Multimodal Technologies and Interaction

MDPI

Article

Using Shallow and Deep Learning to Automatically Detect Hate Motivated by Gender and Sexual Orientation on Twitter in Spanish

Carlos Arcila-Calderón *, Javier J. Amores, Patricia Sánchez-Holgado and David Blanco-Herrero

Faculty of Social Sciences, Unamuno Campus, University of Salamanca, 37007 Salamanca, Spain; javieramores@usal.es (J.J.A.); patriciasanc@usal.es (P.S.-H.); david.blanco.herrero@usal.es (D.B.-H.)
* Correspondence: carcila@usal.es

Abstract: The increasing phenomenon of "cyberhate" is concerning because of the potential social implications of this form of verbal violence, which is aimed at already-stigmatized social groups. According to information collected by the Ministry of the Interior of Spain, the category of sexual orientation and gender identity is subject to the third-highest number of registered hate crimes, ranking behind racism/xenophobia and ideology. However, most of the existing computational approaches to online hate detection simultaneously attempt to address all types of discrimination, leading to weaker prototype performances. These approaches focus on other reasons for hate—primarily racism and xenophobia—and usually focus on English messages. Furthermore, few detection models have used manually generated databases as a training corpus. Using supervised machine learning techniques, the present research sought to overcome these limitations by developing and evaluating an automatic detector of hate speech motivated by gender and sexual orientation. The focus was Spanish-language posts on Twitter. For this purpose, eight predictive models were developed from an ad hoc generated training corpus, using shallow modeling and deep learning. The evaluation metrics showed that the deep learning algorithm performed significantly better than the shallow modeling algorithms, and logistic regression yielded the best performance of the shallow algorithms.

Keywords: Twitter; hate speech; gender discrimination; gender identity; sexual orientation; feminism; misogyny; machine learning; deep learning; supervised classification

Citation: Arcila-Calderón, C.; Amores, J.J.; Sánchez-Holgado, P.; Blanco-Herrero, D. Using Shallow and Deep Learning to Automatically Detect Hate Motivated by Gender and Sexual Orientation on Twitter in Spanish. *Multimodal Technol. Interact.* **2021**, *5*, 63. https://doi.org/10.3390/mti5100063

Academic Editors: Elisabetta Fersini, Sasa Arsovski and Derek L. Hansen

Received: 31 May 2021
Accepted: 22 September 2021
Published: 13 October 2021

1. Introduction

Although hate speech is not a new phenomenon, it appears to have become more concerning recently, due to its uncontrolled spread in the digital environment. The development and penetration of the internet and ICTs has allowed this new form of hate—online hate speech or cyberhate—to steadily increase. Social media appears to be the platform via which this online hate spreads in the most extensive and problematic manner, despite the recent efforts of technology companies to impose controls. In this regard, authors, such as Müller and Schwarz [1], have stated that there may be a direct correlation between the increase in the spread of hate speech through social media and the hate crimes committed in certain regions, based on specific types of discrimination. Hence, it is considered highly important to study and detect online hate speech, and thus to understand its spread, attempt to reduce it, and, most importantly, prevent and counteract its effects.

Among social media platforms, hate speech appears to have increased most obviously on Twitter. On this platform, messages that express hate, rejection, intolerance, or discrimination towards vulnerable groups have continued to increase, leading to a persistent polarization of public opinion. This increase in the spread of hate on social media platforms such as Twitter is shown by the latest reports from the Anti-Defamation League on online hate and harassment [2,3]. The data offered by the ADL indicate an exponential increase

in all forms of cyberhate on most social networks since 2018. In particular, these reports show that women and the LGTBIQ+ community are two of the groups most victimized by online hate crimes. This increase in online hate speech is also linked to a growing trend in the volume of hate crimes around Europe [4], which seems to support the connection observed by Müller and Schwarz [1].

Due to the recent increase in hate speech and the aforementioned implications, it is important to study new strategies for the detection and prevention of hate speech at the international level. This is also the case in Spain, where the increase in the online and offline expression of hate reflects the absence of an independent and articulated national strategy aimed at its prevention, despite the fact that the rest of Europe has been working on this issue for some time. This highlights the necessity of implementing new techniques to help in identifying and monitoring hate messages in the Spanish context—automatically and on a large scale—in order to prevent and combat them. Therefore, in September 2018, the government of Spain signed an institutional cooperation agreement with the General Council of the Judiciary and the Attorney General's Office to fight racism, xenophobia, LGBTI-phobia, and other types of intolerance, thus renewing the 2015 framework agreement that had become obsolete. In this legal context, private institutions are involved in significant efforts to detect and counter online hate speech. However, the increasing amount of data and information transmitted on the internet makes it very difficult to identify and block all hateful content. As a result, in a digital environment that is free of surveillance and regulation, the number of victims of online hate speech continues to grow. This is evidenced by the most recent RAXEN report [5], despite the fact that most incidents may be unrecorded.

In this context, there is an urgent need to develop new computational strategies for detecting the main types of online hate speech that are spread through social media in the Spanish language. According to the report on the evolution of hate crimes in Spain, published by the Ministry of the Interior [6], consistent with international trends, the greatest number of hate crimes registered each year in Spain relates to the categories of racism and xenophobia, followed by ideology, sexual orientation, gender identity, and discrimination based on sex and/or gender. Thus, several researchers, including the authors of this work, have begun to work specifically on the study, analysis, and detection of online hate based on racist and xenophobic reasons, the most concerning category of hate crime internationally [7,8], in addition to that related to ideological reasons [9]. However, few studies have focused on cyberhate aimed at women or LGTBIQ+ groups, and even fewer have focused on developing computational strategies that allow hate to be automatically detected in tweets in Spanish.

For these reasons, the objective of this study was to develop and evaluate a detector of hate speech on Twitter and in Spanish that was based specifically on gender and sexual orientation. The computational strategy was developed using natural language processing and supervised machine learning, and with the support of the Cloud Computing service of the Supercomputing Center of Castilla y León—Scayle—for the analysis and monitoring of the massive amounts of data collected. In addition to automatically performing detection, the detector is expected to acquire empirical knowledge about the type of cyberhate that is spread via Twitter, the communities at which it is aimed, the types of sources or profiles that are potential propagators of this hate, and, finally, how this type of speech may ultimately be related to hate crimes committed in particular regions that are motivated by the same forms of discrimination and directed towards the same vulnerable groups.

Online Hate Detection Based on Reasons of Gender and Sexual Orientation

One of the most widely accepted attempts to define hate speech was made by the Council of Europe, whose Recommendation No. R (97)20 [10] defines hate speech as "all forms of expression which spread, incite, promote, or justify racial hatred, xenophobia, anti-Semitism, or other forms of hatred based on intolerance". In its Recommendation No. 15 [11], the European Commission against Racism and Intolerance specified that hate

may be motivated "on the grounds of 'race', color, descent, national or ethnic origin, age, disability, language, religion or belief, sex, gender, gender identity, sexual orientation, and other personal characteristics or status". In a similar manner, Gagliardone et al. [12] included all forms of expression that directly or indirectly promote discriminatory acts.

In the context of Spain, the Ministry of the Interior [6] mentioned the following 11 categories of discrimination for classifying hate crimes towards vulnerable groups: 1. racism/xenophobia, 2. political ideology, 3. sexual orientation and gender identity, 4. religious beliefs and practices, 5. gender reasons, 6. disability, 7. antisemitism, 8. aporophobia, 9. anti-Roma sentiment, 10. age discrimination, and 11. discrimination based on illness. From these, the first three are the ones with the largest numbers of hate crimes in Spain, whereas gender is usually transversally present, and, in many European countries, it is placed in the fourth position.

Aside from the differentiation made in these and similar reports about the motivations behind hate crimes, the particularities of online hate speech as a particular type of hate crime should also be noted. As stated before, cyberhate [13] has the particularity of being produced and spread through online media, allowing faster, wider, and uncontrolled propagation. For this reason, the automated detection of online hate speech has been one of the most common lines of work in recent years, and there have been some interesting attempts, such as those of Malmasi and Zampieri [14] or Salminen et al. [15], who used computational techniques similar to those that will be employed in the present work. In the same line, one of the most relevant prototypes in the Spanish context is that of Pereira Kohatsu et al. [16].

These studies have some limitations because they used non-supervised approaches that were usually based on lexicon-based dictionaries, or used supervised classification with already-existing databases or ad hoc databases that were generated by only one coder—normally the main researcher—without control for reliability. Another limitation is their generic approach to hate, without focusing on specific discriminatory categories, which limits their efficacy and reliability, given that each type of hate usually has some associated concepts, features, and linguistic aspects [9]. With this in mind, and as already mentioned, the authors of this work have been working on computational strategies for detecting specific types of hate with training corpuses that were developed and validated ad hoc [7–9]; however, to date, the focus has not been on gender identity and sexual orientation.

Despite the fact that Hewitt et al. [17] and Ahluwalia et al. [18] mentioned problems when detecting misogynistic language on Twitter and other social media, the work of Anzovino et al. [19] offered promising results for the classification and detection of misogynistic tweets. The same applies for Şahi et al. [20], who dealt with the automated detection of hate speech against women on Twitter in Turkey, as well as for Basile et al. [21], who focused on the multilingual detection of hate speech against immigrants and women on Twitter. Fuchs and Schäfer [22] also used computational corpus linguistic methods to detect and analyze the use of abusive language against female politicians in Japan. As can be observed, this type of hate speech is often studied in combination with other dimensions; although they did not focus on detection, this also happened in the work of Southern and Harner [23], who observed a large presence of misogynistic and objectifying tweets among the messages sent to female MPs in the UK, or in the study by Gallego et al. [24], who studied the use of gender discriminatory arguments in the discourse about rejection, with respect to refugees. Finally, in the Spanish setting, despite their lesser use of computational methods, the studies of Núñez Puente and Fernández Romero [25], and Villar-Aguilés and Pecourt Gracia [26] should be mentioned, as they also dealt with misogynist and antifeminist attacks on Twitter.

The other element of our analysis, rejection based on sexual orientation or gender identity, has been a frequent object of study from the perspectives of sociology of [27], representation of [28], and violence towards these groups [29]. Nonetheless, there have not been many attempts to approach automated detection; one of the most relevant exceptions considered homosexuality as one of the multiple protected characteristics for the identifica-

tion of cyberhate on Twitter [30]. Without dealing with detection, Lingierdi et al. [31] used a lexicon-based method of semantic content analysis to study different groups of victims of hatred, including women and homosexuals.

As observed in some of the aforementioned literature, these two types of hatred are studied as two different, but strongly interconnected, categories. In fact, the work of Şahi et al. [20] specifically found samples of homophobic discourse within the hate speech aimed at women on Twitter, which highlighted this interconnection. Given the similarities between both categories, and with the goal of attaining a larger training corpus—not as broad and imprecise as the generic ones, but not so overfitted that it isolates the two interconnected categories—this work studied both categories together.

In this way, we intended to fill the gaps in the existing knowledge and methodologies about the detection of hate based on sex/gender, sexual orientation, and gender identity on Twitter. The novelty of this work can be found in the use of supervised machine learning techniques to generate specific databases, which were created ad hoc with real examples that were manually classified in a pairwise fashion and with complete intercoder agreement, to be used as a training corpus. Additionally, the generation of this training corpus, with specific examples of hatred based on gender and sexual orientation, will allow more reliable and adequate predictive models to be created. To develop this corpus, the manual classification of previously filtered and downloaded examples from Twitter's API was necessary, which is why the following research question is presented.

RQ1: What frequency and percentage of tweets that include hate based on gender and sexual orientation will be found after the manual classification of a previously filtered sample?

An additional novelty of this research is the use of deep learning to generate the final detection model, which was expected to offer significant advantages, in terms of performance, compared to traditional classification algorithms [32,33]. Specifically, recurrent neural networks were used. This allowed us to formulate the following research questions.

RQ2a: What machine learning algorithm offers the best performance for the generation of a predictive model that is able to detect hate speech based on gender and sexual orientation motivations on Twitter in Spanish?

RQ2b: Does deep learning offer a better performance than shallow modeling for the generation of predictive models that are able to detect hate speech based on motivations related to gender and sexual orientation on Twitter in Spanish?

2. Materials and Methods

This automatic detection strategy led to the development of a prototype based on data-intensive computing strategies, for which the supercomputing infrastructure of Castilla y León—Scayle—was used to implement natural language processing techniques and supervised machine learning. The work was divided into two main phases—one was dedicated to the creation of the training corpus, and the other was dedicated to the generation of predictive models.

2.1. Creation of the Training Corpus

This phase focused on the creation of ad hoc databases from reliable examples of tweets containing hatred based on gender and sexual orientation. These databases would serve as a training corpus with which predictive models would be generated that would finally allow the automatic and large-scale detection of hateful messages. This stage allowed the project to overcome the limitations of the previously developed prototypes that used dictionaries or general databases. The generation of the corpus was subdivided into a series of sub-stages, which are described below.

2.1.1. Creation of a Filter Dictionary and Downloading of the Filtered Tweets

In order to create the training corpus with real examples of hate, the first step was to locate these examples on Twitter and compile them. With this goal in mind, we created

a dictionary of words and word combinations that would allow us locate hate messages based on gender and sexual orientation on this social network. Bearing that in mind, a definition of what would be considered hate speech motivated by any form of sex/gender or sexual orientation was created, and a compilation of related derogatory terms, expressions, accounts, and hashtags was made. To make this compilation, generic keywords with which the potential victims of this type of hate were mentioned in some way (mainly women and LGTBIQ+ groups) were used. Later, these messages were manually classified according to whether they only referred to those audiences or if they also included hate.

At the same time, after the previous exploration stage, the definition of hate speech was broadened as much as possible to encompass the three levels of cyberhate collected in the taxonomy offered by Miró Llinares [34]—from offensive messages to those directly promoting physical violence. Thus, our classification included all types of hate for reasons of gender and sexual orientation that could constitute a crime, but also the subtler ones that, a priori, could be considered within the scope of freedom of expression. This decision was made in order to be able to detect all possible hateful, discriminatory, or rejective messages, including the subtler ones, as they could be the basis for stronger, more hateful, and more violent ones in later stages [7]. Furthermore, the most explicit hate was expected to be a minority in the Spanish context, which had already been determined in previous exploratory stages, as well as in other studies [9,34]. Finally, as absolute intercoder agreement was requested, it was expected that, in the manual classification process, the subtler messages and those that were not hateful enough would not result in adequate agreement and would be discarded; thus, only the most explicit hate messages would be selected.

Once the levels and types of hate speech to be identified had been defined and the terms and combinations of words representative of our target group had been compiled, a dictionary of words that would serve as a filter when downloading example tweets was developed. For this, the final selection of search terms was made according to the distinction made by Kalampokis, Tambouris, and Tarabanis [35]. This dictionary of hateful terms was developed ad hoc with the intention of being able to access tweets with a higher probability of containing hate for reasons of gender or sexual orientation, thus optimizing the tagging process. The filters were translated into computer language in order to be able to download the necessary number of messages from Twitter's API. Although a greater number of messages was downloaded, a final sample of 24,000 tweets was eventually collected and compiled into a database for manual sorting. Below is a fragment of the final script used to download and collect the filtered tweets, where the key terms—both derogatory and hateful—as well as the representative target groups can be found.

- *word = ['comerabo', 'comepolla', 'chupapolla', 'muerdealmohada', 'machirul', 'marimach', 'hermafrodit', 'travelo', 'gay\nvicio', 'gay\ndesviac', 'gay\ndesviad', 'gay\npervert', 'gay\npervers', 'gay\ndepravac', 'gay\ndepravad', 'gay\npromiscu', 'gay\nlibertin', 'gay\nenferm', 'gay\nput', 'gay\nmaldit', 'gay\nsuci', 'gay\npluma', 'gay\ntijer', 'gay\ncoño', 'gay\npotorro', 'gay\npierdeaceit', 'gay\nmierda', 'gay\nbasura', 'gay\ngentuza', 'gay\nasco', 'gay\nlacra', 'gay\nescoria', 'gay\ncontagi', 'gay\ndestroz', 'gay\nreventar', 'gay\nrevient', 'gay\nmata', 'gay\nextermin', 'maric\ndesviac', 'maric\ndesviad', 'maric\npervert', 'maric\npervers', 'maric\ndepravac', 'maric\ndepravad', 'maric\npromiscu', 'maric\nlibertin', 'maric\nenferm', 'maric\nput', 'maric\nmaldit', 'maric\nsuci', 'maric\npluma', 'maric\ntijer', 'maric\ncoño', 'maric\npotorro', 'maric\npierdeaceit', 'maric\nmierda', 'maric\nbasura', 'maric\ngentuza', 'maric\nasco', 'maric\nlacra', 'maric\nescoria', 'maric\ncontagi', 'maric\ndestroz', 'maric\nreventar', 'maric\nrevient', 'maric\nmata', 'maric\nextermin' 'mariq\ndesviac', 'mariq\ndesviad', 'mariq\npervert', 'mariq\npervers', 'mariq\ndepravac', 'mariq\ndepravad', 'mariq\npromiscu', 'mariq\nlibertin', 'mariq\nenferm', 'mariq\nput', 'mariq\nmaldit', 'mariq\nsuci', 'mariq\npluma', 'mariq\ntijer', 'mariq\ncoño', 'mariq\npotorro', 'mariq\npierdeaceit', 'mariq\nmierda', 'mariq\nbasura', 'mariq\ngentuza', 'mariq\nasco', 'mariq\nlacra', 'mariq\nescoria', 'mariq\ncontagi', 'mariq\ndestroz', 'mariq\nreventar', 'mariq\nrevient', 'mariq\nmata', 'mariq\nextermin', 'lesbi\ndesviac', 'lesbi\ndesviad', 'lesbi\npervert', 'lesbi\npervers', 'lesbi\ndepravac', 'lesbi\ndepravad', 'lesbi\npromiscu', 'lesbi\nlibertin', 'lesbi\n-*

enferm', 'lesbi\nput', 'lesbi\nmaldit', 'lesbi\nsuci', 'lesbi\npluma', 'lesbi\ntijer', 'lesbi\ncoño', 'lesbi\npotorro', 'lesbi\npierdeaceit', 'lesbi\nmierda', 'lesbi\nbasura', 'lesbi\ngentuza', 'lesbi\nasco', 'lesbi\nlacra', 'lesbi\nescoria', 'lesbi\ncontagi', 'lesbi\ndestroz', 'lesbi\nreventar', 'lesbi\nrevient', 'lesbi\nmata', 'lesbi\nextermin', 'trans\ndesviac', 'trans\ndesviad', 'trans\npervert', 'trans\npervers', 'trans\ndepravac', 'trans\ndepravad', 'trans\npromiscu', 'trans\nlibertin', 'trans\nenferm', 'trans\nput', 'trans\nmaldit', 'trans\nsuci', 'trans\npluma', 'trans\ntijer', 'trans\ncoño', 'trans\npotorro', 'trans\npierdeaceit', 'trans\nmierda', 'trans\nbasura', 'trans\ngentuza', 'trans\nasco', 'trans\nlacra', 'trans\nescoria', 'trans\ncontagi', 'trans\ndestroz', 'trans\nreventar', 'trans\nrevient', 'trans\nmata', 'trans\nextermin', 'drag\ndesviac', 'drag\ndesviad', 'drag\npervert', 'drag\npervers', 'drag\ndepravac', 'drag\ndepravad', 'drag\npromiscu', 'drag\nlibertin', 'drag\nenferm', 'drag\nput', 'drag\nmaldit', 'drag\nsuci', 'drag\npluma', 'drag\ntijer', 'drag\ncoño', 'drag\npotorro', 'drag\npierdeaceit', 'drag\nmierda', 'drag\nbasura', 'drag\ngentuza', 'drag\nasco', 'drag\nlacra', 'drag\nescoria', 'drag\ncontagi', 'drag\ndestroz', 'drag\nreventar', 'drag\nrevient', 'drag\nmata', 'drag\nextermin' (. . .)].

2.1.2. Manual Peer Classification

The filtered messages were manually classified by using the Doccano platform, which facilitated the task of tagging the texts with multiple coders. Thus, all the tweets were coded by the main coder, a project researcher, and one out of eight secondary coders, who classified sub-samples of 3000 tweets each. In order to be able to cross-check the results later and make the resulting messages more reliable, the secondary judges had to be outsiders of the project, so undergraduate and graduate students from the University of Salamanca were selected after being informed about and trained for the coding. During the classification process, messages were labeled in a binary fashion as "hate" and "not hate", but the main coder also had the possibility of discarding messages that were not valid for the corpus because they belonged to other categories of hate or to other groups, or because they applied to other themes, contexts, or settings from outside of Spain that could contaminate the final models.

2.1.3. Checking the Intercoder Agreement

Once all of the tweets were classified by both coders, the results were cross-examined to check the intercoder reliability, and only reliable tweets for which total agreement was attained were collected, only keeping messages that were classified with the same label by both coders and discarding the messages for which there was no such agreement, as well as those that were previously discarded in the classification process. This step, in addition to ensuring the quality of the coding, allowed one of the main limitations of some prototypes [16] to be overcome, as they used dictionaries or an ad hoc training corpus, but it was generic and developed by a single coder; therefore, it was conditioned by his/her subjective understanding of hate speech and cognitive biases.

2.1.4. Cleaning and Compilation of the Final Database

After the classification and the reliability check, the databases were cleaned, leading to a training corpus in which a total of 11.6% of the reliable hateful tweets (N = 2773) and 33.7% of the reliable non-hateful tweets (N = 8082) were included. The remaining tweets were discarded so that they could not contaminate the resulting sample. This distribution can be observed in Figure 1, where the frequencies and percentages of manually classified and rejected tweets are shown.

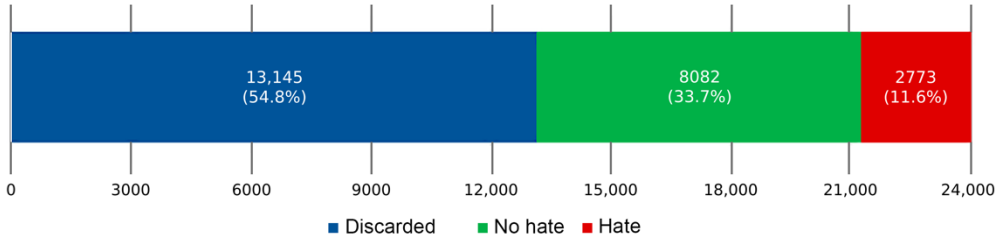

Figure 1. Frequencies and percentages of manually classified political tweets.

2.2. Generation of the Predictive Models

Once the validated training corpus was ready, the last step was to use it to train and generate predictive models that would finally allow the automatic and large-scale detection of hate speech in Spanish on Twitter for the reasons of gender and sexual orientation. The generated database was used to provide the algorithms with the necessary examples so that they could create rules that would lead to predictive classification models. Using the generated and validated training corpus, a total of eight predictive models were trained and generated. Six of them were generated by using shallow learning algorithms, another model was generated from the votes of the previous models, and a final model was generated by using deep learning.

2.2.1. Shallow Modeling

The six predictive models that were generated by using traditional classification algorithms were based on "Bag of Words" as a representation of the text, from which each word was taken as a vector. Python's NLTK and SciKit-Learn libraries were used to generate binary classification models by using the following conventional shallow learning algorithms: original Naïve Bayes, Naïve Bayes for multinomial models, Naïve Bayes for Bernoulli's multivariate models, Logistic Regression, Lineal classifiers with Stochastic Gradient Descent training, and Support Vector Machines. Natural language processing (NLP) techniques were also applied to extract the features of the tagged set of messages. In the process of training the models, the most frequently repeated words from the set of examples that made up the training corpus were tokenized and converted into quantitative features or vectors with which the predictive models could work. In this modeling process, the corpus was randomly divided into the following two sub-groups for each of the algorithms: 70% were dedicated to training, and 30% were dedicated to the testing and validation of the models. This way, optimized classifiers were generated for each of the six aforementioned algorithms, and they were implemented on the training corpus in order to generate six predictive models that were capable of detecting hate speech in tweets in Spanish for reasons of gender and sexual orientation. Once these models were developed, a final classifier was generated based on the vote of each of the six previous models. This classifier chose the category—hate/not hate—that the most models predicted, adding a confidence indicator based on the proportion of said agreement (the number of votes for the majority class/number of possible votes), which allowed the establishment of a confidence threshold that was greater than 80% (0.8) for each prediction. Finally, each of the six classifiers, as well as the one based on the voting of the other models, were evaluated by using the 30% of the corpus that was dedicated to testing so that we could compare the manual classification of that sample with the predictions produced by the models.

2.2.2. Deep Modeling

In addition to shallow modeling, a second strategy was developed for the classification of texts based on deep modeling, which made it possible to generate the eighth and last of the predictive models with the intention of improving the evaluation metrics of the previous models. This last model was generated by using embeddings as a form of representation of the text, as well as a recurrent neural network (RNN). Specifically, TensorFlow (v2) and the Keras environment were used to create a sequential model with the following four layers:

- The first input layer converted each word into embeddings, which were dense vectors that represented the categorical value of any given word. The embeddings were trained using the 10,000 most common words of the created vocabulary plus 1000 out-of-vocabulary buckets. Thus, the matrix of embeddings included one row for each of these 11,000 words and one column for each of the six embedding dimensions (this hyperparameter was tuned several times and obtained the best performance with size = 6).
- The second and the third hidden layers included were gated recurrent unit (GRU) layers with 128 neurons each. GRUs are simplified versions of traditional LSTM cells. Despite the fact that both perform quite well for text classification (converging quickly and detecting long-term dependencies), we decided to use GRUs instead of LSTM given that the simplified version performs as well as the original one.
- The last output layer consisted of a dense layer with just one neuron, and it used sigmoid activation to predict the probability of a message being a hateful message.

To compile the model, standard loss with binary cross-entropy and the Adam optimizer were used. Finally, the training corpus was fit for five epochs, and the test set was used for validation (30 steps). Because neural networks require a high computing capacity and there was a need to scale the processes from the local to the distributed, the evaluation of the deep model was carried out remotely in virtual computers by using the aforementioned computing services of the Castilla y León Supercomputing Center.

3. Results

To the best of the authors' knowledge, this work developed the first validated corpus and the first prototype for the automatic detection of hate speech that is motivated by gender and sexual orientation, and spread via Twitter in Spanish. As previously mentioned, six classification models were generated from shallow algorithms that were traditionally used for binary classification—with Bag of Words as the text representation—in addition to an extra model based on the votes of those models, and a model based on embeddings and deep learning. For the generation of the deep model, a recurrent neural network (RNN) was used, which, as observed in Figure 2, generally improved upon the evaluation metrics of the previous models. However, before reviewing the performance of each of the generated models, it is important to analyze the results of the manual classification carried out in the generation of the training corpus.

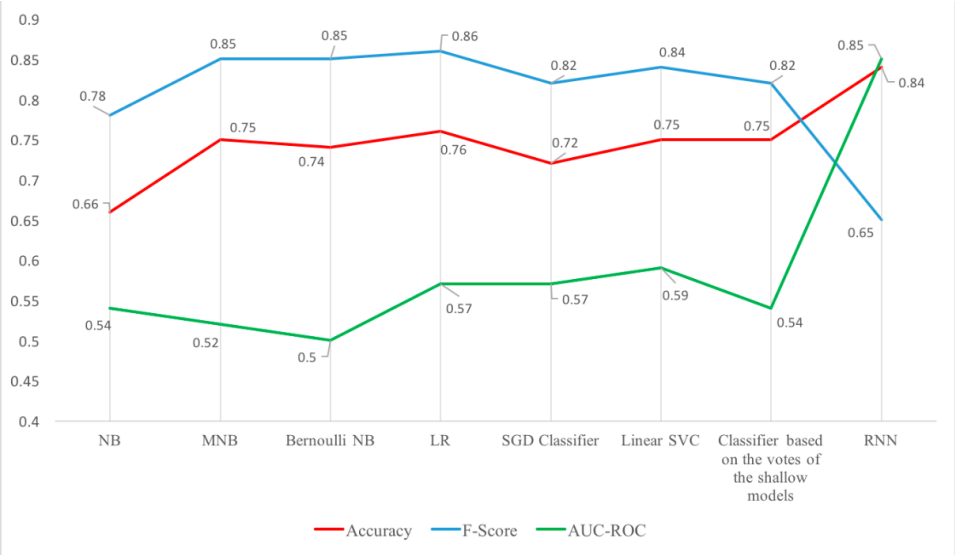

Figure 2. Evaluation metrics of the models generated with each of the algorithms.

It should be noted that the percentage of tweets that were finally discarded was high—a total of 54.8% (N = 13,145) of the sample, including those that did not have inter-coder agreement and those that were discarded in the classification process for not referring to the selected hate category or to the Spanish context. This shows that the percentage of hate tweets motivated by gender or sexual orientation, and validated with full agreement, was considerably reduced, despite the previous filtering of messages. Specifically, as an answer to RQ1, 11.6% of the sample was validated as hateful tweets (N = 2773), compared to 33.7% of the sample that comprised non-hateful messages (N = 8082). This means that, according to the validated messages that actually referred to the target groups in the Spanish context, 25.5% included hatred. Below, a series of examples of real tweets transmitting hate for reasons of gender and sexual orientation, which were validated and included in the final dataset, can be consulted. In the examples, we have highlighted the keywords that are indicative of the type of hate studied by making them bold; these terms found in these validated messages were part of the filter dictionary used for the collection of the tweets. Next to the examples is an English translation that is as literal as possible.

- *Vaya cara de* ***chupapollas*** *que tienes* ***perra*** (what a cocksucker face you have bitch);
- ***>Maricones*** *más* ***tóxicos*** *que Chernobyl no, gracias* (fags more toxic than Chernobyl, no thanks);
- *Yo veo a alguien que es* ***gay*** *un hombre que le da por* ***homofobico*** *o algo y lo corto en pedazos* (I see someone who is a gay, a man who thinks he is homophobic or something and I cut him to pieces);
- *Esos* ***maricones*** *de* ***mierda****, son* ***escorias*** (those shitty fags, they're slag);
- *Hermosa en tus sueños pedazo de* ***travesti*** *eres un* ***asco****!* (beautiful in your dreams piece of transvestite you are disgusting!);
- *Que* ***asco*** *son* ***lesbianas*** *ojalá las quemen como hacía Hitler* (how disgusting they are lesbians, I hope they burn them like Hitler did);
- *Que te haces la pacifista* ***puta*** *de* ***mierda travesti*** *política* (you play the pacifist political transvestite fucking whore);
- *Es un* ***comepollas*** *de primera... como su* ***mujer****, que le caben cuatro de golpe en la boca* (he is a first-rate cocksucker... like his wife, who can fit four at once in her mouth);
- *No me vayas a contagiar tu enfermedad* ***lesbiana asquerosa*** (don't give me your disgusting lesbian disease);
- ***Muerete puta Lesbiana*** *de* ***mierda*** (die fucking lesbian whore);
- *Deja de llorar puto* ***maricon*** *de* ***mierda*** (stop crying you fucking fag shit);
- ***Puta*** *por que le contabas a mi madre las* ***pollas*** *que te chupabas y como follabas con los tios eh* ***puta chupapollas*** *Hija de puta culera jodete* ***perra puta guarra*** (you are a whore because you told my mother about the cocks you sucked and how you fucked with the guys, eh whore cocksucker daughter of a bitch fuck you slut, whore, hooker);
- *Que no te pille por ahí* ***maricona*** *de* ***mierda*** *que te* ***reviento*** (better that I don't catch you out there, you fucking faggot I'll bust you);
- *Al punto que llegan estos Progres....realmente una vergüenza esta tropa de* ***Travestis*** (to the point that these progressives come... they really are a shame this troop of transvestites);
- *Tu lo q eres es una* ***puta marica*** *pasiva* (what you are is a bottom faggot whore).

First, the figures show that, no matter how complete and complex the linguistic filter dictionaries and the detection techniques based on expressions and keywords are, they do not offer an effective method for identifying online hate messages. Nonetheless, they served as an aid in order to limit and optimize the process because, without these linguistic filters, it would have been very difficult to find examples of hate for reasons of sex, gender, sexual orientation, or gender identity by using Twitter's API. Secondly, it could be concluded that, although it was notably present in one quarter of the validated sample, the amount of hate motivated by gender and sexual orientation that spreads through Twitter is not as predominant as would be expected, though it is usually noisier and more effective, and its potential harm should not be underestimated. Thirdly, comparing these results with those of the authors' previously developed prototype [9], which was focused on ideological

hate, it can be observed that a larger number of hateful tweets were discovered in that case, even though the sample size was the same (24,000 tweets). In the said prototype, based on political intolerance, more hateful tweets (3879) and fewer non-hateful ones (6334) were found; both cases had rather similar samples of discarded examples—13,145 in the case of hate motivated by gender and sexual orientation, and 13,787 in the case of political intolerance. This might indicate that hatred based on gender or sexual orientation might be less present or subtler than political intolerance—despite the fact that this type is slightly more difficult to identify with filter dictionaries—which is something reasonable in the current scenario of growing affective polarization [36].

Next, for the evaluation of the predictive models generated, we applied three of the most commonly used evaluation metrics in supervised machine learning, which are as follows: the accuracy; the harmonic mean; the F-score—which offers a balanced metric calculated from the precision and recall; and the AUC-ROC, which shows the performance of the classification models at all classification thresholds. All of the values produced by these evaluation metrics were acceptable—in most cases, they were above 0.70. When comparing the performance of each of the algorithms in the generation of these predictive models, it should be noted that the accuracy and AUC-ROC values were considerably higher in the model generated with the recurrent neural network, which confirmed the comparative advantage of the application of deep learning to the classification of texts. Thus, answering RQ2, it can be concluded that—focusing specifically on shallow modeling—the traditional classification algorithm that offers the best performance in this case is logistic regression, followed by Naïve Bayes for multinomial models and support vector machines. However, in general terms, it is the deep model—in this case, the model generated with the recurrent neural network—that seems to offer better performance than the models generated with shallow algorithms. Furthermore, as Figure 2 shows, although the F-Score decreased considerably in this model, both the accuracy and the AUC-ROC had significantly better coefficients than those of any of the shallow algorithms.

4. Discussion and Conclusions

This paper presents the first prototype for the automatic detection of hate speech on Twitter in Spanish that is specifically motivated by gender and sexual orientation; this prototype was modeled by using a manually and pairwise-generated ad hoc dataset, and by making use of machine learning and deep learning algorithms, thus improving upon and complementing previously developed prototypes. The main techniques used for the development of this prototype were natural language processing for the analysis and processing of unstructured data, and the classification of texts with supervised machine learning. This work confirmed that it is possible to train predictive models that allow the detection of hate speech on Twitter based on gender and/or sexual orientation, which also makes it possible to more precisely narrow down and specify the training of the models, leading to a solid performance and acceptable precision. In addition, a specific database was created for the training of predictive models, thus making it possible to improve the reliability of the detector when applied to this specific context, and overcoming the possible internal validity problems of previous prototypes. In this regard, it should be noted that, although the final percentages of hate and non-hate messages may seem small according to the training corpus, the most important thing in this process is having quality examples, not just finding large corpuses. The training corpus provided can always be updated (which is convenient given the constant evolution of language and social factors) with more reliable examples of this kind of hate. For this reason, the main effort of this work was focused on generating an initial reliable and validated corpus.

It was observed that, of the six machine learning algorithms used for shallow modeling, the one that offered the best performance was logistic regression, followed by Naïve Bayes for multivariate models. However, in general terms, it was verified that deep learning worked considerably better than conventional classification algorithms for the detection of this type of hate speech on Twitter, although the F-Score metric was lower than that in

the shallow models. We admit this limitation, which is due to the fact that the precision, which indicates the percentage of positive predictions that were correct, did not exhibit the performance that was expected, and we will continue to improve the training corpus, in order to obtain a detector with better precision and recall. However, this is not very serious because we understand that it can be improved. Thus, observing that the RNN showed better results for the rest of the metrics, we consider deep learning to definitely be better than shallow learning for this kind of detection. In any case, future studies should conduct an external validation to confirm the performance of this type of deep modeling, and, if necessary, to expand and improve the training corpus to improve the evaluation metrics. This could also help in overcoming the main limitation of this study, which is the collection of a limited—but large—sample of messages from a particular moment in time.

Aside from the methodological aspects, the manual classification developed to generate the corpus allowed us to observe the notable presence of hate speech in the dataset of filtered tweets; this presence was found in 11.6% of the total sample, and 25.5% of the tweets that were classified with agreement and selected for the corpus. This makes it possible to contribute to theoretical discussions on the definition and taxonomy of hate speech [34]; on the limits to freedom of expression [13]; on the different forms of hate motivated by gender, sexual orientation, and gender identity [29]; and on the implications of these forms of verbal violence for their victims [1], as well as on the possible quantification of hate speech on social platforms such as Twitter. The tasks of detecting and quantifying this type of hate, which are especially complex due to the volatility of speech [7], can benefit from a validated and specific tool, such as this one, so that specific types of hate can be measured in different periods of time, thus helping in measuring their evolution.

In short, it can be stated that this work presents a methodological contribution in the form of the large-scale detection strategy, the generation of an ad hoc training corpus, and the models that were developed with supervised machine learning techniques; it also presents a theoretical advancement in the study of hate crimes and, specifically, hate speech on Twitter for reasons of gender and sexual orientation, as well as a practical application because the technology developed can be implemented in different institutions such as government agencies, private companies, consultancies, research groups, and non-profit organizations will be able to benefit from it. This last aspect is the most relevant, due to this work's potential application by social networks themselves, as well as by public, private, or third-sector institutions, including the media and even political parties, to locate and reduce the presence of hate. Ultimately, it is hoped that this tool can help all of these actors to make decisions based on data that allow more effective combating and countering of this type of hate, thus promoting less radicalized and polarized spaces and more social networks.

Author Contributions: Conceptualization, C.A.-C., J.J.A., P.S.-H. and D.B.-H.; methodology, C.A.-C., J.J.A., P.S.-H. and D.B.-H.; software, C.A.-C., J.J.A., P.S.-H. and D.B.-H.; validation, C.A.-C., J.J.A., P.S.-H. and D.B.-H.; formal analysis, C.A.-C., J.J.A., P.S.-H. and D.B.-H.; investigation, C.A.-C., J.J.A., P.S.-H. and D.B.-H.; resources, C.A.-C., J.J.A., P.S.-H. and D.B.-H.; data curation, C.A.-C., J.J.A., P.S.-H. and D.B.-H.; writing—original draft preparation, C.A.-C., J.J.A., P.S.-H. and D.B.-H.; writing—review and editing, C.A.-C., J.J.A., P.S.-H. and D.B.-H.; visualization, C.A.-C., J.J.A., P.S.-H. and D.B.-H.; supervision, C.A.-C., J.J.A., P.S.-H. and D.B.-H.; project administration, C.A.-C., J.J.A., P.S.-H. and D.B.-H.; funding acquisition, C.A.-C., J.J.A., P.S.-H. and D.B.-H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Regional Development European Fund and the Junta de Castilla y León via the TCUE plan of the Fundación General de la Universidad de Salamanca, reference PC-TCUE_18-20_016, and also by the European Union, within the Rights, Equality and Citizenship programme REC-RRAC-RACI-AG-2019 (GA n. 875217).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data will be available at https://github.com/carlosarcila/hate_gender_LGTB.

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

References

1. Müller, K.; Schwarz, C. Fanning the Flames of Hate: Social Media and Hate Crime. *J. Eur. Econ. Assoc.* **2021**, *19*, 2131–2167. [CrossRef]
2. Anti-Defamation League. *Online Hate and Harassment. The American Experience 2020*. The ADL Center for Technology and Society: 2020. Available online: https://www.adl.org/media/14643/download (accessed on 15 February 2021).
3. Anti-Defamation League. *Online Hate and Harassment. The American Experience 2021*. The ADL Center for Technology and Society: 2021. Available online: https://www.adl.org/media/16033/download (accessed on 29 May 2021).
4. Hate Crime Reporting. Available online: https://hatecrime.osce.org (accessed on 31 May 2021).
5. Movimiento contra la Intolerancia. *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*; Movimiento contra la Intolerancia: Madrid, Spain, 2019.
6. Ministerio del Interior de España. Informe de Evolución de los Delitos de Odio en España. 2020. Available online: http://www.interior.gob.es/documents/642012/3479677/Informe+sobre+la+evolución+de+delitos+de+odio+en+Espa~na%2C%20a~no+2019/344089ef-15e6-4a7b-8925-f2b64c117a0a (accessed on 6 February 2021).
7. Arcila-Calderón, C.; Blanco-Herrero, D.; Valdez-Apolo, M.B. Rechazo y discurso de odio en Twitter: Análisis de contenido de los tuits sobre migrantes y refugiados en español. *REIS Rev. Española Investig. Sociológicas* **2020**, *172*, 21–40. [CrossRef]
8. Valdez-Apolo, M.B.; Arcila-Calderón, C.; Amores, J.J. El discurso del odio hacia migrantes y refugiados a través del tono y los marcos de los mensajes en Twitter. *Rev. Asoc. Española Investig. Comun.* **2019**, *6*, 361–384. [CrossRef]
9. Amores, J.J.; Blanco-Herrero, D.; Sánchez-Holgado, P.; Frías-Vázquez, M. 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* **2021**, *49*, 98–124. [CrossRef]
10. Council of Europe. Recommendation No. R (97) 20 of the Committee of Ministers to Member States on "Hate Speech". Council of Europe: Strasbourg, France, 1997.
11. European Commission against Racism and Intolerance. *ECRI General Policy Recommendation N° 15 on Combating Hate Speech*; Council of Europe: Strasbourg, France, 2016.
12. Gagliardone, I.; Gal, D.; Alves, T.; Martinez, G. *Countering Online Hate Speech*; Unesco Publishing: Paris, France, 2015.
13. Moretón Toquero, M.A. El «ciberodio», la nueva cara del mensaje de odio: Entre la cibercriminalidad y la libertad de expresión. *Rev. Jurídica Castilla León* **2012**, *27*. Available online: https://dialnet.unirioja.es/servlet/articulo?codigo=4224783 (accessed on 12 December 2020).
14. Malmasi, S.; Zampieri, M. Detecting hate speech in social media. *arXiv* **2017**, arXiv:1712.06427. Available online: https://arxiv.org/abs/1712.06427 (accessed on 8 November 2020).
15. Salminen, J.; Hopf, M.; Chowdhury, S.A.; Jung, S.G.; Almerekhi, H.; Jansen, B.J. Developing an online hate classifier for multiple social media platforms. *Hum. -Cent. Comput. Inf. Sci.* **2020**, *10*, 1–34. [CrossRef]
16. Pereira-Kohatsu, J.C.; Quijano-Sánchez, L.; Liberatore, F.; Camacho-Collados, M. Detecting and monitoring hate speech in Twitter. *Sensors* **2019**, *19*, 4654. [CrossRef]
17. Hewitt, S.; Tiropanis, T.; Bokhove, C. The Problem of Identifying Misogynist Language on Twitter (and Other Online Social Spaces). In *WebSci'16. Proceedings of the 2016 ACM Web Science Conference*; Nejdl, W., Hall, W., Parigi, P., Staab, S., Eds.; ACM: New York, NY, USA, 2016.
18. Ahluwalia, R.; Shcherbinina, E.; Callow, E.; Anderson, C.; Nascimento, A.; De Cock, M. Detecting Misogynous Tweets. In Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval), Sevilla, Spain, 18 September 2018.
19. Anzovino, M.; Fersini, E.; Rosso, P. Automatic Identification and Classification of Misogynistic Language on Twitter. In *Natural Language Processing and Information Systems. NLDB 2018. Lecture Notes in Computer Science*; Silberztein, M., Atigui, F., Kornyshova, E., Métais, E., Meziane, F., Eds.; Springer: Berlin/Heidelberg, Germany, 2018; Volume 10859. [CrossRef]
20. Şahi, H.; Kılıç, Y.; Sağlam, R.B. Automated Detection of Hate Speech towards Woman on Twitter. In Proceedings of the 3rd International Conference on Computer Science and Engineering (UBMK), Sarajevo, Bosnia and Herzegovina, 20–23 September 2018; pp. 533–536. [CrossRef]
21. Basile, V.; Bosco, C.; Fersini, E.; Nozza, D.; Patti, V.; Rangel, F.; Rosso, P.; Sanguinetti, M. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 6 June 2019; pp. 54–63.
22. Fuchs, T.; Schäfer, F. Normalizing misogyny: Hate speech and verbal abuse of female politicians on Japanese Twitter. *Jpn. Forum* **2020**. [CrossRef]
23. Southern, R.; Harmer, E. Twitter, Incivility and "Everyday" Gendered Othering: An Analysis of Tweets Sent to UK Members of Parliament. *Soc. Sci. Comput. Rev.* **2019**, *39*, 259–275. [CrossRef]
24. Gallego, M.; Gualda, E.; Rebollo, C. Women and Refugees in Twitter: Rhetorics on Abuse, Vulnerability and Violence from a Gender Perspective. *J. Mediterr. Knowl.* **2017**, *2*, 37–58. [CrossRef]

25. Núñoz Puente, S.; Fernández Romero, D. Posverdad y victimización en Twitter ante el caso de La Manada: Propuesta de un marco analítico a partir del testimonio ético. *Investig. Fem.* **2019**, *10*, 385–398. [CrossRef]
26. Villar-Aguilés, A.; Pecourt Gracia, J. Antifeminismo y troleo de género en Twitter. Estudio de la subcultura trol a través de #STOPfeminazis. *Teknocultura* **2021**, *18*, 33–44.
27. Alden, H.L.; Parker, K.F. Gender role ideology, homophobia and hate crime: Linking attitudes to macro-level anti-gay and lesbian hate crimes. *Deviant Behav.* **2005**, *26*, 321–343. [CrossRef]
28. Carratalá, A. Audiencias críticas en Twitter frente a coberturas tránsfobas: La identidad de género como nuevo derecho y su tratamiento periodístico. In *Más Sobre Periodismo y Derechos Humanos Emergentes*; Gómez y Méndez, J.M., Turón-Padial, M.C., Cartes Barroso, M.J., Eds.; Universidad de Sevilla: Sevilla, Spain, 2020; pp. 64–78.
29. Blondeel, K.; De Vasconcelos, S.; García-Moreno, C.; Stephenson, R.; Temmerman, M.; Toskin, I. Violence motivated by perception of sexual orientation and gender identity: A systematic review. *Bull. World Health Organ.* **2018**, *96*, 29–41. [CrossRef]
30. Burnap, P.; Williams, M.L. Us and them: Identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Sci.* **2016**, *5*, 1–15. [CrossRef]
31. Lingiardi, V.; Carone, N.; Semeraro, G.; Musto, C.; D'Amico, M.; Brena, S. Mapping Twitter hate speech towards social and sexual minorities: A lexicon-based approach to semantic content analysis. *Behav. Inf. Technol.* **2020**, *39*, 711–721. [CrossRef]
32. Cai, J.; Li, J.; Li, W.; Wang, J. Deep learning model used in text classification. In Proceedings of the 15th International Computer Conference on Wavelet Active Media Technology and Information Proccessing, Chengdu, China, 14–16 December 2018; pp. 123–126.
33. Li, Q.; Peng, H.; Li, J.; Xia, C.; Yang, R.; Sun, L.; Yu, P.S.; He, L. A survey on text classification: From shallow to deep learning. *arXiv* **2020**, arXiv:2008.00364.
34. Miró-Llinares, F. Taxonomía de la comunicación violenta y el discurso del odio en Internet. *IDP Rev. Internet Derecho Política* **2016**, *22*, 82–107. [CrossRef]
35. Kalampokis, E.; Tambouris, E.; Tarabanis, K. Understanding the predictive power of social media. *Internet Res.* **2013**, *23*, 544–559. [CrossRef]
36. Martini, S.; Torcal, M. Trust across political conflicts: Evidence from a survey experiment in divided societies. *Party Politics* **2016**, *25*, 126–139. [CrossRef]