*Accuracy*	*F-Score*	*AUC-ROC*
Original Naïve Bayes	.66	.73	.63
Multinomial Naïve Bayes	**.68**	**.78**	**.60**
Bernoulli Naïve Bayes	.62	**.76**	.50
Logistic regression	**.70**	**.78**	**.65**
Linear regression with stochastic gradient descent	.67	.75	.63
Support vector machines	.67	.75	**.64**
Model based on the votes of the shallow models	**.70**	.75	**.64**
Deep learning	***Accuracy***	***F-Score***	***AUC-ROC***
Recurrent neural networks	**.81**	**.77**	**.89**

Table 2. Evaluation metrics of the models generated with each of the algorithms

Source: Own elaboration.

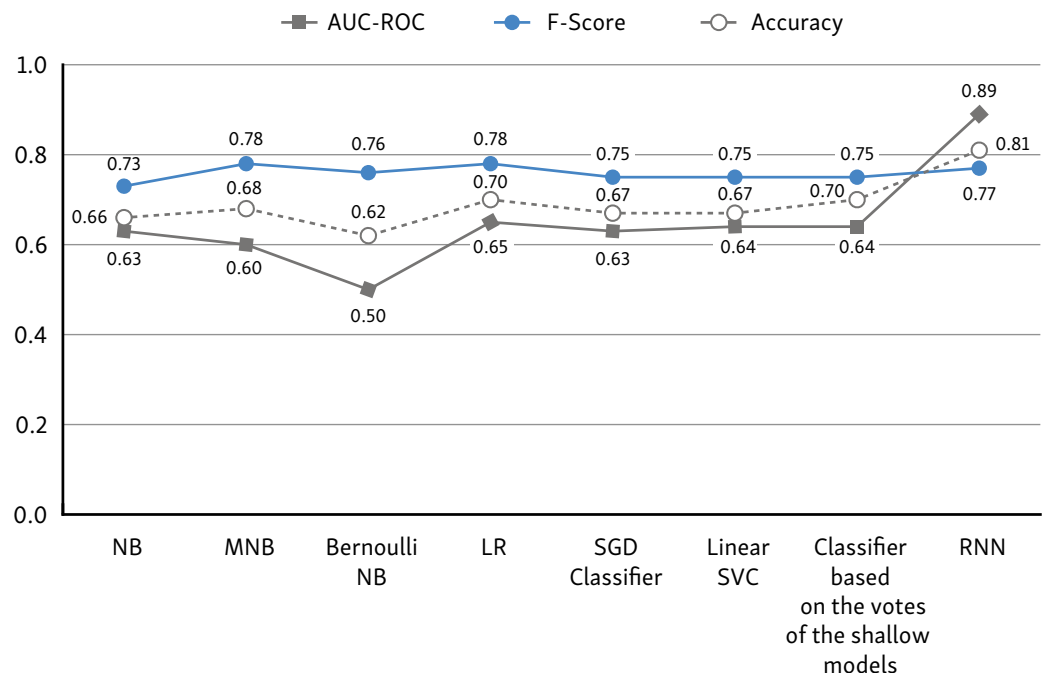

Figure 5. Evaluation metrics of the models generated with each of the algorithms

Source: Own elaboration.

CONCLUSIONS AND DISCUSSION

This article presents the first prototype for the automatic detection of political ideology-based hate speech on Twitter in Spanish, modeled from a manually generated ad-hoc training corpus, and also using deep learning, an innovation compared to previous prototypes. The main techniques used to develop this prototype have been natural language processing, to analyze and process unstructured data, and text classification with supervised machine learning, to detect hate based on political ideology. The computational strategy developed for the final detector involves downloading the messages from the Twitter streaming API and their direct and massive processing at the Supercomputing Center of Castilla y León, Scayle, where the trained and validated predictive models are applied to generate datasets with the messages finally classified as reliable, in hate and non-hate groups, for observation by the end user.

This work confirms that it is possible to train predictive models that allow detecting hate speech on Twitter due to a specific type of discrimination, such as political ideology, which also allows to better limit and specify the models' training, resulting in a solid performance, with a more than acceptable reliability and accuracy. In addition, a specific dataset has been created to train the predictive models, which allows improving the reliability of the detector applied to this specific context, overcoming previous prototypes' possible problems of internal validity. It should be noted that, although the final percentage of hate and non-hate messages with agreement in the training corpus may seem low, the most important is to have

quality examples, rather than quantity. This is because, although the evaluation metrics could be acceptable, if the examples are not completely reliable, the internal validity of the prototype could be damaged, contaminated with false positives or negatives. Thus, the focus was on generating a reliable and validated training corpus, since, in addition, the quantity can be easily expanded with new classified examples.

In short, we have resolved that, of the six machine learning algorithms used in shallow modeling, the one that offers the best performance is logistic regression, followed by Multinomial Naïve Bayes. Nevertheless, in general terms, we verified that deep learning works considerably better than conventional classification algorithms to detect this type of hate speech on Twitter, since the model trained with neural networks presented better evaluation metrics.

Beyond technical and methodological issues, the study has made it possible to observe a notable presence of hate speech –16.2% of the total sample and 38% of the tweets classified with certainty– on a sample previously selected with keywords. This allows contributing to theoretical discussions, not only about its definition and taxonomy (Miró Llinares, 2016), the limits to freedom of expression (Moretón Toquero, 2012) and the implications (Müller & Schwarz, 2020), but also about its quantification. This task, especially complex due to the volatility of this discourse (Arcila Calderón et al., 2020), can benefit from a validated and specific tool like this one, so that the same type of hate –political ideology– can be measured in different periods, helping to measure its evolution.

It can be affirmed that this work presents a methodological contribution thanks to the large-scale detection strategy, the generation of the ad-hoc training corpus, and the models developed with supervised machine learning techniques; a theoretical advance in the study of hate crimes and, specifically, political ideology-based hate speech on Twitter for reasons, and a practical application, since the technology developed here can be implemented in various public and private spheres. The latter is the most relevant, due to its potential application by social platforms to locate and reduce the presence of hate, by public, private or third-sector institutions, including the media and even political parties, that try to promote spaces less radicalized and polarized. Thus, the prototype could also be useful in projects that seek precisely to combat hate speech or polarization in social platforms, such as WONT-HATE[1] or TRI-POL[2].

1. Led by the Universidad de Navarra, financed by the H2020 program. https://cordis. europa.eu/project/id/795937/es

2. Led by Universidad Pompeu Fabra, financed by the Ministry of Science and Innovation and the BBVA Foundation: https://www.upf.edu/web/tri-pol

LIMITATIONS AND FUTURE RESEARCH LINES

Despite these contributions, this project has several limitations to face. First, the developed classification models have acceptable evaluation metrics; however, the prototype requires a validation that proves its reliability by being implemented in a practical way, with new real cases, and being compared with a new validated manual classification. This would require collecting a new sample of messages in a different context, coding it manually in the same way as it was done with the corpus (two coders) and, on that same sample, running the models and comparing the results of each classification, to later extract coefficients according to manual and automatic classification.

Secondly, the developed prototype can only detect political ideology-based hate speech in Twitter messages and in Spanish, which has allowed the development of more reliable models, but only applicable in this context; thus, it would be advisable to train and develop models based on this same strategy to detect hate speech on Twitter for other discrimination reasons, as well as in other languages and contexts, tasks on which the article's authors are working on.

Finally, the prototype is limited to detecting hate speech only on Twitter, so it should be extended to more sources, including social media such as YouTube or Instagram, political parties and associations blogs or websites, as well as digital media platforms. In this regard, although it is recognized that Twitter is not representative of public opinion (nor any isolated social media), its content tends to impact and go viral, reaching all kinds of people, with or without an account on the social platform.

REFERENCES

Alonso González, M. (2017). Predicción política y Twitter: Elecciones generales de España 2015 (Political prediction and Twitter: Spanish legislative elections 2015). *ZER: Revista de Estudios de Comunicación = Komunikazio Ikasketen Aldizkaria, 22*(43), 13-30. https://doi.org/10.1387/zer.16298

Amores, J. J., Arcila-Calderón, C. A., & Stanek, M. (2019). Visual frames of migrants and refugees in the main Western European media. *Economics & Sociology, 12*(3), 147-161. https://doi.org/10.14254/2071-789X.2019/12-3/10

Amores, J. J., Arcila-Calderón, C., & Blanco-Herrero, D. (2020). Evolution of negative visual frames of immigrants and refugees in the main media of Southern Europe. *Profesional de la Información, 29*(6). Retrieved from https://revista.profesionaldelainformacion.com/index.php/EPI/article/view/80525

Anti-Defamation League. (2020). *Online Hate and Harassment. The American Experience 2020*. The ADL Center for Technology and Society. Retrieved from https://www.adl.org/media/14643/download

Anti-Defamation League. (2021). *Online Hate and Harassment. The American Experience 2021*. The ADL Center for Technology and Society. Retrieved from https://www.adl.org/media/16033/download

Arcila-Calderón, C., Blanco-Herrero, D., & Valdez-Apolo, M. B. (2020). 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). *REIS: Revista Española de Investigaciones Sociológicas*, *172*, 21-40. https://doi.org/10.5477/cis/reis.172.21

Arcila-Calderón, C., Ortega-Mohedano, F., Amores, J. J., & Trullenque, S. (2017). Análisis supervisado de sentimientos políticos en español: clasificación en tiempo real de tweets basada en aprendizaje automático (Supervised sentiment analysis of political messages in Spanish: Real-time classification of tweets based on machine learning). *Profesional de la Información*, *26*(5), 973-982. https://doi.org/10.3145/epi.2017.sep.18

Arroyo, S. C. (2017). El concepto de delitos de odio y su comisión a través del discurso: especial referencia al conflicto con la libertad de expresión (The concept of hate crimes and their execution through speech: special reference to the conflict with freedom of speech). *Anuario de derecho penal y ciencias penales, 70*(1), 139-225. Retrieved from http://agora.edu.es/servlet/articulo?codigo=6930585

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017, April). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion* (pp. 759-760). https://doi.org/10.1145/3041021.3054223

Bane, K. C. (2019). Tweeting the agenda: How print and alternative web-only news organizations use Twitter as a source. *Journalism Practice*, *13*(2), 191-205. https://doi.org/10.1080/17512786.2017.1413587

Benesch, S. (2014). Defining and diminishing hate speech. In P. Grant (Ed.), *State of the World's Minorities and Indigenous Peoples* (pp. 18-25). Retrieved from https://minorityrights.org/publications/state-of-the-worlds-minorities-and-indigenous-peoples-2014-july-2014/

Calvert, C. (1997). Hate speech and its harms: A communication theory perspective. *Journal of Communication, 47*(1), 4-19. https://doi.org/10.1111/j.1460-2466.1997.tb02690.x

Carmona, O. I. (2010). Internet 2.0: El territorio digital de los prosumidores (Web 2.0: the digital territory of prosumers). *Revista Estudios Culturales*, (5), 43-64. Retrieved from http://servicio.bc.uc.edu.ve/multidisciplinarias/estudios_culturales/

Chetty, N. & Alathur, S. (2018). Hate speech review in the context of online social networks. *Aggression and violent behavior, 40*, 108-118. https://doi.org/10.1016/j.avb.2018.05.003

Council of Europe. (1997). *Recommendation No. R (97) 20 of the Committee of Ministers to member states on "hate speech"*. Council of Europe, Committee of Ministers. Retrieved from https://search.coe.int/cm/Pages/result_details.aspx?ObjectID=0900001680505d5b

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). *Automated hate speech detection and the problem of offensive language*. In *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1). Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/14955

ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W. Y., & Belding, E. (2018). *Hate lingo: A target-based linguistic analysis of hate speech in social media*. In *Proceedings of the International AAAI Conference on Web and Social Media, 12*(1). Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/15041

European Commission against Racism and Intolerance. (2016). *ECRI General Policy Recommendation N.° 15 on Combating Hate Speech*. Council of Europe. Retrieved from https://book.coe.int/en/human-rights-and-democracy/7180-pdf-ecri-general-policy-recommendations-no-15-on-combating-hate-speech.html

Ferreira, C. (2019). Vox como representante de la derecha radical en España: un estudio sobre su ideología (Vox as representative of the radical right in Spain: A study of its ideology). *Revista Española de Ciencia Política*, (51), 73-98. https://doi.org/10.21308/recp.51.03

Jubany, O. & Roiha, M. (2018). *Las palabras son armas. Discurso de odio en la red* (Words are weapons. Hate speech online). Barcelona, Spain: Edicions Universitat Barcelona.

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). *Countering online hate speech*. Paris, France: Unesco Publishing.

García-Ortega, C. & Zugasti-Azagra, R. (2018). Gestión de la campaña de las elecciones generales de 2016 en las cuentas de Twitter de los candidatos: entre la autorreferencialidad y la hibridación mediática (The management of the candidates' Twitter accounts in the Spanish 2016 general elections: Between self-referentiality and media hybridization). *Profesional de la Información*, *27*(6), 1215-1224. https://doi.org/10.3145/epi.2018.nov.05

Isasi, A. C. & Juanatey, A. G. (2017). *El discurso del odio en las redes sociales: Un estado de la cuestión* (Hate speech on social media: A state of the art). Barcelona, Spain: Ajuntament de Barcelona Progress Report. Retrieved from https://ajuntament.barcelona.cat/bcnvsodi/wp-content/uploads/2015/03/Informe_discurso-del-odio_ES.pdf

Kalampokis, E., Tambouris, E., & Tarabanis, K. (2013). Understanding the predictive power of social media. *Internet Research*, *23*(5), 544–559. https://doi.org/10.1108/IntR-06-2012-0114

Krippendorff, K. (2010). *On communicating: Otherness, meaning, and information*. Routledge.

Leader Maynard, J. & Benesch, S. (2016). Dangerous speech and dangerous ideology: An integrated model for monitoring and prevention. *Genocide Studies and Prevention*, *9*(3), 70-95. https://doi.org/10.5038/1911-9933.9.3.1317

López-García., G. (2016). 'New' vs 'old' leaderships: the campaign of Spanish general elections 2015 on Twitter. *Communication & Society, 29*(3), 149-168. https://doi.org/10.15581/003.29.3.149-168

López-Meri, A. (2015). Twitter como fuente informativa de sucesos imprevistos: el seguimiento de hashtags en el caso #ArdeValencia (Twitter as an Information Source of Unexpected Events: Following Hashtags in the Case #ArdeValencia). *Disertaciones: Anuario electrónico de estudios en Comunicación Social, 8*(1), 27-51. https://doi.org/10.12804/disertaciones.01.2015.02

Malmasi, S. & Zampieri, M. (2017). *Detecting hate speech in social media*. arXiv preprint:1712.06427. Retrieved from https://arxiv.org/abs/1712.06427

Marín Dueñas, P. P. & Díaz Guerra, A. (2016). Uso de Twitter por los partidos y candidatos políticos en las elecciones autonómicas de Madrid 2015 (Use of Twitter by political parties and candidates in the 2015 Madrid regional elections). *Ámbitos: Revista Internacional de Comunicación,* (32), 1-15. Retrieved from https://revistascientificas.us.es/index.php/Ambitos/article/view/10436

Ministerio del Interior de España (Ed.). (2020). *Informe de Evolución de los Delitos de Odio en España* (Report on the Evolution of Hate Crimes in Spain). Retrieved from http://www.interior.gob.es/documents/642012/3479677/Informe+sobre+la+evolución+de+delitos+de+odio+en+España%2C%20año+2019/344089ef-15e6-4a7b-8925-f2b64c117a0a

Ministerio de Empleo, Migraciones y Seguridad Social. (2018). *Acuerdo de cooperación institucional con el Consejo General del Poder Judicial y la Fiscalía General del Estado, para luchar contra el racismo, la xenofobia, la LGBTIfobia y otras formas de Intolerancia* (Institutional cooperation agreement with the General Council of the Judiciary and the State Attorney General's Office, to fight against racism, xenophobia, LGBTIphobia and other forms of Intolerance). Retrieved from http://www.inclusion.gob.es/oberaxe/ficheros/ejes/cooperacion/Acuerdo_insterinsticuional_original.pdf

Miró Llinares, F. (2016). Taxonomía de la comunicación violenta y el discurso del odio en Internet (Taxonomy of violent communication and the discourse of hate on the internet). *IDP. Revista de Internet, Derecho y Política*, (22), 82-107. Retrieved from https://www.raco.cat/index.php/IDP/article/view/n22-miro/408486

Mondal, M., Silva, L. A., & Benevenuto, F. (2017, July). A measurement study of hate speech in social media. In *Proceedings of the 28th ACM conference on hypertext and social media* (pp. 85-94). https://doi.org/10.1145/3078714.3078723

Moretón Toquero, M. A. (2012). El «ciberodio», la nueva cara del mensaje de odio: entre la cibercriminalidad y la libertad de expresión (Cyberhate, the new face of the hate message: between cybercrime and freedom of expresión). *Revista Jurídica de Castilla y León, 27*, 1-18.

Movimiento contra la Intolerancia. (2019). *Informe Raxen: Racismo, Xenofobia, Antisemitismo, Islamofobia, Neofascismo y otras manifestaciones de intolerancia a través de los hechos. Especial 2019. Por un Pacto de Estado contra la Xenofobia y la Intolerancia* (Raxen Report: Racism, Xenophobia, Anti-Semitism, Islamophobia, Neo-fascism and other manifestations of intolerance through facts. Special 2019. For a State Pact against Xenophobia and Intolerance). Movimiento contra la Intolerancia. Retrieved from https://www.inclusion.gob.es/oberaxe/ficheros/documentos/InformeRaxen.pdf

Müller, K. & Schwarz, C. (2020). Fanning the Flames of Hate: Social Media and Hate Crime. *Journal of the European Economic Association*, jvaa045. https://doi.org/10.1093/jeea/jvaa045

Newman, N., Fletcher, R., Kalogeropoulos, A., & Nielsen, R. (2019). *Reuters Institute Digital News Report 2019*. Reuters Institute for the Study of Journalism. Retrieved from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/inline-files/DNR_2019_FINAL.pdf

Organization for Security and Cooperation in Europe. (2020). *OSCE - ODIHR. Hate Crime Reporting*. Retrieved from https://hatecrime.osce.org/

Pereira Kohatsu, J. C. (2017). *Construcción de modelos de clasificación automática para discursos de odio* (Building automatic classification models for hate speech) (Master's thesis). Retrieved from https://repositorio.uam.es/handle/10486/680053

Pereira-Kohatsu, J. C., Quijano-Sánchez, L., Liberatore, F., & Camacho-Collados, M. (2019). Detecting and monitoring hate speech in Twitter. *Sensors, 19*(21), 4654. https://doi.org/10.3390/s19214654

Rodríguez, R. & Ureña, D. (2011). Diez razones para el uso de Twitter como herramienta en la comunicación política y electoral (Ten reasons to use Twitter as a tool for politicaland electoral communication). *Comunicación y pluralismo*, (10), 89-116. Retrieved from https://summa.upsa.es/viewer.vm?id=30573&view=main&lang=es

Said-Hung, E. M., Prati, R. C., & Cancino-Borbón, A. (2017). La orientación ideológica de los mensajes publicados en Twitter durante el 24M en España (The Ideological Orientation of Messages Posted on Twitter during the 24M in Spain). *Palabra Clave*, *20*(1), 213-238. https://doi.org/10.5294/pacla.2017.20.1.10

Salminen, J., Hopf, M., Chowdhury, S. A., Jung, S. G., Almerekhi, H., & Jansen, B. J. (2020). Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences, 10*, 1. https://doi.org/10.1186/s13673-019-0205-6

Tamarit Sumalla, J. M. (2018). Los delitos de odio en las redes sociales (Hate crimes on social networks). *IDP: Revista de Internet, Derecho y Política*, *27*, 17-29. Retrieved from https://www.raco.cat/index.php/IDP/article/view/n27-tamarit

Valdez-Apolo, M. B., Arcila-Calderón, C., & Amores, J. J. (2019). El discurso del odio hacia migrantes y refugiados a través del tono y los marcos de los mensajes en Twitter (Hate speech against migrants and refugees through the tone and frames of Twitter messages). *Revista de la Asociación Española de Investigación de la Comunicación*, *6*(12). https://doi.org/10.24137/raeic.6.12.2

ACKNOWLEDGMENTS/FUNDING

This work was developed within the framework of the *STOPHATE project, Development and Evaluation of an online hate speech detector in Spanish* [PC-TCUE18-20_016], competitive proof of concept led by Dr. Carlos Arcila Calderón, and funded by the European Regional Development Fund and the Junta de Castilla y León through PLAN T-CUE of the Fundación General of the Universidad de Salamanca (2018-2020). The authors especially thank Dr. Félix Ortega Mohedano and all the students of the Universidad de Salamanca who collaborated in the coding tasks for their participation and involvement in the project, without whom this work could not have been conducted.

ABOUT THE AUTHORS

JAVIER JIMÉNEZ AMORES, researcher member of the Observatorio de los Contenidos Audiovisuales (Audiovisual Content Observatory). He holds degree in Audiovisual Communication and a master in Research in Audiovisual Communication from the Universidad de Salamanca; he is currently developing his doctoral thesis at the same university, with the financial support of the Junta de Castilla y León and the European Social Fund. His research focuses on social media and network analysis, social communication, hate speech, and computational methods in social sciences.

https://orcid.org/0000-0001-7856-5392

DAVID BLANCO-HERRERO, doctoral student at the Universidad de Salamanca, where he develops his thesis with an FPU scholarship. He holds a degree in Journalism (Universidad a Distancia de Madrid) and Business Administration (Universidad de León) and a master's degree in Audiovisual Communication (Universidad de Salamanca). He is a member of the Audiovisual Content Observatory; his research lines are journalistic ethics, disinformation, and hate speech. He is editorial assistant in the *Anuario Electrónico de Estudios en Comunicación Social "Disertaciones"*.

https://orcid.org/0000-0002-7414-2998

PATRICIA SÁNCHEZ-HOLGADO, researcher at the Universidad de Salamanca and member of the Audiovisual Content Observatory. She has a degree in Advertising and Public Relations (Universidad Complutense de Madrid). She is an associate professor in the Faculty of Languages and Education of the Universidad Nebrija de Madrid and in the Faculty of Communication of the Universidad Pontificia de Salamanca. She is an expert in Big Data (Universidad Pontificia de Salamanca) and has a master's degree in Science, Technology, and Innovation Studies (Universidad de Oviedo).

http://orcid.org/0000-0002-6253-7087

MAXIMILIANO FRÍAS VÁZQUEZ, doctoral student at the Universidad de Salamanca and researcher member of the Observatory of Audiovisual Content. Holds a degree in Communication Sciences from the Universidad La Salle, Mexico, and a master's in Audiovisual Communication Research from the Universidad de Salamanca; his research lines are migration and hate speech, social network analysis, and big data.

https://orcid.org/0000-0001-9750-6136