1. Introduction

The growing access to combined telecommunication along with the increase of electronic social network adoption has granted users a convenient method of sharing posts and comments on the Internet. However, even if this is an improvement in human communication, this environment also has provided proper conditions resulting in serious negative consequences to the welfare of society, due to a type of user who posts offensive comments and does not care about the psychological impact of his/her words, harming other users feelings. This phenomenon is called cyber-aggression [1]. Cyber-aggression is a frequently used keyword in the literature to describe a wide range of offensive behaviors other than cyber-bullying [2–5].

Unfortunately, this problem has spread into a wide variety of mass media; Bauman [6] says that some of the most-used digital media for cyber-aggression are: social networks (e.g., Facebook and Twitter), short message services, forums, trash-polling sites, blogs, video sharing websites, and chat rooms, among others. Their accessibility and fast adoption are a double-edged sword because it is impossible to have moderators to keep an eye on every post made and filter its content. Therefore, cyber-aggression [7] has become a threat to society’s welfare, generating electronic violence.

When cyber-aggression is constant, then it becomes cyber-bullying, mainly characterized by the invasion of privacy, harassment, and use of obscene language against one user [8], in most of the cases...
against a minor [9]. Unlike bullying, cyber-bullying can happen 24/7, and the consequences for the victim can be more dramatic, because this not only creates insecurity, trust issues, and depression, but can also create suicidal thoughts [10] with fatal consequences. Both cyber-bullying and cyber-aggression can harm a user by mocking and ridiculing them an indefinite number of times [11,12]; therefore it is crucial to support the increase of detection of these social problems. According to Ditch The Label and their Cyber-Bullying Survey [13] some of the social networks with the highest number of users that have reported cases of cyber-aggression are Facebook, Twitter, Instagram, YouTube, and Ask.fm.

Currently, there is a wide variety of research aiming to mitigate cyber-bullying. Those from psychology or social sciences aim to detect criminal behavior and provide prevention strategies [14], and those from the field of computer science seek to develop effective techniques or tools to support the prevention of cyber-aggression and cyber-bullying [15]. Nevertheless, it is still complicated to identify a unique behavior pattern of aggression. It is also necessary to highlight that most of the research is carried out by English-speaking countries, so it is difficult to obtain free resources on the Internet for the development of tools that allow the analysis of comments written in Spanish. The Spanish language has an extensive vocabulary with multiple expressions, words that vary in different Spanish-speaking countries, synonyms and colloquialisms that are different in different countries and regions. For this reason, we consider it prudent to focus on the Spanish language of Mexico. However, we do not rule out the possibility of extending this research in other Spanish-speaking countries.

There are different types of cyber-aggression. Bauman [6] identifies and describes some forms of cyber-bullying that are used to attack victims by digital means, such as flaming, harassment, denigration, masquerading, outing, trickery, social exclusion, and cyber-stalking. Peter [16] highlights death threats, homophobia, sexual acts, the threat of physical violence, and damage to existing relationships, among others. Ringrose and Walker [17,18] indicate that women and children are the most vulnerable groups in case of cyber-aggression.

In Mexico, there has been a wave of hate crimes; in 2014, the World Health Organization reported that Mexico occupies second place in the world of hate crimes for factors such as “fear”. In Mexico, the case of violence based on sexual orientation is one of the most vulnerable groups, according to the National Survey on Discrimination [19]. As found in an investigation by a group of civil society organizations [20], in Mexico and Latin America, 84% of lesbian, gay, bisexual, and transgender (LGBT) population have been verbally harassed. On the other hand, violence against women is an exponential problem in Mexico has affected young women between 18 and 30 [21]. Mexico has a public organization known as Instituto Nacional de Estadística y Geografía (INEGI) [22,23], responsible for regulating and coordinating the National System of Statistical and Geographic Information, as well as conducting a national census. INEGI [22] has reported that nine million Mexicans have suffered at least one incident of digital violence in one of its different forms. In 2016, INEGI reported that of the approximately 61.5 million women in Mexico, at least 63% aged 15 and older had experienced acts of violence. Racism is another problem in Mexico that has been growing through offensive messages related to “the wall” with the United States (US), the mobilization of immigrant caravans and discrimination by skin color. According to an INEGI study, skin color can affect a person’s job growth possibilities, as well as their socioeconomic status [23].

The American Psychological Association [24] says that different combinations of contextual factors such as gender, race, social class, and other sources of identity can result in different coping styles, and urges psychologists to learn guidelines for psychological practice with people with different problems (e.g., lesbian, gay, and bisexual patients).

For many years, problems from various research areas have been addressed through Artificial Intelligence (AI) techniques. AI is a discipline that emerged in the 1950s and is basically defined as the construction of automated models that can solve real problems by emulating human intelligence. Some recent applications include the following: in [25], two chemical biodegradability prediction models are evaluated against another commonly used biodegradability model; Li et al. [26] applied Multiscale Sample Entropy (MSE), Multiscale Permutation Entropy (MPE), and Multiscale Fuzzy
Entropy (MFE) feature extraction methods along with Support Vector Machines (SVMs) classifier to analyze Motor Imagery EEG (MI-EEG) data; Li and coworkers [27] proposed the new Temperature Sensor Clustering Method for thermal error modeling of machine tools, then the weight coefficient in the distance matrix and the number of the clusters (groups) were optimized by a genetic algorithm (GA); in [28], fuzzy theory and a genetic algorithm are combined to design a Motor Diagnosis System for rotor failures; the aim in [29] is to design a new method to predict click-through rate in Internet advertising based on a Deep Neural Network; Ocaña and coworkers [30] proposed the evolutionary algorithm TS-MBFOA (Two-Swim Modified Bacterial Foraging Optimization Algorithm) and proved a real problem that seeks to optimize the synthesis of a four-bar mechanism in a mechatronic systems.

In this research, we used IA algorithms to classify comments in three cases of cyber-aggression: racism, violence based on sexual orientation, and violence against women. The accurate identification of cybernetic aggression cases is the first step of a process to reduce the incidence of this phenomenon. We applied IA techniques, specifically Random Forest, Variable Importance Measures (VIMs), and OneR. The comments used to create the data set were collected from Facebook, considering specific news related to the cyber-aggression cases included in this study.

In line with the aims of this study, we propose to answer the following research questions:

1. Is it possible to create a model of automatic detection of cyber-aggression cases with high precision? To answer this question, we will experiment with two classifiers of different approaches and compare their performance.

2. What are the terms that allow detection of cyber-aggression cases included in this work effectively? We will seek the answer to this question using methods to identify the relevant features for each cyber-aggression case.

The present work is organized as follows: Section 2 presents a review of related research works, Section 3 describes the materials and methods used in this research, Section 4 shows the architecture of the proposed computational model, Section 5 describes the experiments and results obtained, and the last section concludes the article.

2. Related Research

Due to the numerous victims of cyber-bullying and the fatal consequences it causes, today there is a need to study this phenomenon in terms of its detection, prevention, and mitigation. The consequences of cyber-bullying are worrisome when victims cannot cope with the emotional stress of abusive, threatening, humiliating, and aggressive messages. Presently there are several types of research to avoid or reduce online violence; even so, it is necessary to develop more precise techniques or online tools to support the victims.

Raisi [31] proposes a model to detect offensive comments on social networks, in order to intervene by filtering or advising those involved. To train this model, they used comments with offensive words from Twitter and Ask.fm. Other authors [8,32] have developed conversation systems, based on intelligent agents that allow supportive emotional feedback to victims who suffer from cyber-bullying. Reynolds [33] proposed a system to detect cyber-bullying in the Formspring social network, based on the recognition of violence patterns in the user posts, through the analysis of offensive words; in addition, it uses a ranking level of the detected threat. Likewise, it obtained an accuracy of 81.7% with J48 decision trees.

Ptaszynski [34] describes in his research the development of an online application in Japan, for school staff and parents with the duty of detecting inappropriate content on unofficial secondary websites. The goal is to report cases of cyber-bullying to federal authorities; in this work they used SVMs and got 79.9% accuracy. Rybnicek [12] proposes an application for Facebook to protect minor users from cyber-bullying and sex-teasing. This application aims to analyze the content of images and videos, as well as the activity of the user to record changes in behavior. In another study, a collection
Different studies fight against cyber-bullying by supporting the classification of situations, topics or types of it. For example, at the Massachusetts Institute of Technology, [35] developed a system to detect cyber-bullying in YouTube video comments. The system can identify the topic of the message, such as sexuality, race, and intelligence. The overall success of this experiment was 66.7% accuracy using SVMs. Similarly, the study carried out by Nandhini [36] proposes a system to detect cyber-bullying activities and classify them as flaming, harassment, racism, and terrorism. The author uses a fuzzy classification rule; however, the accuracy of the results are very low (around 40%), but he increased the efficiency of classifier up to 90% using a series of rules. In the same way, Chen [37] proposes an architecture to detect offensive content and identify potential offensive users in social media. The system achieves an accuracy of 98.24% in sentence offensive detection and an accuracy of 77.9% in user offensiveness detection. Similar work to this is presented by Sood [38], in which comments were tagged from a news site using Amazon’s Mechanical Turk to create a profanity-labeled data set. In this study, they use SVMs with a profanity list-based and Levenshtein edit distance tool, getting an accuracy of 90%.

In addition to studies that aim to detect messages with offensive content or classify messages into types of cyber-bullying, other studies are trying to prove that a system dedicated to taking care of user behavior can reduce situations of cyber-bullying. Bosse [8] performed an experiment consisting of a normative multi-agent game with children 6 to 12 years old. In the experiment, the author highlights a particular case: a girl who, regardless of the agent’s warnings, continued to violate rules and was removed from the game. However, she changed her attitude and began to follow the rules of the game. Through this research, the author shows that in the long term, the system manages to reduce the number of rule violations. Therefore it is possible to affirm that research with technological proposals using sentiment analysis, text mining, multi-agent, or other AI techniques can support the reduction of violence online.

3. Materials, Methods and Metrics

3.1. Materials

Study data. Since we did not find a Spanish data set for the study, we had to collect data through the Facebook API from relevant news in Latin America related to three cases of cyber-aggression: racism, violence based on sexual orientation, and violence against women. We collected 5000 comments; however, we used only 2000, those free of spam (spam characterizes in this study as texts with rare characters, images of expression or humorous thought such as memes, empty spaces, or comments unrelated to the problem). We then grouped the comments (instances) as follows: 700 comments about violence based on sexual orientation, 700 comments about violence against women and 600 racist comments.

Labeling process. In the literature, we found that some researchers [33,38–40] used the web service of Amazon’s Mechanical Turk and paid for anonymous online workers to manually label comments, reviews, words, images, among others. However, in [41], tagging by workers from Amazon’s Mechanical Turk showed that at least 2.5% of reviews, classified as “no cyber-bullying” should be tagged as “cyber-bullying”, for this reason, they sought support from graduate students and some undergraduates of psychology.

Due to the above, we decided to use a group of three teachers with experience in machine-learning algorithms supported by psychologists with experience in evaluation and intervention in cases of bullying in high schools to manually tag the comments. The psychologists explained to professors when a comment is considered offensive in the three cases of cyber-aggression and how to label the comment according to the predefined range of values. So, the purpose of the labeling process was to add an offensive value to each comment considering the case of cyber-aggression and a predefined
numerical scale. In the cases of comments with violence against women we used a scale from zero to two, with the lowest value of those comments the least offensive. For comments about violence based on sexual orientation, a scale of values of four to six was used, in the same way with the lowest value for the least offensive comments. Finally, a scale of eight to ten was used for racist comments, with the lowest value for the least offensive comments and the highest value for the most. As a result, we obtained the data set of offensive comments and then used them in the feature-selection procedure and training process. This data set consisted of two columns; the first column contained the instance or comment and second the offensive value according to each comment. We describe, in the following section, the algorithms and methods used in this research for the feature-selection procedure and training process.

3.2. Methods

Random Forest. The Random Forest classification algorithm, developed by Breiman [42], is a set of decision trees that outputs a predictive value and is robust against over-fitting. This algorithm has two parameters, mtry and ntree, which may vary to improve its performance. The first represents the number of input variables chosen at random in each division and the second the number of trees.

Random Forest obtains a class vote of each tree and proceeds to classify according to the vote of the majority. We show the functioning of Random Forest in Algorithm 1.

Algorithm 1: Random Forest for classification

1. Let \( N \) be the number of training cases, and \( M \) the number of variables.
2. Select the value of \( m \), in order to determine the decision in a given node, which must be less than the value of \( M \).
3. Vary and choose \( N \) for the tree (completing all training cases) and use the rest of the test cases to estimate the error.
4. For each node of the tree, randomly choose \( m \). Obtain the vote of each tree and classify according to the vote of the majority.

In classification processes, the default value for \( m \) is calculated by \( \sqrt{p} \) or \( \log_2 p \), where \( p \) is the number of features. Random Forest provides the VIMs method, through which it is possible to rank the importance of variables in regression or classification problems.

Variable Importance Measures (VIMs). VIMs, based on CART classification trees [43], allows calculation of the weight to identify key attributes or features. In VIMs there are two embedded methods for measuring variable importance proposed by Breiman: Mean Decrease Impurity importance (MDI) and Mean Decrease Accuracy (MDA). MDI ranks each feature importance as the sum over the number of splits that include the feature, proportionally to the number of samples it splits; this method is used in classification cases and applied the Gini index to measure the node impurity [44,45]. Random Forest also uses the Gini index for determining a final class in each tree. MDA is also called the permutation importance and is based on out-of-bag (OOB) [46] samples to measure accuracy; this method is suitable for regression problems [43]. In this work we used MDI with Gini index to classify comments in cases of cyber-aggression, which is defined by:

\[
G(t) = 1 - \sum_{k=1}^{Q} p^2(k|t)
\]

where \( Q \) is the number of classes, \( p(k|t) \) is the estimated class probability for feature \( t \) or node \( t \) in a decision tree and \( k \) is an output class.

At the end of the process, VIMs calculate a score for each variable, normalized by the standard deviation. The higher the score, the more important the feature. In this work, we used R [47] and applied the randomForest [48] library to develop the classification model.
OneR stands for One Rule. OneR is a rule-based classification algorithm. A classification rule has the form: if attribute1 <relational operator> value1 <logical operator> attribute2 <relational operator> value2 <...> then decision-value. OneR generates one rule for each predictor in the data set, and among them all, it selects the rule with the smallest total error as the only one rule [49]. One advantage of rules is that they are simple for humans to interpret.

3.3. Evaluation Metrics

Confusion Matrix. It shows the performance of the prediction model, comparing the results of the predictive model against real values.

Accuracy (ACC). This is the evaluation of the predictive model performance. In binary classification, the accuracy is calculated by dividing the number of correctly identified cases among the total cases. It is computed by:

\[
ACC = \frac{TP + TN}{T + N}
\]  

(2)

Sensitivity (TRP). This metric measures the proportion of true positives which were correctly classified and is calculated by:

\[
TRP = \frac{TP}{TP + FN}
\]  

(3)

Specificity (SPC). This metric measures the proportion of true negatives, which were correctly classified. It is given by:

\[
SPC = \frac{TN}{FP + TN}
\]  

(4)

where \(TP = \) True Positive, \(TN = \) True Negative, \(FP = \) False Positive and \(FN = \) False Negative, in Equations (2)-(4).

Kappa statistic. This method introduced by Cohen [50] and Ben-David [51] is used to measure the accuracy of machine-learning algorithms. Kappa measures the agreement between the classifier itself and the ground truth corrected by the effect of the agreement between them by chance. This method is defined by:

\[
K = \frac{P_o + P_c}{1 - P_c}
\]  

(5)

where \(P_o\) is the proportion of agreement between the classifier and the ground truth, \(P_c\) is the proportion of agreement expected between the classifier and the ground truth by chance.

4. Experimental Procedure

4.1. Random Forest

This section describes the steps followed to carry out the experimental procedure with Random Forest (see Figure 1).

Phase 1. Comment Extraction. We detail this step in the Materials section (study data).

Phase 2. Feature Selection. In this phase, we applied the VIMs method to identify the key attributes for each cyber-aggression case [43]. The feature-selection process is shown in Figure 2. We applied this process for each cyber-aggression case. The first step, after we cleaned the comments, was to separate them according to a numerical range, as follows: 0–2 for comments related to violence based on sexual orientation, 4–6 for those about violence against women, and 8–10 for those related to racism. Then, we identified the frequent terms to eliminate those repeated and not important. Later, we partitioned the set of comments (corpus) using 2/3 for training and 1/3 for testing. At the same time, we applied the VIMs method along with the Gini index to obtain the terms with the higher
weight to create the feature corpus. After this process, we configured a vector of 30 seeds, and then we applied the `random_forest.importance` instruction. We executed this process 30 times, considering a different seed to train with different partitions of the training corpus, in order to evaluate the results of Random Forest. We considered high-performance values greater than 70%, and then it was possible to use the corpus of features in the parameter optimization process, which we applied to find the most appropriate value of `mtry`. In this process, we used the `Fselector` library in the R system for statistical computing [47].

Phase 3. Classification. This step is the training process considering the feature corpus, identified by VIMs (applied in phase 2 of the Model). We used 10-fold cross-validation and Random Forest algorithm to adjust the `mtry` and `ntree` parameters. After the training process, we carried out the classification process using Random Forest, which we executed another 30 times, implementing in
each execution a different seed, generated by the Mersenne-Twister method [52]. To test the result of the classification in different executions with different sections of training-sets, a variety of seeds is important. Figure 3 represents the steps followed in this phase.

**Figure 3. Classification process.**

4.2. OneR

This section describes the experimental procedure for OneR classifier.

For each aggression case, we first calculated the frequency of appearance of each term and the average frequency of appearance of all the terms. Then, we selected the terms above this average to use the most frequent terms as predictors. Afterwards, we made a subsetting of the original data set containing only the selected terms and created a new data set. From this new data set, we created three data sets to perform the classification experiments. In the first data set, we kept the LGBT class and marked the other two as ALL; in the second we kept the Machismo class and marked the other two as ALL; and also with the third. Finally, we performed 30 independent runs, where we calculated the average values of the metrics used across all the study.

5. Results

5.1. Feature Selection Using VIMs

We used the VIMs method to identify the key features of each cyber-aggression case, as shown in Figure 2 and described in Phase 2 (feature selection). The result of this process (see Table 1) was made in two executions—initial and final. In the initial execution, we showed the results without a cleaning process; in the final execution, the results are shown considering a cleaning process and the application of VIMs method. Column 1, “Case”, of Table 1, represents the cyber-aggression case, where VS is violence based on sexual orientation, VM is violence against women and R is racism. The second column, “execution”, indicates the initial and final executions from each cyber-aggression case. The third column, “Potential features”, presents the arithmetic mean of potential features in each cyber-aggression case. Column 4 shows the arithmetic mean of weights obtained from all terms, where we can see that values from the final execution increase in comparison with the result from the initial execution. In column five, “Maximum weight” represents the maximum weights obtained in each cyber-aggression case and each type of execution; this term appears in Table 2, in the column of Importance from each type of cyber-aggression. The final column, “Minimum weight”, is the minimum weights obtained in each case of cyber-aggression and each case of execution.
Table 1. Results of the feature-selection procedure.

| Case | Execution | Potential Features | Average Weights | Maximum Weight | Minimum Weight |
|------|-----------|--------------------|-----------------|----------------|---------------|
| VS   | Initial   | 33.33%             | 5.43            | 29.44          | 1.00          |
| VS   | Final     | 89.33%             | 6.44            | 33.45          | 0.13          |
| VM   | Initial   | 29.16%             | 4.33            | 21.09          | 0.02          |
| VM   | Final     | 88.49%             | 5.24            | 27.04          | 0.075         |
| R    | Initial   | 48.06%             | 4.65            | 23.66          | 1.00          |
| R    | Final     | 89.10%             | 6.87            | 28.04          | 0.20          |

The processing and cleaning of comments influenced the results of the final execution, because in the first execution some irrelevant words such as “creo” (“I think” in English) obtained a high value of 11.70, “comentarios” (“comments” in English) obtained a value of 5.36 and “cualquiera” (“anyone” in English) obtained a value of 2.34.

We measured the importance of the terms with VIMs using the Gini index. Table 2 shows the most important terms that were identified by this method, which are those with the highest value. In the case of violence based on sexual orientation, the most important term was “asco” (“disgust”) with a value of 33.4539, in the case of violence against women, the most important term was “estúpida” (“stupid”) with a value of 27.0473, and in racist comments, the most important term was “basura” (“garbage”) with a value of 28.0465. We show in Table 2 an extract of important term results (feature selection) for each cyber-aggression case after the cleaning of comments; some terms include an “*”, which is not part of the original term, but was used to censor the offense.

Table 2. Results of the importance of the terms in the feature-selection procedure.

| VS Term | Importance | VM Term | Importance | R Term | Importance |
|---------|------------|---------|------------|--------|------------|
| asc*    | 33.4539    | est*pida| 27.0473    | b*sura | 28.0465    |
| circo   | 28.0842    | c*chetadas| 19.5111    | escl*vos| 26.3688    |
| d*generados| 24.1965  | c*ger   | 17.7479    | color   | 22.8754    |
| ab*rración| 23.9967  | c*lera  | 16.3509    | ch*nos  | 20.9468    |
| an*rmal | 18.0855    | estorban| 13.0006    | r*za    | 18.9392    |
| asqu*roso| 17.8648   | p*rras  | 12.7073    | n*gro   | 12.4821    |
| e*n*remedado| 17.6001 | cocina | 12.6314    | en*no   | 12.2413    |
| dep*rvados| 14.4772   | drama   | 12.3626    | am*al   | 10.6022    |
| an*rmales| 13.5536   | br*ta   | 11.6915    | b*gotonas| 10.9998    |
| mar*cas | 12.6927    | c*lebra  | 11.0332    | cr*stianos| 9.2143    |
| est*pidez| 11.0387   | p*ta    | 10.3944    | pr*etos | 7.3209     |
| biblia  | 10.9030    | ab*sivas| 10.3768    | t*roristas| 6.6306    |
| cabr*nes| 10.8184    | b*rra   | 9.5017     | c*zar   | 6.2852     |
| enf*mos | 10.6553    | débil    | 8.9978     | m*tar   | 6.0199     |
| cerrada | 10.2412    | descuidada| 8.7899    | enemigos| 5.5840     |
| an*males| 9.5388     | arg*endo| 8.4999     | ch*cha  | 4.6080     |
| arrepientanse| 9.1431 | asqu*rosas| 8.1761    | ind*o   | 4.4520     |
| d*sviados| 8.9325     | c*gan   | 7.8184     | d*lincuente| 4.3617    |
| deberian| 8.4167     | m*tenidas| 7.6039    | gr*ngo  | 4.3958     |

Once we identified the features for each problem, we carried out the training process (phase 3), where we evaluated the results using the metrics described in Section 3.3 (evaluation metrics). Finally, we applied the classifier model, if the accuracy exceeded 70%.

5.2. Training Process and Parameter Tuning

To allow the algorithm to be prepared to perform a better classification, we carried out a training process with Random Forest. We selected the 50 features with the highest weight of each problem; also, we performed 10-fold cross-validation and a variation of the value of the mtry parameter using
random numbers between zero and 50. We used these random numbers in the same way for each problem. We measured the training performance of the algorithm using the accuracy and Kappa metrics. In Table 3, we show the overall results obtained in this process, where R is racism, VS is violence based on sexual orientation and VM is violence against women. In column two (Mtry), we show the optimal values of the mtry parameter obtained for each case of cyber-aggression, column three presents the accuracy obtained with respect of the value of the mtry parameter, and column four shows the value of Kappa with respect of the value of the mtry parameter. The values used in the mtry parameter were in a range of 23 to 44. Figure 4 shows the complete results for the adjustment of the mtry parameter.

Table 3. Results of the adjustment of the mtry parameter.

| Case | Mtry | Accuracy | Kappa |
|------|------|----------|-------|
| VS   | 50   | 0.8211   | 0.5789|
| VM   | 23   | 0.8327   | 0.6217|
| R    | 44   | 0.9169   | 0.7737|

Figure 4. Graph of adjustment of the results of the mtry parameter.

In the case of the ntree parameter, we considered two values: 100 and 200, based on the investigations reported in [53,54]. Oshiro [53] concludes that the variation of the mtry with values that indicate a higher number of trees only increases the computational cost and does not have a significant improvement ratio.

Once we obtained the value of mtry for each case of cyber-aggression, as well as the value of ntree, we carried out the classification process with Random Forest. It is essential to highlight that we made this process with 30 executions varying the training and test set, in order to test the performance of Random Forest with different sections of the data set (explained in phase three of the architecture of the model).

Based on the results obtained with accuracy and confusion matrices, in the 30 executions, metrics such as balanced accuracy, sensitivity, and specificity were obtained. These results are presented in Table 4, as well as the standard deviation (sd) of each metric depending on cyber-aggression case.
As can be seen, the model obtains a better performance in the presented metrics, even though in the three problems it highlights the sensitivity with low values and the specificity with high values. The performance of the model indicates that negative comments were correctly classified; this proves that the results were very close to the expected performance because 95% of the comments in this data set were offensive. Since this work focuses on detecting potential features to classify comments on three problems of cyber-aggression (racism, sexual violence, and violence against women), it was justified in implementing a data set composed primarily of negative comments.

Table 4. Results of comment classification. Standard deviation (sd) is also shown.

| Case | Accuracy | Balanced Accuracy | Sensitivity | Specificity |
|------|----------|-------------------|-------------|-------------|
| R    | 0.9030   | 0.9237            | 0.6783      | 0.9883      |
| sd   | 0.0129   | 0.0129            | 0.0466      | 0.0060      |
| VS   | 0.8036   | 0.8313            | 0.5320      | 0.9591      |
| sd   | 0.0162   | 0.0203            | 0.0355      | 0.0115      |
| VM   | 0.8204   | 0.8233            | 0.6325      | 0.9261      |
| sd   | 0.0148   | 0.0187            | 0.0311      | 0.0165      |

5.3. Results with OneR

As Table 5 shows, we found an average frequency of terms for VS aggression of 13.41, 21.61 for VM, and 18.60 for R. This means that for example in VS, terms appear 13.41 times as an average, where the most frequent term is “mar*c*nes”, with 29 appearances. In VM, terms appear 21.61 times as an average, the most frequent term being “vieja” with 76 appearances. Finally, in R, terms appear 18.60 times as an average, with “negros” as the most frequent with 63 appearances. We show the selected terms in the same table, i.e., those with a frequency of appearance greater than the average frequency of appearance of all the terms. Figure 5, known as word clouds, show the importance of the terms for each type of aggression. The bigger the term in the figure, the more relevant it is. These pictures match the terms shown in Table 5. As expected, the most common offensive terms in Spanish are present. The most common offensive words are “maric*nes”, “asco", “put*s", “gays", “gay", “mi*rda" for VS; “vieja", “pinch*", “mujeres", “put*", “viejas", “pinch*s" for VM, and “negros", “mi*rda", “indio", “negro", “judíos", “musulmanes" for R. We used the terms from Table 5 as input to OneR classifier to build the classification models for each cyber-aggression case.

Table 5. Selected terms by their average frequency for each type of aggression.

| Case | Average Frequency of Terms | Selected Terms (Quantity) |
|------|-----------------------------|---------------------------|
| VS   | 13.41                       | maric*nes, asco, put*s, gays, gay, mi*rda, hombres, pinch*s, jot*s, maric*s, dios, mundo, gente, homosexuales, ser, respeto, put* (17) |
| VM   | 21.61                       | vieja, pinch*, mujeres, put*, viejas, pinch*s, bien, mujer, p*nd*ja, put*s, solo, asi (12), negros, mi*rda, indio, negro, judíos, musulmanes, |
| R    | 18.6                        | malditos, p*nd*ejo, mexicanos, pinch* (10) |
Figure 5. Word clouds for each cyber-aggression case. (a) Violence based on sexual orientation (VS). (b) Violence against women (VM). (c) Racism (R).

Table 6 shows the classification results with OneR using only the most relevant terms found, as described in the Experimental Procedure section. We show the results averaged over 30 independent runs. We also show the standard deviation (sd). We found an accuracy above 0.90 for all types of aggression. Moreover, all metrics reached 0.90 in most cases. Results prove the efficiency of the methodology applied.

Table 6. Classification results using only the most relevant terms for each type of aggression. The results are shown averaged over 30 runs. Standard deviation (sd) is also shown.

| Case | Accuracy | Balanced Accuracy | Sensitivity | Specificity | Kappa |
|------|----------|-------------------|-------------|-------------|-------|
| R    | 0.9623   | 0.9569            | 0.9712      | 0.9426      | 0.9119|
| sd   | 0.0080   | 0.0103            | 0.0071      | 0.0184      | 0.0188|
| VS   | 0.9525   | 0.9192            | 0.8383      | 1.0000      | 0.8795|
| sd   | 0.0102   | 0.0174            | 0.0348      | 0.0000      | 0.0274|
| VM   | 0.9545   | 0.9615            | 0.9274      | 0.9957      | 0.9066|
| sd   | 0.0099   | 0.0112            | 0.0138      | 0.0237      | 0.0206|
In Table 7, we show the best classification rule generated by OneR for each type of aggression. Also, the number of correctly classified instances and the balanced accuracy in the test set represents the effectiveness of each rule. We can see that OneR selected the same most relevant terms for each type of aggression. Some irrelevant terms in Spanish were included, such as “así”, “bien”, “solo”, among others. They represent adverbs and prepositions instead of nouns or adjectives.

| Classification | VS | VM | R |
|----------------|----|----|----|
| Rule asco -> lgbt | asco -> ALL | asco -> ALL |
| Dios -> lgbt | dios -> ALL | dios -> ALL |
| G*y -> lgbt | g*y -> ALL | g*y -> ALL |
| G*y's -> lgbt | g*y's -> ALL | g*y's -> ALL |
| Gente -> lgbt | gente -> ALL | gente -> ALL |
| Hombres -> lgbt | hombres -> ALL | hombres -> ALL |
| Homosexuales -> lgbt | homosexuales -> ALL | homosexuales -> ALL |
| J*tos -> lgbt | j*tos -> ALL | j*tos -> ALL |
| M*ric*s -> lgbt | m*ric*s -> ALL | m*ric*s -> ALL |
| M*ric*nes -> lgbt | m*ric*nes -> ALL | m*ric*nes -> ALL |
| Mi*rda -> ALL | mi*rda -> ALL | mi*rda -> ALL |
| Mundo -> lgbt | mundo -> ALL | mundo -> ALL |
| Pinch*s -> ALL | pinch*s -> ALL | pinch*s -> ALL |
| P*ta -> ALL | p*ta -> ALL | p*ta -> ALL |
| P*tos -> lgbt | p*tos -> ALL | p*tos -> ALL |
| Respeto -> lgbt | respeto -> ALL | respeto -> ALL |
| Ser -> lgbt | ser -> ALL | ser -> ALL |
| Así -> ALL | así -> ALL | así -> ALL |
| Bien -> ALL | bien -> ALL | bien -> ALL |
| Mujer -> ALL | mujer -> ALL | mujer -> ALL |
| Mujeres -> ALL | mujeres -> ALL | mujeres -> ALL |
| P*nd*ja -> ALL | p*nd*ja -> ALL | p*nd*ja -> ALL |
| P*nche -> ALL | p*nche -> ALL | p*nche -> ALL |
| P*tas -> ALL | p*tas -> ALL | p*tas -> ALL |
| Solo -> ALL | solo -> ALL | solo -> ALL |
| Vieja -> ALL | vieja -> ALL | vieja -> ALL |
| Viejas -> ALL | viejas -> ALL | viejas -> ALL |
| Indio -> ALL | indio -> ALL | indio -> ALL |
| Judios -> ALL | judios -> ALL | judios -> ALL |
| Malditos -> ALL | malditos -> ALL | malditos -> ALL |
| Mexicanos -> ALL | mexicanos -> ALL | mexicanos -> ALL |
| Musulmanes -> ALL | musulmanes -> ALL | musulmanes -> ALL |
| Negro -> ALL | negro -> ALL | negro -> ALL |
| Negros -> ALL | negros -> ALL | negros -> ALL |
| P*nd*jo -> ALL | p*nd*jo -> ALL | p*nd*jo -> ALL |

Correct instances:
- VS: 772/818
- VM: 775/818
- R: 782/818

Balanced Accuracy:
- VS: 0.9458
- VM: 0.9776
- R: 0.9774

6. Discussion and Conclusions

Cyber-aggression has increased negatively as the use of social networks increases, which is why in this work we have sought to develop computational tools analyzing offensive comments and classified into three categories.

There are already a variety of software tools or applications that operate under AI techniques to detect offensive comments, filter them or send messages of support to the victim. However, it is still necessary to improve the performance of these tools to get more effective predictions. The development of tools that work in the Spanish language is also required, since most of the research targets English-speaking countries, which is why it is difficult to obtain resources for algorithm training, such as data sets, lexicons, corpora, among others. Moreover, it is crucial to consider certain idioms or
colloquialisms of the region where the model applies, so the translation of the available resources of the Web is not always convenient. Therefore it is necessary to create resources in the Spanish language. According to the need for a data set of offensive comments in Spanish, and considering the example of other related research, the authors have created their own data sets of comments using social networks such as Twitter [55], Formspring.me [35], Facebook [56], Ask.fm [31] and others where cyber-bullying has been increasing [57].

We decided to create a data set of offensive comments using Facebook. At first, Twitter was used to extract comments on these three cases of cyber-aggression, but most of the comments were irrelevant. The Twitter API allows download of comments using a hashtag, but there are few users who make offensive comments and use a hashtag to identify the comment. For this reason, we decided to use news on Facebook about marriage between people of the same sex, publications about women’s triumphs, or reports of physical abuse and news about Donald Trump’s wall. Besides, Facebook is the most-used social network in Mexico [58]. We believe that a gathering of more relevant news comments such as abortion, adoption by same-sex couples, feminicides and other related news, not only from Mexico but also from Latin America, as well as the inclusion of experts who study the Spanish language and colloquialisms, can improve the performance of the classifier.

This paper describes the development of a model to classify cyber-aggression cases applying Random Forest and OneR. We seek to initially impact Mexico, where there has been a wave of hate crimes. Our contribution is as follows: (1) We created a data set with cyber-aggression cases from social networks. (2) We focused on cyber-aggression cases in our native language, i.e., Spanish. (3) Specifically, we were interested in the most representative types of cyber-aggression in our country of origin, Mexico. (4) We identified the most relevant terms for the detection of cyber-aggression cases included in the study. (5) We created an automatic detection model of cyber-aggression cases with high precision and interpretable by the human being (rule-based).

The results obtained in this work with Random Forest support the identification of relevant features to classify offensive comments into three cases: racism, violence based on sexual orientation, and violence against women. Nevertheless, OneR outperformed Random Forest in identifying types of cyber-aggression, in addition to providing a simple classification rule with the most relevant terms for each type of aggression. Even when we obtained high-performance classification models with this particular data, it is essential to highlight that the classifiers used in this study have a better performance classifying offensive comments against the LGBT population and racists comments. Therefore the exploration of other machine-learning techniques and the continuous update of the offensive data set may not be ruled out in future work, thus allowing the analysis of another kind of cyber-aggression case, e.g., those suffered by children. Also, it is important to continue improving the feature-selection process. On the other hand, building an automatic labeling system for offensive comments made by social networks users, and thus minimizing human error, will be of great help. Finally, we will seek to identify the victims of cyber-aggression to provide them with psychological attention according to the case of harassment that they suffer.

Author Contributions: Conceptualization—G.O.G.-E.; Methodology—G.O.G.-E., J.H.-T.; Software—G.O.G.-E.; Supervision—M.V.-A.; Validation—G.O.G.-E.; formal analysis—J.H.-T.; writing—original draft preparation—G.O.G.-E., M.V.-A.; writing—review and editing—G.O.G.-E., M.V.-A.

Funding: This research received no external funding.

Acknowledgments: We want to extend our appreciation to Consejo Nacional de Ciencia y Tecnología (CONACYT) (National Council for Science and Technology) under the ‘Catedras CONACYT’ programme, No. 1591.

Conflicts of Interest: The authors declare no conflicts of interest.
References

1. Kowalski, R.M.; Limber, S.P.; Limber, S.; Agatston, P.W. Cyberbullying: Bullying in the Digital Age; John Wiley & Sons: Hoboken, NJ, USA, 2012.

2. Grigg, D.W. Cyber-aggression: Definition and concept of cyberbullying. *J. Psychol. Couns. Sch.* 2010, 20, 143–156. [CrossRef]

3. Kopecky, K.; Szotkowski, R. Cyberbullying, cyber aggression and their impact on the victim—The teacher. *Telemat. Inform.* 2017, 34, 506–517. [CrossRef]

4. Corcoran, L.; Fredstrom, B.K.; Gilman, R. Electronic and school-based victimization: Unique contexts for adjustment difficulties during adolescence. *J. Youth Adolesc.* 2011, 40, 405–415. [CrossRef]

5. Bosse, T.; Stam, S. A normative agent system to prevent cyberbullying. In Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Volume 02. IEEE Computer Society, Lyon, France, 22–27 August 2011; pp. 425–430.

6. Smith, P.K. Cyberbullying and cyber aggression. In *Handbook of School Violence and School Safety*; Routledge: London, UK, 2012; pp. 21–26.

7. Ringrose, J.; Gill, R.; Livingstone, S.; Harvey, L. A Qualitative Study of Children, Young People and ’Sexting’: A Report Prepared for the NSPCC; National Society for the Prevention of Cruelty to Children: London, UK, 2012.

8. Walker, S.; Sanci, L.; Temple-Smith, M. Sexting: Young women’s and men’s views on its nature and origins. *J. Adolesc. Health* 2013, 52, 697–701. [CrossRef]
22. INEGI. Microdatos del Módulo sobre Ciberacoso (MOCIBA). 2015. http://www.beta.inegi.org.mx/contenidos/proyectos/investigacion/ciberacoso/2015/doc/mociba2015_principales_resultados.pdf (accessed on 16 November 2018).

23. INEGI. Módulo de Movilidad Social Intergeneracional (MMSI). 2016. https://www.inegi.org.mx/programas/mmsi/2016/ (accessed on 11 February 2019).

24. American Psychological Association. Guidelines for psychological practice with lesbian, gay, and bisexual clients. Am. Psychol. 2012, 67, 10. [CrossRef]

25. Baker, J.R.; Gamberger, D.; Mihelcic, J.R.; Sabljic, A. Evaluation of Artificial Intelligence Based Models for Chemical Biodegradability Prediction. Molecules 2004, 9, 989–1003. [CrossRef] [PubMed]

26. Li, M.a.; Liu, H.n.; Zhu, W.; Yang, J.F. Applying Improved Multiscale Fuzzy Entropy for Feature Extraction of MI-EEG. Appl. Sci. 2017, 7, 92. [CrossRef]

27. Li, F.; Li, T.; Wang, H.; Jiang, Y. A Temperature Sensor Clustering Method for Thermal Error Modeling of Heavy Milling Machine Tools. Appl. Sci. 2017, 7, 82. [CrossRef]

28. Kuo, C.C.; Liu, C.H.; Chang, H.C.; Lin, K.J. Implementation of a Motor Diagnosis System for Rotor Failure Using Genetic Algorithm and Fuzzy Classification. Appl. Sci. 2017, 7, 31. [CrossRef]

29. Wang, Q.; Liu, F.; Xing, S.; Zhao, X. A New Approach for Advertising CTR Prediction Based on Deep Neural Network via Attention Mechanism. Comput. Math. Methods Med. 2018, 2018, 1–11. [CrossRef]

30. Hernández-Ocaña, B.; Pozos-Parra, M.D.P.; Mezura-Montes, E.; Portilla-Flores, E.A.; Vega-Alvarado, E.; Calva-Yáñez, M.B. Two-Swim Operators in the Modified Bacterial Foraging Algorithm for the Optimal Synthesis of Four-Bar Mechanisms. Comput. Intell. Neurosci. 2016, 2016, 1–18. [CrossRef]

31. Raisi, E.; Huang, B. Cyberbullying identification using participant-vocabulary consistency. arXiv 2016, arXiv:1606.08084.

32. Van der Zwaan, J.M.; Dignum, V.; Jonker, C.M. A conversation model enabling intelligent agents to give emotional support. In Modern Advances in Intelligent Systems and Tools; Springer: Berlin/Heidelberg, Germany, 2012; pp. 47–52.

33. Reynolds, K.; Kontostathis, A.; Edwards, L. Using machine learning to detect cyberbullying. In Proceedings of the 2011 10th International Conference on Machine learning and applications and workshops (ICMLA), Honolulu, HI, USA, 18–21 December 2011; Volume 2, pp. 241–244.

34. Ptaszyński, M.; Dybala, P.; Matsuba, T.; Masui, F.; Rzepka, R.; Araki, K. Machine learning and affect analysis against cyber-bullying. In Proceedings of the 36th AISB, Leicester, UK, 29 March–1 April 2010; pp. 7–16.

35. Dinakar, K.; Reichart, R.; Lieberman, H. Modeling the detection of Textual Cyberbullying. Soc. Mob. Web 2011, 11, 11–17.

36. Nandhini, B.S.; Sheeba, J. Online social network bullying detection using intelligence techniques. Procedia Comput. Sci. 2015, 45, 485–492. [CrossRef]

37. Chen, Y.; Zhou, Y.; Zhu, S.; Xu, H. Detecting offensive language in social media to protect adolescent online safety. In Proceedings of the 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing, Amsterdam, The Netherlands, 3–5 September 2012; pp. 71–80.

38. Sood, S.O.; Antin, J.; Churchill, E.F. Using Crowdsourcing to Improve Profanity Detection. In Proceedings of the AAAI Spring Symposium: Wisdom of the Crowd, Palo Alto, CA, USA, 26–28 March 2012; Volume 12, p. 6.

39. Rosa, H.; Pereira, N.; Ribeiro, R.; Ferreira, P.; Carvalho, J.; Oliveira, S.; Coheur, L.; Paulino, P.; Simão, A.V.; Trancoso, I. Automatic cyberbullying detection: A systematic review. Comput. Hum. Behav. 2019, 93, 333–345. [CrossRef]

40. Potha, N.; Maragoudakis, M. Cyberbullying detection using time series modeling. In Proceedings of the 2014 IEEE International Conference on Data Mining Workshop, Shenzhen, China, 14 December 2014; pp. 373–382.

41. Ptaszyński, M.; Leliwa, G.; Piech, M.; Smywiński-Pohl, A. Cyberbullying Detection–Technical Report 2/2018, Department of Computer Science AGH, University of Science and Technology. arXiv 2018, arXiv:1808.00926.

42. Breiman, L. Random forests. Mach. Learn. 2001, 45, 5–32. [CrossRef]

43. Hjelte, A. Computing Random Forests Variable Importance Measures (VIM) on Mixed Numerical and Categorical Data; KTH, School of Computer Science and Communication (CSC): Stockholm, Sweden, 2016.

44. Han, H.; Guo, X.; Yu, H. Variable selection using mean decrease accuracy and mean decrease gini based on random forest. In Proceedings of the 2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, 26–28 August 2016; pp. 219–224.
45. Breiman, L.; Cutler, A. Breiman and Cutler's Random Forests for Classification and Regression. 2018. Available online: https://cran.r-project.org/web/packages/randomForest/randomForest.pdf (accessed on 12 April 2019).
46. Breiman, L. Out-Of-Bag Estimation; University of California: Berkeley CA, USA, 1996. Available online: https://www.stat.berkeley.edu/~breiman/OOBestimation.pdf (accessed on 12 April 2019).
47. R Core Team. R: A Language and Environment for Statistical Computing; R Foundation for Statistical Computing: Vienna, Austria, 2013.
48. Liaw, A.; Wiener, M. Classification and regression by randomForest. R News 2002, 2, 18–22.
49. Witten, I.; Frank, E.; Hall, M. Practical Machine Learning Tools and Techniques; Morgan Kaufmann: Burlington, MA, USA, 2011.
50. Cohen, J. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 1960, 20, 37–46. [CrossRef]
51. Ben-David, A. About the relationship between ROC curves and Cohen’s kappa. Eng. Appl. Artif. Intell. 2008, 21, 874–882. [CrossRef]
52. Matsumoto, M.; Nishimura, T. Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. ACM Trans. Model. Comput. Simul. 1998, 8, 3–30. [CrossRef]
53. Oshiro, T.M.; Perez, P.S.; Baranauskas, J.A. How many trees in a random forest? In International Workshop on Machine Learning and Data Mining in Pattern Recognition; Springer: Berlin/Heidelberg, Germany, 2012; pp. 154–168.
54. Torruco, J.H. Descriptive and Predictive Models of Guillain-Barré Syndrome Based on Clinical Data Using Machine Learning Algorithms. Ph.D. Thesis, Universidad Juárez Autónoma de Tabasco, Tabasco, México, 2015.
55. Sanchez, H.; Kumar, S. Twitter bullying detection. Ser. NSDI 2011, 12, 15.
56. Hamouda, A.E.D.A.; El-taher, F. Sentiment analyzer for arabic comments system. Int. J. Adv. Comput. Sci. Appl. 2013, 4. [CrossRef]
57. Van Hee, C.; Jacobs, G.; Emmery, C.; Desmet, B.; Lefever, E.; Verhoeven, B.; De Pauw, G.; Daelemans, W.; Hoste, V. Automatic detection of cyberbullying in social media text. PLoS ONE 2018, 13, e0203794. [CrossRef] [PubMed]
58. INFOTEC. 13o. Estudio sobre los Hábitos de los Usuarios de Internet en México. 2018. Available online: https://www.infotec.mx/work/models/infotec/Resource/1012/6/images/Estudio_Habitos_Usuarios_2017.pdf (accessed on 11 February 2019).

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).