applied sciences

MDPI

Article

A Language Model for Misogyny Detection in Latin American Spanish Driven by Multisource Feature Extraction and Transformers

Edwin Aldana-Bobadilla [1,2], Alejandro Molina-Villegas [1,3,*], Yuridia Montelongo-Padilla [2], Ivan Lopez-Arevalo [2] and Oscar S. Sordia [3]

[1] CONACYT, Mexico City 03940, Mexico; edwyn.aldana@cinvestav.mx
[2] Centro de Investigación y de Estudios Avanzados del I.P.N., Unidad Tamaulipas, Victoria 87130, Mexico; yuridia.montelongo@cinvestav.mx (Y.M.-P.); ilopez@cinvestav.mx (I.L.-A.)
[3] Centro de Investigación en Ciencias de Información Geoespacial, Merida 97302, Mexico; osanchez@centrogeo.edu.mx
* Correspondence: amolina@centrogeo.edu.mx

Abstract: Creating effective mechanisms to detect misogyny online automatically represents significant scientific and technological challenges. The complexity of recognizing misogyny through computer models lies in the fact that it is a subtle type of violence, it is not always explicitly aggressive, and it can even hide behind seemingly flattering words, jokes, parodies, and other expressions. Currently, it is even difficult to have an exact figure for the rate of misogynistic comments online because, unlike other types of violence, such as physical violence, these events are not registered by any statistical systems. This research contributes to the development of models for the automatic detection of misogynistic texts in Latin American Spanish and contributes to the design of data augmentation methodologies since the amount of data required for deep learning models is considerable.

Keywords: automatic hate speech detection; multisource feature extraction; Latin American Spanish language models; natural language processing

Citation: Aldana-Bobadilla, E.; Molina-Villegas, A.; Montelongo-Padilla, Y.; Lopez-Arevalo, I. and S. Sordia, O. A Language Model for Misogyny Detection in Latin American Spanish Driven by Multisource Feature Extraction and Transformers. *Appl. Sci.* **2021**, *11*, 10467. https://doi.org/10.3390/app112110467

Academic Editor: Arturo Montejo-Ráez

Received: 21 September 2021
Accepted: 2 November 2021
Published: 8 November 2021

1. Introduction

According to a recent report released by the World Health Organization, "*Physical or sexual violence is a public health problem that affects more than one third of all women globally*" [1]. Nevertheless, the problem seems even more prominent in Latin America when looking at the regional data. For instance, the regional prevalence rate for sexual violence among all women older than 15 years is 36.1% for the Americas region and 27.2% for Europe [2]. The Commission on the Status of Women (CSW), one of America's leading promoters of women's human rights, have been covering issues related to women's social and economic rights and, very recently, online violence. The sixty-fifth session of the CSW revolved around the theme of "*The participation of women in public life and the elimination of violence*" [3] in response to an increasingly online, gender-based abuse, cyberbullying, and sexual harassment. Out of all the recommendations to prevent and eliminate violence against women in public life (https://undocs.org/E/CN.6/2021/3, accessed on 2 November 2021, Par. 65), we can highlight the following three given their relationship with online violence against women (the original labels are used):

(i) reform legal frameworks to criminalize violence against women in political and public life, both online and offline, and to end impunity;

(o) set standards on what constitutes online violence against women in public life so that the media and companies running social media platforms can be held accountable for such content; and

(p) increase the capacity of national statistical systems to collect data regularly and systematically (both online and offline) on violence against women in public life.

It is of great importance that online violence against women is included in the recommendations above since many authors have considered that subtle violence can stratify to more severe violence. Johan Galtung, a renowned Norwegian pacifist and sociologist, assures that "*Cultural violence makes direct and structural violence appear, and even perceived, as charged with reason—or at least not bad*" [4]. Thus, for Galtung, hate speech, such as misogyny, precisely represents expressions of cultural violence because, through language, the misogynistic expressions legitimize and naturalize rejection and contempt towards women. Similarly, for Michel Foucault [5], discursive practices produce effects on the world, so that hate speech not only involves violence in itself but also implies the risk of generating direct violence on disadvantaged groups in addition to the fact that makes structural violence invisible.

However, creating effective mechanisms to detect misogyny online automatically represents significant scientific and technological challenges. The complexity of recognizing misogyny through computer models lies in the fact that it is a subtle type of violence, it is not always explicitly aggressive, and it can even hide behind seemingly flattering words, jokes, parodies, and other expressions (see Reference [6]). Currently, it is even difficult to have an exact figure for the rate of misogynistic comments online because, unlike other types of violence, —such as physical violence—, these events are not registered by any statistical systems.

Given this scenario, recent efforts to quantify and visualize the incidence of hate speech in digital media have recently been made mainly by the Natural Language Processing community, as is described in the Related Work Section 2.

Our research contributes to the development of models for the automatic detection of hate speech, particularly misogynistic texts, and to the design of Spanish data augmentation methodologies (since the amount of data required for deep learning models is considerable). However, in addition to the scientific contribution, we have the goal of doing science with social relevance. We seek raise awareness about the proliferation of misogyny in social networks in Latin America.

2. Related Work

Several recent studies evidence the growing interest of the scientific community on automatic detection of hate speech, mainly for English [7–12]. This research area has grown mainly thanks to the competitions organized at SemEval [13] (e.g., HatEval, OffensEval, and Toxic Spans Detection) and other venues, such as TRAC [14] and HASOC [15]. These competitions are essential since they provided participants with widely used benchmark datasets (e.g., OLID [16]). Regarding aggressiveness detection for Latin American Spanish, the most relevant competition is MEX-A3T track at IberLEF 2019 [17], where the organizers considered two tasks focused on the authorship and aggressiveness in Mexican tweets, and IberEval 2018 [18], with the first shared task on Automatic Misogyny Identification.

Two very notorious aspects emerge from the state of the art on detection of hate speech: the target language defines the degree of maturity of the existing models and the target group to which the hate speech is directed defines the specific challenges. Regarding the first aspect, there are several research in different languages, most of them including data compilation: German [19,20]; French [21]; Danish[22]; Greek [23]; Italian [24]; Hindi [20]; Arabic [21,25]; Indonesian [26]; Polish [27]; Turkish [28]; and Spanish [29]. However, the lack of language-specific corpora for all possible languages and variants have being created an important gap between the research maturity and results in English face to other languages but also had motivated innovation in research to deal with this challenge. To cope with data scarcity, researchers have explored different solutions, such as feature engineering, data augmentation, and multilingual models [30,31].

The other aspect is that, in hate speech, there are few specific groups to which the attacks are systematically directed (women and immigrants, for instance [32]). This is why talking about hate speech is still generally in the context of the state of the art on automatic misogyny detection. In this sense, below, we will pay greater attention to the state of the art in the specific task of detecting and/or classifying misogynistic language.

The authors of Reference [33] present an experimental analysis using different NLP features and ML models to detect misogynous tweets in English labeled from different perspectives and an exploratory investigation using NLP features and ML models to detect and classify misogynistic language. Several ML models from scikit-learn were used: Linear Support Vector Machine (SVM), Random Forest (RF), Naïve Bayes (NB), and Multi-layer Perceptron Neural Network (MPNN). The best reported model, SVM, arises an accuracy of 0.7995. In addition, for automatic identification of misogynistic language in English, in Reference [34], the authors propose a Long Short-Term Memory (LSTM) classifier using a pretrained LSTM-based Language Model to build an accurate classification model with a small training set. A "Bayesian interpretation" of Transfer Learning is presented as a regularization technique to estimate the uncertainty in the pre-training. The method is relevant since misogynistic tweet is a highly unbalanced class against general tweets, so the regularization proposal avoids overfitting. The best model arises the following scores: accuracy 0.846, precision 0.806, and F1-score 0.781. In Reference [35], the authors present an exploratory work detecting Misogyny for English and Spanish using the IberEval 2018 data [18]. They test different ML classifiers obtaining the best result using an ensemble technique (majority voting) to combine the predictions of SVM, Random Forest, and Gradient Boosting classifiers. The best model arises accuracy 87.05 for English and 81.35 for Spanish. To the best of our knowledge, the work presented in Reference [36] is the only that compiled messages harassing women in Spanish from Latin America. The proposal created the MisoCorpus-2020, a balanced corpus regarding misogyny in Spanish. The authors also present models combining word embeddings and linguistic features for three ML classifiers: Random Forest (RF), a decision tree classifier, Sequential Minimal Optimization (SMO), and a Support Vector Machine achieving the best accuracy of 85.17%.

In a deeper research, Fulper et al. [37] explored whether social media can be used as an indicator of sexual violence in the U.S., by tracking misogynistic tweets. Using the FBI Uniform Crime Reports provides rape statistics in the U.S. at the state level and a 10% sample of the Twitter stream produced during 2012. The authors manually compiled a list of 90 terms that are commonly used as misogynistic insults. With such filters, they obtained georeferenced tweets that contain misogynistic language and location (either latitude/longitude or a free-form location string) and mapped them to states, using state boundary data from the U.S. Census Bureau. The final dataset contains roughly 170 million georeferenced tweets, of which 1.2 million contain misogynistic language. As a result, the authors found a significant association between tweets that contain misogynistic language and rape crime statistics for each state in the U.S. A similar project is presented in [38], where the author delves into the relationship between the rate of misogynistic tweets and the rate of femicides in Mexico. Data consisted of femicide reports from the Executive Secretariat of the National Public Security System of México (SESNSP) and about twelve million georeferenced tweets in 2017–2018. Some regions were found to have particularly high rates of both misogyny and femicides. Furthermore, the Spearman correlation coefficient between both variables is 0.2515 with a significance level of 0.16; in other words, there is an interdependence, although very low, between both indicators, but the risk of concluding that there is a correlation, when, in reality, there is not, is only 16%.

3. Misogyny Detection Approach

The first challenge in recognizing written misogyny is obtaining a representative dataset containing a wide set of examples of what may or may not be a manifestation of violence. Another challenge that arises is the complexity of extracting the convenient features of the text from which is possible to create computational models able to recognize manifestations of written violence. As mentioned before, the complexity lies in the fact that the written violence is subtle; it is not always explicitly aggressive. Guided by these challenges, we have designed an integrated proposal consisting of several techniques, from gathering data to training a model capable of recognizing misogynistic manifestations.

Our proposal includes three main stages: *Gathering*, *Feature Extraction*, and *Modeling*, illustrated in Figures 1 and 2 described in the subsequent sections.

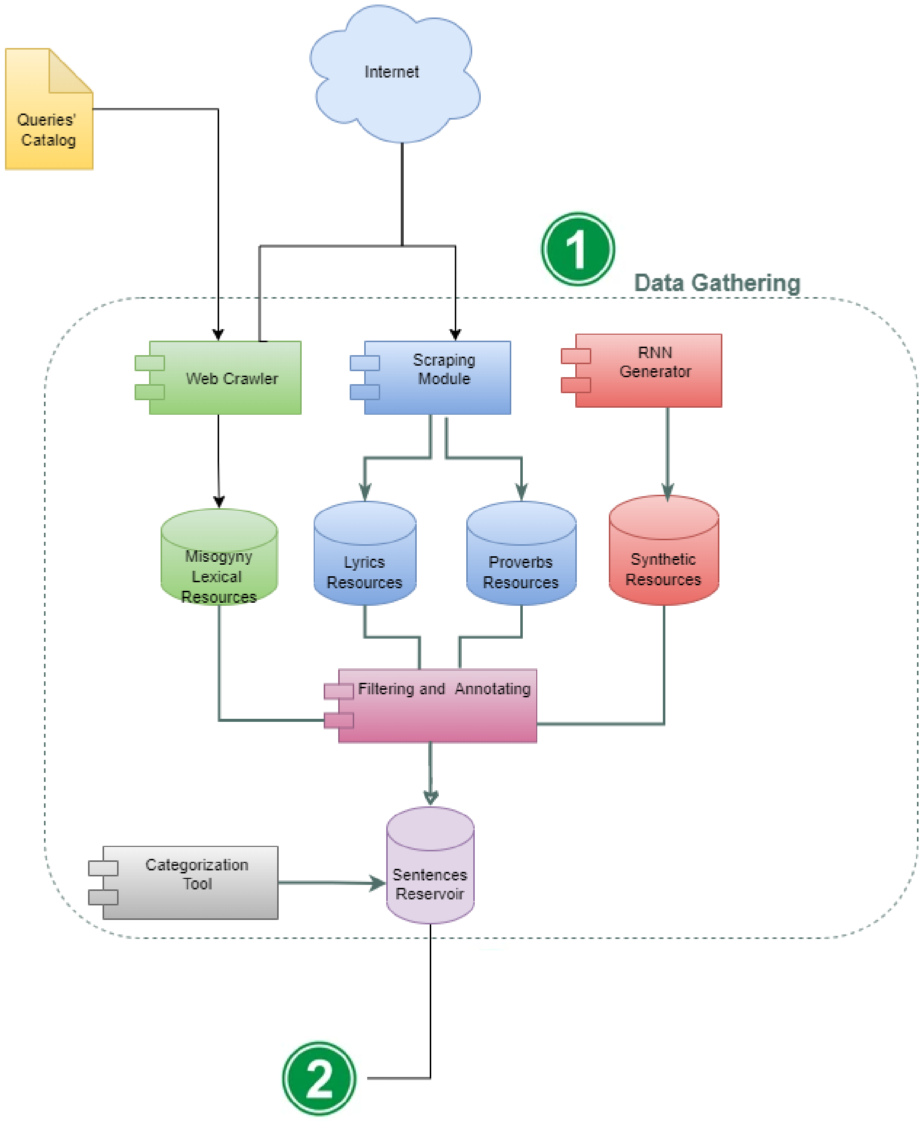

Figure 1. Overview of data gathering stage. As result of this stage, there is a corpus containing annotated sentences which will be the input of the subsequent stage (Feature Extraction).

3.1. Data Gathering

This stage comprises a set of components that allow us to obtain an appropriate set of documents for our purpose. The components of Gathering, illustrated inside the upper box in Figure 1, are:

- Web Crawler: It allows us to seek and obtain documents (in HTML and PDF format) containing misogynistic expressions. The search of documents is guided by a set of what we have called *Queries Catalog*. This catalog contains 64 sentences in Spanish made up of key words allusive to misogyny, that intend to focus the search on those documents containing misogynistic elements. For example, sentences of the form: *comportamiento misógino (misogynistic behavior)*, *discriminación y violencia contra la mujer (discrimination and violence against women)*, *chistes misóginos (misogynistic jokes)*, *misoginia en la política (misogyny in politics)*, and so on. Additionally, the catalog also contains a set of *n*-grams that frequently appeared in text with misogynistic bias, according to a preliminary study reported in [38]. For example, the *n*-gram *eres una puta (you are a whore*) and *malditas feminazis (fucking feminazis*), among others. From the described catalog, the web crawler could find an initial set of 991 documents of different length containing text in Spanish with a high probability of having misogynistic expressions.

- Scraping Module: Unlike the web crawling, scraping is a process that allows us to obtain the content of prior identified resources. Relying on the assumption that the text of several songs could be a suitable resource to find misogynistic expressions (an interesting study is reported in Reference [39]), we focus on identifying lyrics in Spanish singled out as resources with sexist content and violence against women, under the perception of some group of people. In this regard, we use a set of keywords on the search engines of Google, Facebook, Twitter, and YouTube to identify those songs that are commonly associated with misogynistic content. From the results of the searches, we build a catalog of 163 titles of songs with a high probability of including valuable sentences for our purposes; this resource is available at http://shorturl.at/lptzT (accessed on 4 November 2021). Relying on the catalog's titles, the Scraping Module is executed in order to obtain their corresponding lyrics, from the website https://www.letras.com (accessed on 4 November 2021). We also configured this module to scrap and filter documents available at https://proverbia.net (accessed on 4 November 2021), which are short documents (proverbs) just containing expressions that people often quote for giving advice or some philosophical reflection. Those proverbs with misogynistic content were not considered. It is worth mentioning that all the texts, including the proverbs, were manually revised to avoid including misogynistic phrases.
- RNN Generator: Although the above datasets allowed us to obtain an large number of documents, it was insufficient to encompass the study phenomenon. For this reason, our proposal includes a strategy to overcome the lack of data, based on a Recurrent Neural Network (RNN) capable of learning the intrinsic semantic of misogynistic expressions contained within the collected lyrics and generating documents that contain synthetic text. The length of the generated text is determined by a parameter corresponding to the number of words desired. For purposes of our work, we set this parameter to a constant value of 300. Since the quality of the generated text is much lower than that of the lyrics text, a lot of generated documents could not contain valuable sentences to be considered in our corpus. Despite this, we achieved to obtain a valuable set of sentences by exhaustive manual inspection (see Table 1).
- Other components: At this point, we take the collected documents as input to execute the module that we have called *Filtering and Annotating*: Filtering is the process by which expressions and sentences (in general) are extracted from the text of each document; this process was manual and did depend on the criteria from who executes it to define what is a sentence. From this, a set of 7191 "raw sentences" was obtained. Relying on these sentences, we executed an annotation process in which each sentence was annotated by 2 independent annotators judging the presence or absence of misogynistic content (binary decision). Of the 7191 sentences, we obtained 6747 agreements (93.84%) and 3624.8 agreements expected by chance (50.41% of observations) resulting in a kappa value of 0.87 indicating a suitable agreement. For those sentences where the annotators did not agree, there was a third annotator to judge and make a final decision based on the mode of the three annotations. We try to keep the annotation guideline as simple as possible, so that annotators could easily make a decision. The guideline included only three rules regarding a misogynistic sentence:
  - Unigrams sentences containing rudeness that could be directed at a woman. For instance, *puta (whore)*, *fea (ugly)*, *gorda (fat)*, *tonta (silly)*.
  - Any sentence containing one or more rudeness and violent language explicitly directed at a woman or a group of them. For example, *me engañaste pinche*

alcohólica (you cheated on me, fucking alcoholic), *te odio pinche zorra (I hate you fucking bitch)*, etc.

– Any phrase that, in the judgment of the annotator, indicates explicitly or implicitly submission, inferiority or violence, but it does not necessarily include rudeness. For example, *ellas también tiene que respetarse, se visten así y luego se quejan cuando les pasa algo (they have to respect themselves, they dress that way and then complain when something happens to them)*, *vete a la cocina y prepararme un sándwich (go to the kitchen and make me a sandwich)*.

In Table 1, a summary that illustrates the number of documents and annotated sentences gathered via the described components is shown. The documents are of different length; the longest documents were those obtained via the web crawler with a length in the range of 476 to 2492 words. The length of the lyrics is in the range of 80 to 300 words, while the length of the proverbs is in the range of 3 to 50 words.

Table 1. Documents and annotated sentences via data gathering components.

Component	Documents	Sentences
Web Crawler	991	3310
Scraping (Lyrics)	200	733
Scraping (Proverbia)	2196	2196
RNN Generator	1000	952
	Total sentences	7191

The synthetic documents obtained via RNN Generator have a constant length of 300 words. Since the quality of the generated text is much lower than that of the lyrics text, a lot of generated documents could not contain valuable sentences to be considered in the corpus. On the other hand, a single generated document could contain more than one valuable sentence. The above means that some groups of the sentences found come from the same document.

At this point, we have a corpus containing annotated sentences that will be use by the remaining stages, which are illustrated in Figure 2, and described in subsequent sections.

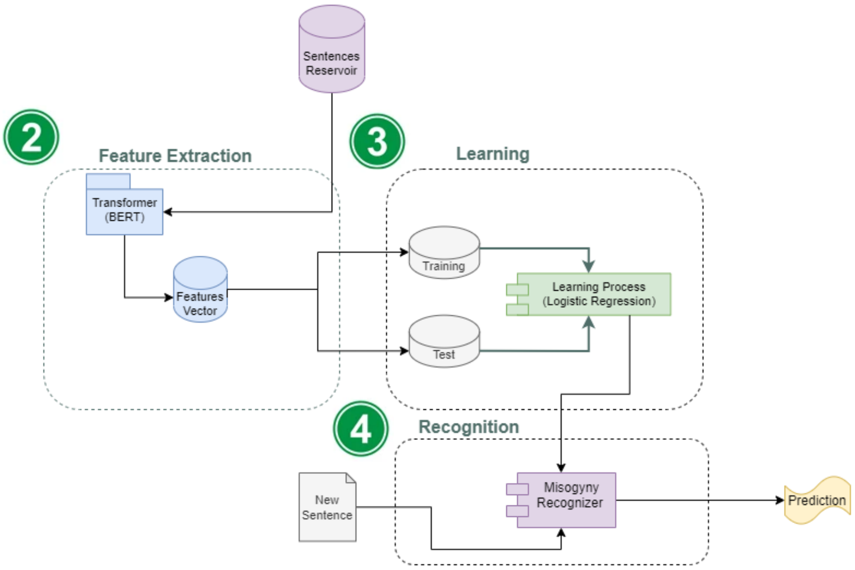

Figure 2. Overview of Feature Extraction, Learning, and Recognition stages. The first two make up the pipeline by the means of which a model of language is obtained on the basis of the inherent knowledge and experience conveyed in the corpus.

3.2. Feature Extraction

Having overcome the lack of data on Latin American Spanish, now, the main challenge is to encode such information into a model able to recognize that misogyny could be a subtle type of violence, and it can even hide behind seemingly flattering words, jokes, parodies, and other expressions.

We resort to recent approaches known as *transformers* [40] which incorporate the so-called *attention mechanisms* to identify these relations. In general, transformers provides thousands of pre-trained models to perform tasks on texts, such as classification, information extraction, question answering, summarization, translation, and text generation, in a lot of languages. They have been trained on large amounts of raw text in a self-supervised fashion (the objective is automatically computed from the inputs of the model). For practical purposes, we can build our models on top of already trained models, reducing the overall compute cost, and this process is usually known as *transfer learning*. We focus on a state-of-the art transformer known as BERT (Bidirectional Encoder Representations from Transformers) [41]. BERT is described as "bidirectional" because, unlike methods, such as Word2Vec, it can read a text or a sequence of words all at once, with no specific direction. Thanks to its bidirectionality, this model can understand the meaning of each word based on context both to the right and to the left of the word.

In general, any model based on a *transformer* architecture involves high computational resources to process and store a huge amount of training data and parameters. This complexity is latent to the resulting pre-trained language models that keep getting larger and heavier to new problems. Under this drawback, we decide to focus on distillation technique [42,43] that allows us to compress a large model, called *the teacher*, into a smaller model, called *the student*. Specifically, we use a "distilled" version of BERT known as DistilBERT, reported in Reference [44] as a suitable approach comparable to the performance of state-of-the-art transformers. The process of transfer learning via BERT allows us to extract the features of the sentences gathered in previous stage and denoted in what follows as $\mathbb{D}$; such a process involves the following steps:

1. Obtaining of an instance of a BERT model pre-trained on a large unlabeled dataset in Spanish.
2. Fine-tuning [45], where the model is initialized with the pre-trained parameters and all of them are fine-tuned using a labeled data regarding sentences previously categorized as misogynistic and non-misogynistic.

So far, $\mathbb{D}$ is a set of sentences that require be prepared in the form that BERT expects. For this aim, we use a tokenizer provided by BERT which we have called *BertTokenizer*, obtaining a set $\mathbb{T}$ containing encoded sentences in the form of multidimensional arrays. For each $\vec{t}$ in $\mathbb{T}$, we pad it with zeros until its length is equal to the longest array in $\mathbb{T}$. As mentioned, one of the main characteristics of BERT is the transformer structure, where its encoder pay attention to the sentence as a whole and not dividing it into tokens. We need to tell BERT which part of the whole tokenized and padded array $\vec{t} \in \mathbb{T}$ contains useful information. So, for each $\vec{t}$, we create a binary vector $\vec{m}$ indicating those positions in which there is a tokenized or padded value (encoded as 1 and 0, respectively). At this point, we have a set of binary vectors known as *attention mask* and denoted as $\mathbb{M}$. Finally, both $\mathbb{T}$ and $\mathbb{M}$ are propagated through BERT model (fine tuning). As a result, we obtain a high-dimensional vector (embedding) for each sentence in $\mathbb{D}$ representing its feature vector; the set of feature vector is denoted as $\mathbb{F}$. The above process is summarized in Algorithm 1 and illustrated in Figure 3.

Algorithm 1: Feature Extraction

```
Data:
$\mathbb{D}$: Sentences reservoir,
Result: Set of feature vector from $\mathbb{D}$
/* Creating instances of Bert Model and Bert Tokenizer */
$BertModel \leftarrow$ get_BertModel();
$BertTokenizer \leftarrow$ get_BertTokenizer();
/* Tokenizing sentences */
$\mathbb{T} \leftarrow BertTokenizer$.encode($\mathbb{D}$);
/* Padding Sentences and Attention Mask */
$maxLen \leftarrow$ max_length($\mathbb{T}$);
$\mathbb{M} \leftarrow \emptyset$;
foreach $t \in \mathbb{T}$ do
    $t \leftarrow$ padding($t, maxLen$);
    $m \leftarrow$ get_mask($t$);
    $\mathbb{M} \leftarrow \mathbb{M} \cup \{m\}$;
end
$\mathbb{F} \leftarrow BertModel$.transform($\mathbb{T}$,$\mathbb{M}$);
return $\mathbb{F}$
```

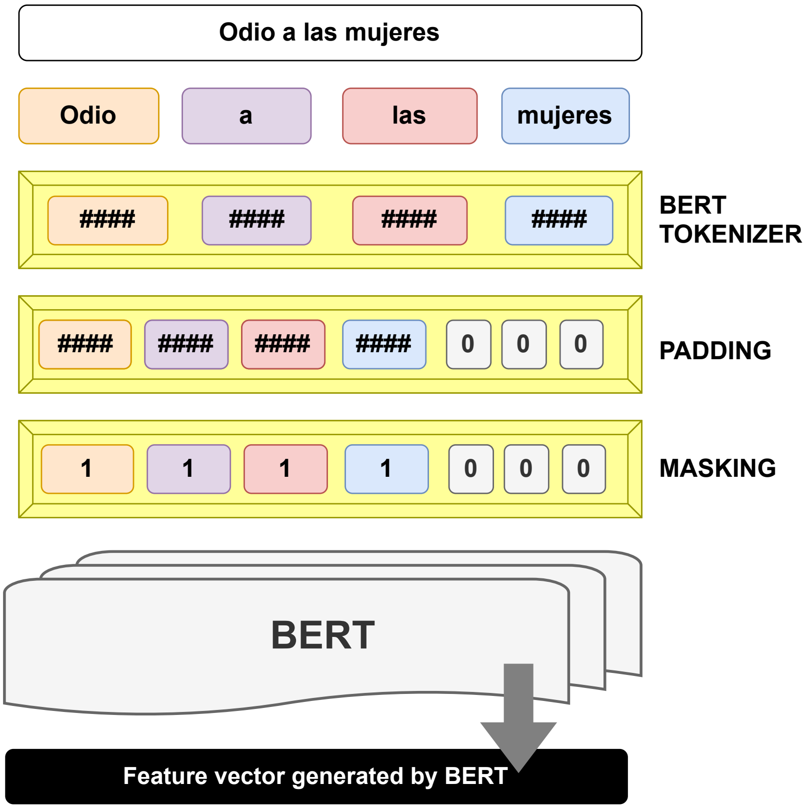

Figure 3. Illustration of feature extraction via BERT for the sentence in Spanish "Odio a las mujeres" (*"I hate women"*).

3.3. Learning Process

Having determined the semantic features of whole sentences in $\mathbb{D}$, we aim to find a model that is able to recognize the presence or absence of misogynistic patterns in the encoded sentences. At this point, we resort to Logistic Regression because it is fast, easily understandable, and appropriate for a dichotomous dependent variable as it is our case. So, from the encoded sentences $\mathbb{F}$ and their corresponding labels $\mathbb{L}$, we now define the so-called training and test datasets denoted as $\mathbb{X}_{train}$, $\mathbb{T}_{train}$ and $\mathbb{X}_{test}$, $\mathbb{Y}_{test}$. We execute an exhaustive search over an specified set of hyper-parameter values $\mathbb{P}$ for obtaining an appropriate model. This search involved cross-validation with $k = 10$ folds on the train set in order to obtain the best parameter values that attain the most reliable model. The set $\mathbb{P}$ included 27 tuples of the form $[optimizer, penalty, C]$ corresponding to optimization algorithm (lbfgs, sag, saga), norm used for penalization (l_1,l_2), and the inverse of regularization strength (1, 5, and 10 in our case), respectively. The above is illustrated in Algorithm 2, when the optimal $p^* \in \mathbb{P}$ is found, we create an instance of Logistic Regression using p^* and fit it to $\mathbb{X}_{train}$. Since this instance has a high degree of certainty of getting the best prediction score on $\mathbb{X}_{test}$, we consider it the best model for our purposes. The score was defined in terms of accuracy, precision, recall, and F1-score; the results regarding these metrics are discussed more fully in Section 4. Finally, the best hyper-parameter values obtained after executing an exhaustive search using cross-validation are: $C = 10$, lbfgs as optimizer and l_2 regularization.

Algorithm 2: Learning Process

```
Data:
$\mathbb{F}$: Encoded Sentences
$\mathbb{L}$: Labels of Sentences
$\mathbb{P}$: Set of tuples of hyper-parameters for estimator (Logistic Regression)
Result: Misogyny Recognizer Model
/* Defining train and test datasets                                    */
$\mathbb{X}_{train}, \mathbb{Y}_{train}, \mathbb{X}_{test}, \mathbb{Y}_{test} \leftarrow \textit{split\_data}(\mathbb{F}, \mathbb{L})$;
/* Applying a stratified sampling with $k$-folds on training data      */
$i \leftarrow 0, k \leftarrow 10$;
$\mathbb{X}_{folds}, \mathbb{Y}_{folds} \leftarrow \textit{get\_folds}(\mathbb{X}_{train}, \mathbb{Y}_{train}, k)$;
$\mathbb{S} \leftarrow \emptyset$;
$score \leftarrow 0$;
foreach $p \in \mathbb{P}$ do
    while $i < k$ do
        /* Creating an instance of the model                           */
        $model \leftarrow LogisticRegression.\mathbf{getInstance}(p)$;
        /* Validation set                                              */
        $\mathbb{X}_{val}, \mathbb{Y}_{val} \leftarrow \mathbb{X}_{folds}[i], \mathbb{Y}_{folds}[i]$;
        /* Training set                                                */
        $\mathbb{X}_{train}, \mathbb{Y}_{train} \leftarrow (\mathbb{X}_{folds} - \mathbb{X}_{folds}[i]), (\mathbb{Y}_{folds} - \mathbb{Y}_{folds}[i])$;
        /* Training Model                                              */
        $model.\mathbf{fit}(\mathbb{X}_{train}, \mathbb{Y}_{train})$;
        /* Validation Model                                            */
        $\mathbb{Y}_{predicted} \leftarrow model.\mathbf{predict}(\mathbb{X}_{val})$;
        $score \leftarrow score + \mathbf{get\_score}(\mathbb{Y}_{val}, \mathbb{Y}_{predicted})$;
        $i \leftarrow i + 1$;
    end
    $avg\_score \leftarrow score/k$;
    $\mathbb{S} \leftarrow \mathbb{S} \cup \{avg\_score\}$;
end
$p^* \leftarrow \mathbf{get\_best\_params}(\mathbb{S}, \mathbb{P})$;
$BestModel \leftarrow LogisticRegression.\mathbf{getInstance}(p^*)$;
/* Training Model                                                      */
$BestModel.\mathbf{fit}(\mathbb{X}_{train}, \mathbb{Y}_{train})$;
/* Testing Model                                                       */
$\mathbb{Y}_{predicted} \leftarrow BestModel.\mathbf{predict}(\mathbb{X}_{test})$;
$score \leftarrow \mathbf{get\_score}(\mathbb{Y}_{test}, \mathbb{Y}_{predicted})$;
return $BestModel$
```

3.4. Recognition Process

Having obtained a reliable model, now, we aim to estimate the degree of misogyny in a sentence s that the model has not seen before. In order to make s amenable to the model, we again use the previous instances of BERT model and BERT tokenizer. At this point, we have an encoded sentence in the form of a high dimensional vector denoted as f which is given as input to the model in order to obtain the prediction probabilities associated with the presence or absence of misogynistic elements in s. The above process is summarized in Algorithm 3.

```
Algorithm 3: Misogyny Recognition
  Data:
  $s$: Sentence to be analyzed,
  Result: Prediction vector
  /* Loading previous models                                  */
  $BertModel \leftarrow \textbf{get\_BertModel}()$;
  $BertTokenizer \leftarrow \textbf{get\_BertTokenizer()}$;
  $MisogynyRecognizer \leftarrow BestModel$;
  /* Tokenizing sentence to be analyzed                       */
  $t \leftarrow BertTokenizer.\textbf{encode}(s)$;
  /* Padding Sentences and Attention Mask                     */
  $maxLen \leftarrow BertModel.\textbf{get\_max\_length}()$;
  $t \leftarrow \textbf{padding}(t, maxLen)$;
  $m \leftarrow \textbf{get\_mask}(t)$;
  $f \leftarrow BertModel.\textbf{transform}(t, m)$;
  $p \leftarrow MisogynyRecognizer.\textbf{predict}(f)$;
  return $p$
```

4. Experiments and Results

The experiment consisted of assessing the performance of our approach from two aspects: (1) the learning ability of the model in terms of the well-known evaluation metrics (accuracy, precision, recall, and F1-score) to predict the misogyny degree on a test dataset consisting of sentences belonging to the corpus gathered via our proposal, and (2) the recognition ability to determine the presence or absence of misogynistic patterns in a real world dataset that includes sentences which the model has not seen during the learning process.

4.1. Learning Ability Assessment

As pointed out in Algorithm 2, the learning process is performed on the sets $\mathbb{F}$ and $\mathbb{L}$ corresponding to the set of encoded sentences in the form of high dimensional vector and their corresponding labels, respectively. From these datasets, we determine training and test datasets. On the training dataset, we performed a sampling strategy known as k-folds (with $k = 10$) in order to find the most reliable model that minimizes the overfitting. At this point, we assessed the yield of the resulting model based on the test dataset. This assessment is defined in terms of metrics, such as *accuracy*, *precision*, *recall*, and *F1-score*. Since a single iteration of the above process does not guarantee the overall performance of the model, we repeated it 100 times by changing random seed values, obtaining a statistical approximation to the real performance values, as it is illustrated in Figure 4.

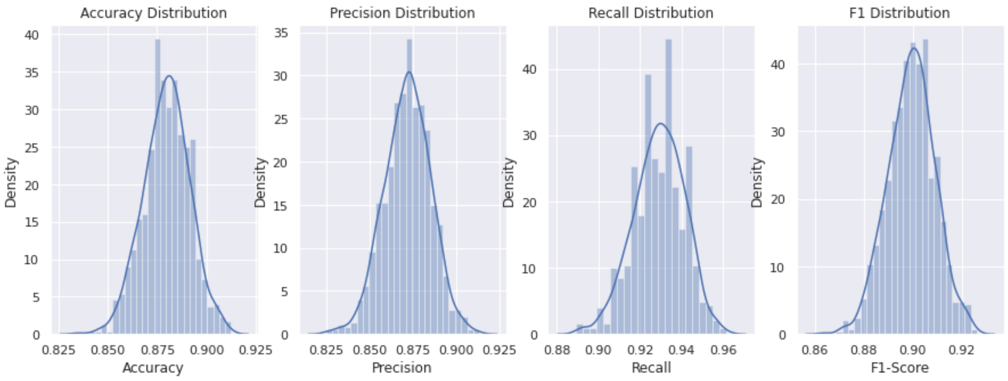

Figure 4. Approximation to the density function of the values corresponding to accuracy, precision, recall, and F1-Score exhibited by the model during the assessment of learning ability.

We analyzed the variability in the performance of our method using quartile summary statistics illustrated in Table 2. From the Interquartile Range (IQR), we can see that there are not large differences in the experiments for a particular metric, and Confidence Interval (CI) allows us to quantify how we expect the average performance to be. In general, the performance of our model was mostly kept inside acceptable range in all metrics.

Table 2. Quartile analysis for the assessment of *learning ability* of the model.

Statistic	Accuracy	Precision	Recall	F1-Score
Q_1	0.87	0.86	0.92	0.89
Q_2	0.88	0.87	0.93	0.89
Q_3	0.88	0.88	0.93	0.90
IQR	0.01	0.01	0.01	0.01
CI	[0.84, 0.91]	[0.83, 0.90]	[0.89, 0.96]	[0.87, 0.92]

4.2. Assessment for Real World Data

Finally, we conducted a set of experiments on a real-world dataset reported in Reference [13]. This work shows the results of different working teams grouped by task. We focused on the work regarding *Task 5* regarding the analysis of hate speech against women and immigrants on a corpus of labeled tweets in Spanish and English. The labels were encoded as three binary values described as follows:

- **HS**—a binary value indicating the presence or absence of hate speech against one of the given target people (women or immigrants).
- **TR**—a binary value indicating if the target of the hate speech is a generic group of people (0) or a specific individual (1).
- **AG**—a binary value indicating if the hate speech present in the tweet is aggressive (1) or not (0).

For our purposes, we filtered the data to retain only those tweets in Spanish and labeled with HS = 1. Then, we selected manually those tweets containing expressions focused on women (most labeled with HS = 1,TR = 1). As result a set of instances containing misogynistic tweets were obtained. On the other hand, instances containing non-misogynistic tweets were obtained simply by filtering tweets tagged with HS = 0. In total, a set of 1774 tweets (with a comparable number of classes) was obtained. On this data, we executed 100 times our model attempting to recognize the presence or absence of misogynistic manifestations. We resort again to the metrics of accuracy, precision, recall, and F1-score to quantify the recognition ability. In Figure 5, the approximation to the density function for the executed experiments is shown.

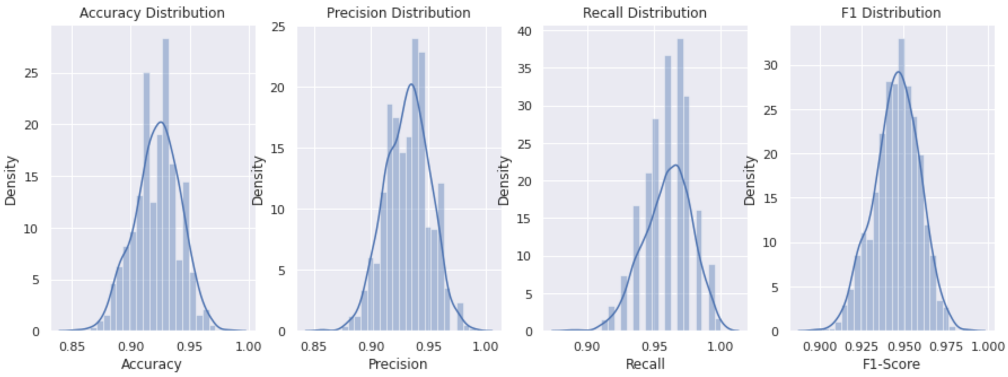

Figure 5. Approximation to the density function of the values corresponding to accuracy, precision, recall, and F1-Score exhibited by the model during the assessment of *recognition ability*.

As in the above assessment, the model exhibited a successful performance. The quartile analysis in Table 3 shows a close variability to that of learning assessment. It means that the recognition ability of our method (on unknown data) is statistically comparable to its ability exhibited during the learning process. To formalize this assumption, we finally conducted a hypothesis test that allowed us to show that there is not a significant difference between what we have called learning ability and recognition ability (p-value > 0.05).

Table 3. Quartile analysis for the assessment for the proposed Latin American Spanish misogyny *recognition* model. The metrics reported are Accuracy, Precision, Recall, and F1-Score.

Statistic	Accuracy	Precision	Recall	F1-Score
Q_1	0.90	0.91	0.93	0.93
Q_2	0.92	0.93	0.94	0.94
Q_3	0.93	0.94	0.95	0.95
IQR	0.02	0.02	0.01	0.01
CI	[0.86, 0.98]	[0.87, 0.98]	[0.90, 0.98]	[0.90, 0.98]

5. Discussion

We have been able to generate a Language Model based on previous knowledge and using extracted data from a pipeline that we designed with the purpose of covering up as much sources as we could where we could find a misogynistic attitude. We must emphasize certain points we think should be considered on this work and in the reported results. First, setting boundaries was very important from the beginning because we are aware of the challenges that NLP tasks have, as well as the subjectivity in opinions while working with topics, such as misogyny. So, we kept following this path knowing we could not get rid of any kind of subjectivity, domain, social and geographic context, polarization, or controversy. On the other hand, we selected Spanish for the lack of particular works on misogyny detection in this language and the importance of pointing out this behavior through social media and its social impact in the Latin America context. On the other hand, we also need to clarify that, even though it is pretty clear we were the ones who selected the data directly, tagging each sentence as misogyny or not misogyny, the process of retrieving has its own subjectivity since what we got is what the internet is indexing as misogyny. So, putting this in mind, we can ensure that the more data we add to the current dataset, the bigger the context will be around this particular type of hate speech. It is also true that our model is capturing exactly what it learned. It is identifying what we taught it was misogyny, considering the full sentence. We know this by comparing the probability that our model generates for each tag in certain sentences. In Table 4, we can see the difference between passive misogyny sentences that do not have any negative word that can be considered as aggressive or insulting, active misogyny that actually uses an insulting word, and a non-misogynistic sentence. The results the model gave us in these sentences show us how it finds more probability of a sentence being misogyny where it is talking directly to women or about women, as it should, but it still does find differences between nice words inside a sentence that itself is misogynistic and one that could be aggressive but that does not specify a gender. We must say the results can improve at using our pipeline for gathering data more than one time, collecting more data from the web that can be used as misogyny sentences, but we are also concerned that our model still lacks in the negative class since we have only trained with proverbs and some other sentences we obtained from the documents we collected. However, at this point, we know the path we should keep working on in order to make our model more robust.

Table 4. Examples of instances used in the Misogyny Recognizer. We can note the difference between passive misogyny sentences, active misogyny that actually use an insulting word and a non-misogynistic sentence.

	Original Tag	Not Misogyny	Misogyny
Y tu eres mi esposa ve a la cocina zorra	1	0.0886	0.9113
Solo sirves para lavar los trastes	0	0.8533	0.1466
Me encanta mi novia porque hace las tareas domésticas	1	0.3099	0.6900
Ella también tiene que respetarse, se visten así y luego se quejan cuando les pasa algo	1	0.0167	0.9832

6. Conclusions

Throughout this article, we have discussed the relevance of having mechanisms for detecting hate speech online, particularly for detecting misogyny. Thanks to the recent scientific advances, it will be possible, in the short term, to use automatic detection to build indicators of violence against women. We have also discussed the challenges of bringing this type of technology to under-resourced languages through Feature Extraction methods, and we have taken Latin American Spanish as a case study.

As part of the technical aspects presented, we consider it most important to highlight the following. We have proposed a pipeline to collect, filter, tag, and generate documents' features for training a recognition model, stating a path where we can be sure we would get a fair quantity of data to use, and it promises to be helpful in future research. The misogynistic corpus we generated is a labeled resource for future investigations, and we plan to expand it for future work. This resource is also available at http://shorturl.at/lptzT (accessed on 4 November 2021).

Our model accurately discerns misogynistic comments based on the data we collected, and, for this, it seems to identify contextual cues of a sentence to generate the probabilities. The model reacts to subtleties in the language. The trained model can be tested at http://contralamisoginia.org/ (accessed on 4 November 2021). We are sure that the continuation of this project and other similar projects will transcend the creation of awareness around misogyny in Latin America.

Author Contributions: Conceptualization, E.A.-B., A.M.-V.; Methodology, E.A.-B., A.M.-V., Y.M.-P.; Software, E.A.-B. and Y.M.-P.; Validation, E.A.-B., Y.M.-P.; Investigation, I.L.-A., E.A.-B., A.M.-V., Y.M.-P.; Resources, I.L.-A., E.A.-B. and O.S.S.; Data curation, E.A.-B. and Y.M.-P.; Writing—original draft preparation, E.A.-B., A.M.-V., Y.M.-P.; Writing—review and editing, I.L.-A., E.A.-B., A.M.-V., Y.M.-P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: http://shorturl.at/lptzT (accessed on 4 November 2021).

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

References

1. WHO. Violence against women: A global health problem of epidemic proportions. In *WHO News Release*; WHO: Geneva, Switzerland, 2013.
2. WHO. *Global and Regional Estimates of Violence against Women: Prevalence and Health Effects of Intimate Partner Violence and Non-Partner Sexual Violence*; World Health Organization: Geneva, Switzerland, 2013.
3. CSW. Report of the Secretary-General of the Commission on the Status of Women, United Nations, Sixty-Fifth Session. 2021. Available online: https://undocs.org/E/CN.6/2021/3 (accessed on 5 August 2021).
4. Galtung, J. Cultural violence. *J. Peace Res.* **1990**, *27*, 291–305. [CrossRef]
5. Foucault, M. *The Order of Discourse (L'ordre du Discours)*; Galimart: Paris, France, 1971. (In French)
6. Hewitt, S.; Tiropanis, T.; Bokhove, C. The problem of identifying misogynist language on Twitter (and other online social spaces). In Proceedings of the 8th ACM Conference on Web Science, Hannover, Germany, 22–25 May 2016; pp. 333–335.
7. Hardaker, C.; McGlashan, M. "Real men don't hate women": Twitter rape threats and group identity. *J. Pragmat.* **2016**, *91*, 80–93. [CrossRef]
8. Waseem, Z.; Hovy, D. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In Proceedings of the NAACL Student Research Workshop, San Diego, CA, USA, 12–17 June 2016; pp. 88–93.
9. Davidson, T.; Warmsley, D.; Macy, M.; Weber, I. Automated hate speech detection and the problem of offensive language. In Proceedings of the International AAAI Conference on Web and Social Media, Montréal, QC, Canada, 15–18 May 2017; Volume 11.
10. Yao, M.; Chelmis, C.; Zois, D.S. Cyberbullying ends here: Towards robust detection of cyberbullying in social media. In Proceedings of the World Wide Web Conference, San Francisco, CA, USA, 13–17 May 2019; pp. 3427–3433.
11. Ridenhour, M.; Bagavathi, A.; Raisi, E.; Krishnan, S. Detecting Online Hate Speech: Approaches Using Weak Supervision and Network Embedding Models. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*; Springer: Berlin, Germany, 2020; pp. 202–212.
12. Lynn, T.; Endo, P.T.; Rosati, P.; Silva, I.; Santos, G.L.; Ging, D. A comparison of machine learning approaches for detecting misogynistic speech in urban dictionary. In Proceedings of the 2019 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (Cyber SA), Oxford, UK, 3–4 June 2019; pp. 1–8.
13. Basile, V.; Bosco, C.; Fersini, E.; Nozza, D.; Patti, V.; Rangel Pardo, 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, Minneapolis, MN, USA, 6–7 June 2019; pp. 54–63. [CrossRef]
14. Kumar, R.; Ojha, A.K.; Lahiri, B.; Zampieri, M.; Malmasi, S.; Murdock, V.; Kadar, D. Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying. In Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, Marseille, France, 11–16 May 2020.
15. Mandl, T.; Modha, S.; Majumder, P.; Patel, D.; Dave, M.; Mandlia, C.; Patel, A. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In Proceedings of the 11th Forum for Information Retrieval Evaluation, Kolkata, India, 12–15 December 2019; pp. 14–17.
16. Zampieri, M.; Malmasi, S.; Nakov, P.; Rosenthal, S.; Farra, N.; Kumar, R. Predicting the type and target of offensive posts in social media. *arXiv* **2019**, arXiv:1902.09666.
17. Aragon, M.; Carmona, M.A.; Montes, M.; Escalante, H.J.; Villaseñor-Pineda, L.; Moctezuma, D. Overview of MEX-A3T at IberLEF 2019: Authorship and aggressiveness analysis in Mexican Spanish tweets. In Proceedings of the 1st SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Bilbao, Spain, 24 September 2019.
18. Fersini, E.; Rosso, P.; Anzovino, M. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. *IberEval@ SEPLN* **2018**, *2150*, 214–228.
19. Bretschneider, U.; Peters, R. Detecting offensive statements towards foreigners in social media. In Proceedings of the 50th Hawaii International Conference on System Sciences, Hilton Waikoloa Village, HI, USA, 4–7 January 2017.
20. Kovács, G.; Alonso, P.; Saini, R. Challenges of Hate Speech Detection in Social Media. *SN Comput. Sci.* **2021**, *2*, *9*. [CrossRef]
21. Ousidhoum, N.; Lin, Z.; Zhang, H.; Song, Y.; Yeung, D.Y. Multilingual and multi-aspect hate speech analysis. *arXiv* **2019**, arXiv:1908.11049.
22. Sigurbergsson, G.I.; Derczynski, L. Offensive language and hate speech detection for Danish. *arXiv* **2019**, arXiv:1908.04531.
23. Pitenis, Z.; Zampieri, M.; Ranasinghe, T. Offensive language identification in Greek. *arXiv* **2020**, arXiv:2003.07459.
24. Bosco, C.; Felice, D.; Poletto, F.; Sanguinetti, M.; Maurizio, T. Overview of the evalita 2018 hate speech detection task. In Proceedings of the EVALITA 2018 Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Turin, Italy, 12–13 December 2018; Volume 2263, pp. 1–9.
25. Albadi, N.; Kurdi, M.; Mishra, S. Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. In Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Barcelona, Spain, 28–31 August 2018; pp. 69–76.
26. Ibrohim, M.O.; Budi, I. Multi-label hate speech and abusive language detection in Indonesian twitter. In Proceedings of the Third Workshop on Abusive Language Online, Florence, Italy, 1–2 August 2019; pp. 46–57.

27. Ptaszynski, M.; Pieciukiewicz, A.; Dybała, P. Results of the Poleval 2019 Shared Task 6: First Dataset and Open Shared Task for Automatic Cyberbullying Detection in Polish Twitter. 2019. Available online: https://ruj.uj.edu.pl/xmlui/bitstream/handle/item/152265/ptaszynski_pieciukiewicz_dybala_results_of_the_poleval_2019.pdf?sequence=1&isAllowed=y (accessed on 4 November 2021).
28. Hussein, O.; Sfar, H.; Mitrović, J.; Granitzer, M. NLP_Passau at SemEval-2020 Task 12: Multilingual Neural Network for Offensive Language Detection in English, Danish and Turkish. In Proceedings of the Fourteenth Workshop on Semantic Evaluation, Barcelona, Spain, 13–14 September 2020; pp. 2090–2097.
29. Pereira-Kohatsu, J.C.; Quijano-Sánchez, L.; Liberatore, F.; Camacho-Collados, M. Detecting and monitoring hate speech in Twitter. *Sensors* **2019**, *19*, 4654. [CrossRef] [PubMed]
30. Corazza, M.; Menini, S.; Cabrio, E.; Tonelli, S.; Villata, S. A multilingual evaluation for online hate speech detection. *ACM Trans. Internet Technol. TOIT* **2020**, *20*, 1–22. [CrossRef]
31. Ranasinghe, T.; Zampieri, M. Multilingual offensive language identification with cross-lingual embeddings. *arXiv* **2020**, arXiv:2010.05324.
32. Pamungkas, E.W.; Basile, V.; Patti, V. Misogyny detection in twitter: a multilingual and cross-domain study. *Inf. Process. Manag.* **2020**, *57*, 102360. [CrossRef]
33. Anzovino, M.; Fersini, E.; Rosso, P. Automatic Identification and Classification of Misogynistic Language on Twitter. In *Natural Language Processing and Information Systems*; Silberztein, M., Atigui, F., Kornyshova, E., Métais, E., Meziane, F., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 57–64.
34. Bashar, M.A.; Nayak, R.; Suzor, N. Regularising LSTM classifier by transfer learning for detecting misogynistic tweets with small training set. *Knowl. Inf. Syst.* **2020**, *62*, 4029–4054. [CrossRef]
35. Frenda, S.; Bilal, G. Exploration of Misogyny in Spanish and English tweets. In Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), Sevilla, Spain, 18 September 2018; Volume 2150, pp. 260–267.
36. García-Díaz, J.A.; Cánovas-García, M.; Colomo-Palacios, R.; Valencia-García, R. Detecting misogyny in Spanish tweets. An approach based on linguistics features and word embeddings. *Future Gener. Comput. Syst.* **2021**, *114*, 506–518. [CrossRef]
37. Fulper, R.; Ciampaglia, G.L.; Ferrara, E.; Ahn, Y.; Flammini, A.; Menczer, F.; Lewis, B.; Rowe, K. Misogynistic language on Twitter and sexual violence. In Proceedings of the ACM Web Science Workshop on Computational Approaches to Social Modeling (ChASM), Bloomington, IN, USA, 23–26 June 2014; pp. 57–64.
38. Molina-Villegas, A. La incidencia de las voces misóginas sobre el espacio digital en México. In *Jóvenes, Plataformas Digitales y Lenguajes: Diversidad Lingüística, Discursos e Identidades*; Pérez-Barajas, A.E., Arellano-Ceballos, A.C., Eds.; Elementum: Pachuca, Mexico, 2021; in press.
39. Cundiff, G. The influence of rap and hip-hop music: An analysis on audience perceptions of misogynistic lyrics. *Elon J. Undergrad. Res. Commun.* **2013**, *4*.
40. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 5998–6008.
41. McCann, B.; Bradbury, J.; Xiong, C.; Socher, R. Learned in Translation: Contextualized Word Vectors. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017.
42. Buciluǎ, C.; Caruana, R.; Niculescu-Mizil, A. Model compression. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, 20–23 August 2006; pp. 535–541.
43. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
44. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv* **2019**, arXiv:1910.01108.
45. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, 2–7 June 2019; Volume 1 (Long and Short Papers); Burstein, J., Doran, C., Solorio, T., Eds.; Association for Computational Linguistics: Cambridge, MA, USA, 2019; pp. 4171–4186. [CrossRef]