mathematics

MDPI

Article

Deep Neural Network for Gender-Based Violence Detection on Twitter Messages

Carlos M. Castorena [1,†,‡], Itzel M. Abundez [1,†], Roberto Alejo [1,*,†,‡], Everardo E. Granda-Gutiérrez [2,†], Eréndira Rendón [1,†] and Octavio Villegas [1,†]

1 Division of Postgraduate Studies and Research, National Technological of Mexico, Instituto Tecnológico de Toluca, Metepec 52149, Mexico; ccastorenal@toluca.tecnm.mx (C.M.C.); iabundezb@toluca.tecnm.mx (I.M.A.); erendonl@toluca.tecnm.mx (E.R.); ovillegasc@toluca.tecnm.mx (O.V.)
2 UAEM University Center at Atlacomulco, Autonomous University of the State of Mexico, Toluca 50450, Mexico; eegrandag@uaemex.mx
* Correspondence: ralejoe@toluca.tecnm.mx; Tel.: +52-722-2816463
† Current address: Av. Tecnológico s/n, Agrícola Bellavista, Metepec 52149, Mexico.
‡ These authors contributed equally to this work.

Abstract: The problem of gender-based violence in Mexico has been increased considerably. Many social associations and governmental institutions have addressed this problem in different ways. In the context of computer science, some effort has been developed to deal with this problem through the use of machine learning approaches to strengthen the strategic decision making. In this work, a deep learning neural network application to identify gender-based violence on Twitter messages is presented. A total of 1,857,450 messages (generated in Mexico) were downloaded from Twitter: 61,604 of them were manually tagged by human volunteers as negative, positive or neutral messages, to serve as training and test data sets. Results presented in this paper show the effectiveness of deep neural network (about 80% of the area under the receiver operating characteristic) in detection of gender violence on Twitter messages. The main contribution of this investigation is that the data set was minimally pre-processed (as a difference versus most state-of-the-art approaches). Thus, the original messages were converted into a numerical vector in accordance to the frequency of word's appearance and only adverbs, conjunctions and prepositions were deleted (which occur very frequently in text and we think that these words do not contribute to discriminatory messages on Twitter). Finally, this work contributes to dealing with gender violence in Mexico, which is an issue that needs to be faced immediately.

Keywords: gender-based violence in Mexico; twitter messages; deep neural networks; class imbalance

Citation: Castorena, C.M.; Abundez, I.M.; Alejo, R.; Granda-Gutiérrez, E.E.; Rendón, E.; Villegas, O. Deep Neural Network for Gender-Based Violence Detection on Twitter Messages. *Mathematics* **2021**, *9*, 807. https://doi.org/10.3390/math9080807

Academic Editors: Florin Leon, Mircea Hulea and Marius Gavrilescu

Received: 26 February 2021
Accepted: 6 April 2021
Published: 8 April 2021

1. Introduction

Gender-based violence (GBV) is a big concern around the globe [1]. The United Nations (UN) recognized GBV as a problem involving health and development [2]. A UN declaration about GBV, specifically the cause to women, describes it as all those acts of violence that results or potentially could lead into physical, psychological or sexual damage or suffering; it also includes the menacing of doing such acts, coercion to perform them and arbitrary deprivation of liberty, no matter if this is done in public or private circumstances [3].

Mexico has shown an escalation in the number of victims of GBV due to its social, economic and political context [4,5]. Moreover, crisis like the recent novel coronavirus disease (COVID-19) outbreak have exposed critical inequalities in the social and economic environments, as well as the health system, which have negatively contributed to the GBV problem [6].

Efforts of scholars and activists have increasingly turned society and government attention to this problem, warning about how certain conditions of power or privilege tend to reproduce broader relations of inequality, domination, exploitation, victimization and,

finally, loss of humanity [1]. In this respect, computer science researchers have developed algorithms and methodologies based on machine learning to address the GBV problem. For example, Ref. [7] presents a camouflaged electronic device to help potential victims of GBV; it allows to send a voice command and Global Positioning System (GPS) location via smartphone to a Control Center, which analyzes the message to properly assist the victim. A similar but more sophisticated work is presented in [8]; it uses two psychological sensors to identify GBV through a robust speaker identification system, based on the evaluation of speech stress conditions by using data augmentation techniques. Rodríguez-Rodríguez et al. [9] used historic open access data to model and forecast GBV through machine learning methods; their methodology produced successful results in three specific Spanish territories with different populations.

GBV has affected many women around the world in online social network environments [10] and several works have been developed to tackle this problem. In Ref. [11], a classification of cyber-bullying detection methods in online social networks was presented; it shows a survey of techniques to automatically identify cyber-bullying through the machine learning algorithms. Another interesting approach is MANDOLA [12]; it is a big-data processing system intended to evaluate the proliferation and effect of online hate-related speech, which is generally inspired by religion beliefs, ethnicity or gender. Gutiérrez-Esparza et al. [13] studied two machine learning algorithms, and the variable importance measures (VIMs) method, to select the best features from the data set, in order to classify situations of cyber-aggression on Facebook for Spanish-language users from Mexico. They collected 2000 Facebook comments, which were manually labeled as racism, violence based on sexual orientation and violence against women, by a group of three machine learning teachers which supported the psychologists who specialized in evaluation and intervention of bullying situation in high schools. Experimental results of these works showed a classification performance greater than 90% in accuracy.

Twitter has been a scenario where violence against women, indigenous, minorities and migrants, is frequent. Consequently, much work has been focused on this problem and the potential use of machine learning has been demonstrated as a methodology in Ref. [14]. In addition, data-mining [15] has been used to detect domestic violence. Other works have been performed for automatic detection of sexual violence [16], cyber-bullying [17], hate expressions [18], offensive or aggressiveness [13,19,20] on the twitter messages' content, in which the feature extraction method, including the appropriate collection of expressions (words), is essential.

On the specific attention to GBV, much research has been performed. Ref. [21] exhibited the use of machine learning methods on Twitter to study about the circumstances implicated in the #MeToo movement (an initiative to denounce GBV), mainly those related to business and marketing activities.

Ref. [22] presented the automatic detection and categorization of misogynous language in Twitter by using different supervised classifiers. Techniques like N-grams, linguistic, syntactic and embedding were used in order to build the feature space of the training data set. One of the main contributions of these work was to make available to the research community a data set of corpus of misogynistic tweets. Ref. [22], and similarly [23] who collect data from Twitter from frequent words in domestic violence, highlighted the importance of building a data set corpus of misogynistic tweets and consider the language regionalization, i.e., the data corpus should be in accordance with regional context [13].

Xue et al. [15] evidenced the viability of employing topic-modeling methods for data-mining on Twitter to identify GBV. An unsupervised algorithm to discover hidden topics in the tweets was used. Twitter messages were converted into a document-term matrix by applying the CountVectorizer method [24], in order to collect words that appear more frequently in domestic violence, which are related to GBV.

In Ref. [16], a deep neural network was applied to identify the risk factor associated with sexual violence on Twitter; however, it did not explain how the messages were pre-processed.

Mohammed et al. [17] recommended an array of unique features obtained from Twitter (based on network, activity type, user as well as the content of the tweet) for the detection of cyber-bullying (which has a direct relationship with GBV). Results showed an AUC of 0.943 indicating that this set of features provides an effective approach for detecting cyber-bullying.

In Ref. [18], an approach to automatic detection of hate expressions on Twitter was shown. Authors collected offensive or hateful expressions for hate speech detection. The pre-processing stage consisted of a cleaning up of the tweets, tokenization, generation of negation expressions (e.g., "not", "never", etc.) and detection of the broadcast of these words. In addition, a feature selection process was done.

Ref. [23] exhibited a technique for detection of xenophobia and misogyny in tweets by using computing methodologies. Authors created a suitable language resource for hate speech recognition in Spanish (Spain), highlighting the importance of language regionalization, i.e., whether it is Spanish from Spain or Mexico.

In [19], an Arabic offensive tweet detector was built. An inherent complexity to classify tweets is noticeable, which is in accordance with the particular language.

In the Mexican Spanish context in Twitter, a few works have been performed for automatic identification of GBV. Most of them have been focused on detection of aggressiveness. Alvarez-Carmona et al. [25] presented an overview of results from MEX-A3T competition (2018), which is addressed to automatic identification of aggressiveness in Mexican Spanish tweets. The competition included two tracks: in the first, author profiling, the aim is to identify the place of residence and occupation of the users; in the second, the goal was detection of aggressiveness in the message. Results showed 76.4% accuracy in the aggressiveness identification task. Results of the deep learning methods used in MEX-A3T did not overcome 68% accuracy [20]. Ref. [20] analyzed the performance of two deep learning models for automatic classification of aggressive Mexican Spanish tweets. It highlighted the low performance of studied deep learning neural networks to identify aggression in Mexican Spanish tweets, i.e., there are still open issues to better understand this topic, thus, they should be addressed.

Based on the previous works, two essential components were identified in the analysis of content in Twitter messages: (a) the suitable collect of expressions (words) related to the topic under study in accordance to regional context, and (b) the extraction features stage by simple techniques like the CountVectorizer method [24], which transforms tweet content into vectors by counting occurrences of each word in each tweet, but also the use of sophisticated methodologies like those presented in [18] or [25].

In relation to the pre-processing stage, it was noted that most of the works need a complex pre-processing or specialized group to manually tag the comments (or use small data sets).

As a relevant concern, it was observed that most recent advances are developed for the English language [25], but the few works performed for other languages agree with the importance of the regional context of the messages in their original tongue [19,23].

In this paper, a simple methodology to identify GBV in Mexican Spanish Twitter messages is studied, which includes three common extraction feature methods: CountVectorizer, TfidfVectorizer and HashingVectorizer. In contrast with other state-of-the-art works, our proposal does not employ a stage to collect expressions related to GBV, but only give to the classifier enough samples previously labeled by human volunteers of tweets containing evidence of GBV or not containing GBV. Thus, the significance of this work can be highlighted as follows:

1. This research contributes to the automatic detection of GBV in Mexican Spanish tweets (specifically contextualized to Mexican language jargon), which is a little faced issue, with the potential use of this work in the early attention of dangerous behaviors in the users.

2. It shows encouraging results in classification of tweets related to GBV. Area under the receiver operating characteristic (AUC) obtained is about 80% by using a deep neural network.
3. Feature extraction method used in this work is very simple, i.e., a minimal pre-processing of the data is needed to classify tweets, which only implies to clean (delete articles, numbers, symbols, conjunction and nonsense words) and tokenize (by means of CountVectorizer, TfidfVectorizer and HashingVectorizer methods) tweets' content.

2. Deep Learning Multilayer Perceptron

Deep learning neural networks are characterized by the increase of the network depth, i.e., the number of hidden layers; then, the multilayer perceptron is a general and intuitive architecture to be transformed to the deep learning multilayer perceptron (DL-MLP) with two or more hidden layers [26].

DL-MLP tries to find a relation between a set of input vectors x and labels id by modifying the parameters linking those sets. The output y_j is a function of x and weight w so that if w is modified, the difference z between the system output and target id could be minimized. DL-MLP uses two or more hidden layers constituted of nodes or neurons. Each neuron is connected with the neurons of the previous layer and the output signal is calculated by combining all the inputs from the preceding layer [27]. The connections between nodes use a neuronal weight (w) to modify the output signal before getting in the neuron; this transformation corresponds to multiply the respective signal (x_i) times the weight (w_i).

The use of multiple layers generates a more complex optimization problem, but gains a reduction in the number of nodes per layer inside the architecture [28]. However, the increase of the computational effort can be overcome by the availability of advanced frameworks like Spark [29] and Tensorflow [30] that provide tools to optimize the cost function of the perceptron. The use of such tools makes possible that the DL-MLP could be used increasingly in big-data problems [31,32], and also increases the capability of abstraction of DL-MLP to complex problems [28].

Usually, DL-MLPs are trained by means the back-propagation algorithm (based on the stochastic gradient descent) [33–35] and initial weights are randomly assigned. One of the most common algorithms of descending gradient optimization is Adam [36], which is based on adaptive estimation of first-order and second-order moments [37]. This algorithm reduces the error between the $f(x, w)$ and $\hat{f}(x, w)$.

Typically, DL-MLP includes different activation functions that modify the linear space to a nonlinear space of the samples x in each hidden layer, namely: Rectified Linear Unit (ReLU) $f(z) = max(0, z)$, tangent function $f(z) = tanh(z)$, Exponential Linear Unit (ELu) $f(z) = z \geq 0 \rightarrow z, z < 0 \rightarrow (e^z - 1)$ and sigmoid function $f(z) = 1/(1 + e^{-z})$.

3. Deep Learning for Natural Language Processing and Sentiment Analysis

The advent of the world wide web and search engines brought with it the emergence of natural language processing (NLP) [38], which allows a machine to process a natural human language and then translates it into a format that is processable and understandable to a computer [39]. This field has received a lot of attention due to the efficiency in language modeling. Some of the NLP models have been applied in various areas, as they provide great mechanisms to analyze text in real time, in addition to the reliability that they also demonstrate in different tasks [40].

Due to the rapid growth of the Internet, the use of social networks, forums, blogs and other platforms where people from all over the world share their ideas, opinions and comments on multiple topics, has increased. Politics, cinema, sports, music, among others, have given rise to a great deal of unstructured information [41]. For this reason, sentiment analysis has become one of the main challenges addressed by NLP, whose main objective is to extract feelings, opinions, attitudes and emotions from the users [42] through a series of methods, techniques and tools on the detection and extraction of subjective information

to detect the polarity of the text, that is, to determine if the given text is positive, negative or neutral [43].

Sentiment analysis has been positioned as one of the essential tools to transform the emotions and attitudes of a text into actionable and understandable information for a machine [44]. It is so important within the NLP that this area has been addressed at 3 different levels [42]: (1) the document level, focused on determining whether an opinion document expresses a positive or negative sentiment, (2) the sentence level, whose task is to check whether each sentence expresses a positive, negative or neutral opinion and (3) the aspect level, responsible for looking directly at the opinion itself.

To address the problems of sentiment analysis, previously, approaches based on machine learning algorithms and the sentiment lexicon have been used. However, these methods have limitations such as limited data, word order and a large number of tagged texts that make them ineffective for NLP tasks [45]. However, for some of these problems, models based on deep learning have been the solution, these methods have been gaining popularity, thus proving to be a better option to face the problem of sentiment analysis and this is attributed to the high performance they show in different tasks of the NLP [46].

For years, the implementation of a deep learning or pattern recognition system in NLP has required careful engineering and extensive experience to design a feature extraction system that can transform raw data into appropriate internal data or in a vector of characteristics that a learning subsystem, generally a classifier, could use to detect patterns [47]. Feature extraction, as a data preprocessing method in the learning algorithm, contributes to performance improvement. The extraction methods used for this task range from simple approaches, such as those based on the bag of words model (like CountVectorizer [24], TfidfVectorizer [48] or HashingVectorizer [49], to more sophisticated approaches, such as transformers [50–53].

3.1. Text Feature Extraction

The CountVectorizer method converts a document d into a numeric vector $d = \{u_1, u_2, \ldots, u_i, \ldots, u_T\}$, where (u_i) is the weight of the word with the number i in the document d. The feature i of the document will be the sum of the times that the word i appears in it; seen in another way, u_i will be made up of the frequency of appearance of each word i in the document d [48].

TfidfVectorizer method uses the CountVectorizer matrix and applies a term frequency-inverse document frequency transformation (TFIDF), which takes a frequency of the word i, and the inverse frequency of occurrences in the document d (Equation (1)), instead of the raw frequencies of occurrence of a token [54].

$$u_i = TF_i * IDF_i, \tag{1}$$

where the weight (u_i) is a function of TF_i (term frequency), i.e., the appearance frequency of the word i in a document d, and IDF_i (inverse document frequency) which is:

$$IDF_i = log(\text{Total of documents}/DF_i), \tag{2}$$

being DF_i (document frequency) the quantity of documents in which the word i appears at least once.

By using IDF, the weight of high frequency words that are not significant (like conjunctions, prepositions or common words) is reduced, because these kinds of words will appear in several documents allowing to identify those with specific relevance in certain documents.

HashingVectorizer implementation works in a similar way to CountVectorizer, but it employs the hashing trick to find the token string name to include integer index mapping, normalized as token frequencies. Thus, there is no way to compute the inverse transformation, i.e., it does not consider inverse document frequency. However, it is very efficient for large data sets [49].

The CountVectorizer, HashingVectorizer and TfidfVectorizer methods can use different forms of assigning the number of the words included in a token (this parameter is N_{gram}). In the present work, tokens with 1, 2 or 3 words were used, which can give more relations between the pattern of the data.

4. Methodology

The methodological aspects of the work are exhibited in this section. Details about data collection, pre-processing, classifier parameters and assessment test are explained in order to allow the replication of the experiments. The source code for this work is accessible through https://github.com/ccastore/GenderViolence (accessed on 1 January 2021).

4.1. Data Collection

Data were collected by using the *twlets* (http://twlets.com) tool. Twitter messages were collected from 18–19 May 2019, taking tweets comments in Spanish language and located in Mexico (coordinates −118.599, 14.388 to −86.493, 32.718). In order to select tweets related to GBV, messages from individual users, companies and organizations that contained words or phrases related to diverse forms of possible GBV were selected. In addition, news pages and political figures were considered.

A total of 1,857,450 messages were retrieved from Twitter. 61,604 of them were manually tagged by human volunteers as follows: messages referring to GBV (those containing possible intention of GBV) and messages not referring to GBV, resulting in 1604 positive and 60,000 negative tweets.

4.2. Data Pre-Processing

Once the messages were retrieved from the Twitter stream, they were pre-processed to transform the input text to a normalized, comprehensible model of numbers sequence, proceeding as follows:

- Cleaning. Deletion of URLs (starting with "http://" or "https://"), tags ("@user"), articles and unrelated expressions (for example words written in languages outside of ANSI coding), exclamation marks, question marks, full stop marks, quotes and others symbols.
- Convert text to uppercase.
- Transform text to matrix of numbers, using CountVectorizer [24], TfidfVectorizer [48] and HashingVectorizer [49] methods, using a N_{gram} token with 1, 2 or 3 words.

Finally, a matrix obtained by CountVectorize, TfidfVectorizer and HashingVectorizer methods were used to build and test the classifier. For this, the hold-out method [27] was applied; it randomly split the original matrix on training (TDS) 70% and testing (TS) 30% data sets, where $\text{TDS} \cap \text{TS} = \emptyset$.

4.3. Sampling Methods

Oversampling methods are popular and successful techniques to deal with the class imbalance [55]. The most common algorithms are: (a) Random Over Sampling (ROS), that randomly duplicates samples from the minority class to mitigate the class imbalance, and (b) SMOTE, which produces artificial samples in the minority class by interpolation of near occurrences [56]. Specifically, for each minority class, they find the k intra-class nearest neighbors and generate synthetic samples in the direction of those nearest neighbors. In this work, k was set to five in SMOTE (as in Ref. [57]) and ROS and SMOTE were applied to the data set to achieve a relatively balanced class distribution.

In particular, for this work, TDS obtained from CountVectorize, TfidfVectorizer and HashingVectorizer methods contains 1122 GBV and 42,000 non-GBV samples (see Sections 4.1 and 4.2); thus, the resultant over-sampled TDS by SMOTE and ROS is composed of 42,000 GBV and 42,000 non-GBV samples approximately, i.e., those methods balance the class distribution.

4.4. Neural Network Set-Up

DL-MLP was developed on Tensorflow 2.0 and Keras 2.3.1, and Adam algorithm [36] was employed to train it. The Adam algorithm is used to calculate the adaptation of the learning rate for each parameter, storing an exponentially decreasing average of past gradients [30]. The learning rate (η) was established as 0.0006, meanwhile the stopping criterion was 20 epochs with a batch size of 150.

DL-MLP was set-up through of the trial and error method, which is usual in neural network environments. For this, we randomly take from TDS a subset ST (about of 20%), that was split into ST_{train} and ST_{test}, where ST $\subseteq$ TDS, and $ST_{train} \cap ST_{test} = \emptyset$. In this process, we use ST_{train} and ST_{test} to assess different configurations of numbers of hidden layers and neurons by layer, and the topology that produced the best classification result was selected. Final architecture was a DL-MLP with six hidden layer and sigmoid activation functions, and the number of hidden nodes for each layer was set as $6, 6, 5, 5, 4$ and 3, respectively.

4.5. Classifier Performance

Classification accuracy and error rate are widely used to assess the performance of learning models. Nevertheless, in class imbalanced scenarios these measures are biased to majority classes or more represented classes (for example, in this work, there are much more non-GBV tweets than GBV tweets). Thus, others metrics should be used.

The receiver operating characteristic curve (ROC) is an appropriate instrument to evaluate the classifiers performance on imbalance scenarios, according to the trade-offs between benefits (true positives) and costs (false positives). The quantitative depiction of ROC is the area under the curve (AUC), calculated as $AUC = (sensitivity + specificity)/2$, where $sensitivity$ is the percentage of correctly predicted $positive$ samples, and $specificity$ is the percentage of negative samples predicted correctly [58] (see Table 1). In this work, $sensitivity$, $specificity$ and the AUC were used to measure the effectiveness of deep learning neural network to identify GBV on Mexican tweets.

Table 1. Confusion matrix for binary classification.

		Predicted Class		
		Positive	**Negative**	
True class	Positive	True Positive (tp)	False Negative (fn)	$sensitivity$ $\frac{tp}{tp+fn}$
	Negative	False Positive (fp)	True Negative (tn)	$specificity$ $\frac{tn}{tn+fp}$

5. Experimental Results and Discussion

The main experimental results in identifying GBV in Mexican tweets are presented in this section. Table 2 summarizes the results in term of features obtained for extraction methods, classification performance measures $sensitivity$, $specificity$ and AUC.

The number of features for HashingVectorizer method was calculated as trial-error for this work. Several values were tested and the best value was determined to be 350 features. For CountVectorizer and TfidVectorizer methods the default parameters were used. Thus, the employed algorithms settled on number of features (see Section 3.1).

In Table 2, is noted that the class imbalance severely affects the classifier overall performance. Results obtained without using any sampling method indicate that the classifier does not learn the minority class (GBV tweets). Thus, this approach is not appropriate to identify GBV on Mexican tweets.

Table 2. Classification results obtained after applying feature extraction and two sampling methods, using a deep learning multilayer perceptron (DL-MLP) as classifier. Best results (in bold) and the best *AUC* for each sampling method (marked with a star) are also indicated.

Sampling	Feature Extraction	N_{gram}	Features	Specificity	Sensitivity	AUC
N/A	CountVectorizer	1	1021	1.0000	0.0000	0.5000
		1, 2	2124	1.0000	0.0000	0.5000
		1, 2, 3	2915	1.0000	0.0000	0.5000
	HashingVectorizer	1	350	1.0000	0.0000	0.5000
		1, 2	350	1.0000	0.0000	0.5000
		1, 2, 3	350	1.0000	0.0000	0.5000
	TfidVectorizer	1	1027	1.0000	0.0000	0.5000
		1, 2	2152	1.0000	0.0000	0.5000
		1, 2, 3	2836	1.0000	0.0000	0.5000
SMOTE	CountVectorizer	1	1016	0.8659	**0.7562**	**0.8111** *
		1, 2	2143	0.8906	0.6883	0.7895
		1, 2, 3	2879	**0.8921**	0.7022	0.7972
	HashingVectorizer	1	350	0.7125	**0.8067**	**0.7596**
		1, 2	350	**0.7235**	0.7490	0.7363
		1, 2, 3	350	0.6773	0.7571	0.7172
	TfidVectorizer	1	1014	0.8714	**0.7449**	**0.8082**
		1, 2	2130	0.8970	0.6838	0.7904
		1, 2, 3	2865	**0.9012**	0.6712	0.7862
ROS	CountVectorizer	1	1018	0.8926	**0.7241**	**0.8083** *
		1, 2	2108	**0.9120**	0.6506	0.7813
		1, 2, 3	2881	0.8980	0.6779	0.7880
	HashingVectorizer	1	350	0.7017	**0.8339**	**0.7678**
		1, 2	350	**0.7500**	0.7384	0.7442
		1, 2, 3	350	0.7100	0.7108	0.7104
	TfidVectorizer	1	1010	0.8861	**0.7291**	**0.8076**
		1, 2	2137	**0.9217**	0.6279	0.7748
		1, 2, 3	2933	0.9128	0.6544	0.7836

Results obtained by employing sampling methods (ROS and SMOTE) indicate that the DL-MLP is effective to learn GBV tweets. However, Table 2 shows that when the minority class has a best performance the majority class performance is reduced, as it can be observed from the *sensitivity* and $specificity$ values. For example, on ROS with HashingVectorizer, and $N_{gram} = 1$, the high value of *sensitivity* is obtained simultaneously with the worst $specificity$ value. A similar performance is observed with SMOTE.

AUC gives a better understanding of the classifier performance for both classes than the *sensitivity* and $specificity$ measures. High *AUC* values imply a best trade-off between benefits (GBV tweets correctly classify) and costs (GBV tweets incorrectly classify). In this respect, it is observed in Table 2 that CountVectorizer with $N_{gram} = 1$ presents the best AUC value. Then, it is suggested that the simplest method obtains the highest score.

A trend in the studied feature extraction methods is that the better values of $specificity$ and *AUC* are obtained when the $N_{gram} = 1$ is used than when applying other values. In other words, experimental results of this work notice that to identify GBV on Mexican tweets, the employment of only the mean of each word is an effective approach.

Table 2 shows that the worst AUC values correspond to the HashingVectorizer method. However, this method was developed to work with big data sets; then, it could explain this behavior because the data set used in this research contains only 61,604 samples.

Finally, with respect to the number of features obtained for the extraction methods (CountVectorizer, HashingVectorizer and TfidVectorizer), there is not evidence in the

obtained results about the relationship between the number of features used and the classifier performance.

6. Conclusions

GBV is a problem that exist on the social network Twitter. Many works have been performed to deal with it along with related issues like hate speech, xenophobia, misogyny, domestic violence, among others. A main stage of that research is the collection of a corpus of words related to particular situations and language. In the Mexican Spanish context, few works have been developed to deal with GBV in Twitter messages and the language regionalization has been recognized as critical. In addition, results of the most of those works need to be improved.

Thus, in this paper, a study to identify GBV on Twitter messages in Mexico is presented. Three common feature extraction methods were used (CountVectorizer, TfidfVectorizer and HashingVectorizer) together with a deep learning multilayer perceptron as the classifier. A data set containing 1604 GBV tweets and 60,000 non-GBV tweets from a total of 1,857,450 messages retrieved from Twitter social network were labeled by human volunteers as GBV or non-GBV messages to train and test the proposed scheme.

Experimental results showed that the class imbalance problem significantly affects the classification of GBV messages. In this sense, oversampling methods, mainly ROS and SMOTE, are effective to overcome this problem. Thus, it was noticed that the CountVectorizer method (and a sampling method) allows DL-MLP to identify GBV on Mexican tweets with about 80% AUC. As a remarkable result, it is worth to mention that only a minimal data set pre-processing was applied to obtain important results. TfidfVectorizer and HashingVectorizer methods show competitive results, but CountVectorizer presented a trend to obtain the best results.

Results of this research give evidence that giving enough labeled samples, obtained from Mexican Spanish Twitter messages and transformed by simple feature extraction method like CountVectorizer to DL-MLP, can produce improved classification results.

GBV is an issue that must be immediately addressed. In this sense, this study could potentially contribute to deal with gender violence in Mexico because it provides the analysis of useful tools to identify GVB in online social networks despite the language jargon. However, the classification results should be improved because the rate of GBV tweets that have been predicted correctly (*sensitivity*) is still low. The analysis in specific variants of Spanish of certain tools for the detection of GBV could help to push further research needed to improve the studied strategies on the identification of GBV in Twitter messages in Mexican Spanish.

Thus, future work should be addressed mainly to reduce the human effort to label the GBV texts and to test advanced deep learning models in order to increase the classifier performance, including more sophisticated natural language processing techniques. Currently, we work in an application on streaming to identify GVB, which uses a DL-MLP with a rejection option, i.e., when the classifier has doubts about a tweet's content it is rejected and sent to a human volunteer to be targeted and included in the training data set. We consider that this procedure will allow to improve the classifier performance.

Author Contributions: C.M.C., R.A.: conceptualization, methodology and experiment; I.M.A.: conceptualization and review; E.R.: supervision; R.A., E.E.G.-G.: writing—review and editing. O.V.: Experiment. All authors have read and agreed to the published version of the manuscript.

Funding: This research did not receive external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This work has been partially supported under grants of project 5046/2020CIC from UAEMex.

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

References

1. Sweeney, B.N. Gender-Based Violence and Rape Culture. In *Companion to Women's and Gender Studies*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2020; Chapter 15, pp. 285–302. [CrossRef]
2. Russo, N.F.; Pirlott, A. Gender-based violence: Concepts, methods, and findings. *Ann. N. Y. Acad. Sci.* **2006**, *1087*, 178–205. [CrossRef]
3. UN. *Declaration on the Elimination of Violence against Women*; UN General Assembly: New York, NY, USA, 1993.
4. Hernández Castillo, R.A. Racialized Geographies and the "War on Drugs": Gender Violence, Militarization, and Criminalization of Indigenous Peoples. *J. Lat. Am. Caribb. Anthropol.* **2019**, *24*, 635–652. [CrossRef]
5. Sanchez, G. Victimization, Offending and Resistance in Mexico: Toward Critical Discourse and Grounded Methodologies in Organized Crime Research. *Vict. Offenders* **2020**, *15*, 390–393. [CrossRef]
6. John, N.; Casey, S.E.; Carino, G.; McGovern, T. Lessons Never Learned: Crisis and gender-based violence. *Dev. World Bioeth.* **2020**. [CrossRef]
7. Domínguez, M.A.; Palomeque, D.; Carrillo, J.M.; Valverde, J.M.; Duque, J.F.; Pérez, B.; Pérez-Aloe, R. Voice-Controlled Assistance Device for Victims of Gender-Based Violence. In *Developments and Advances in Defense and Security*; Rocha, Á., Pereira, R.P., Eds.; Springer: Singapore, 2020; pp. 397–407. [CrossRef]
8. Rituerto-González, E.; Mínguez-Sánchez, A.; Gallardo-Antolín, A.; Peláez-Moreno, C. Data Augmentation for Speaker Identification under Stress Conditions to Combat Gender-Based Violence. *Appl. Sci.* **2019**, *9*, 2298. [CrossRef]
9. Rodrǵuez-Rodrǵuez, I.; José-Víctor, R.; Domingo-Javier, P.-Q.; Heras-González, P.; Chatzigiannakis, I. Modeling and Forecasting Gender-Based Violence through Machine Learning Techniques. *Appl. Sci.* **2020**, *10*, 8244. [CrossRef]
10. Andrada, A.V.; Sanchez, J.J.; Sánchez-Serrano, J.L.S. Gender Violence and New Technologies. In *Qualitative and Quantitative Models in Socio-Economic Systems and Social Work*; Springer International Publishing: Cham, Switzerland, 2020; pp. 375–390. [CrossRef]
11. Vyawahare, M.; Chatterjee, M. Taxonomy of Cyberbullying Detection and Prediction Techniques in Online Social Networks. In *Data Communication and Networks*; Jain, L.C., Tsihrintzis, G.A., Balas, V.E., Sharma, D.K., Eds.; Springer: Singapore, 2020; pp. 21–37. [CrossRef]
12. Paschalides, D.; Stephanidis, D.; Andreou, A.; Orphanou, K.; Pallis, G.; Dikaiakos, M.D.; Markatos, E. MANDOLA: A Big-Data Processing and Visualization Platform for Monitoring and Detecting Online Hate Speech. *ACM Trans. Internet Technol.* **2020**, *20*. [CrossRef]
13. Gutiérrez-Esparza, G.O.; Vallejo-Allende, M.; Hernández-Torruco, J. Classification of Cyber-Aggression Cases Applying Machine Learning. *Appl. Sci.* **2019**, *9*, 1828. [CrossRef]
14. Bellmore, A.; Calvin, A.J.; Xu, J.M.; Zhu, X. The five Ws of bullying on Twitter: Who, What, Why, Where, and When. *Comput. Hum. Behav.* **2015**, *44*, 305–314. [CrossRef]
15. Xue, J.; Chen, J.; Gelles, R. Using Data Mining Techniques to Examine Domestic Violence Topics on Twitter. *Violence Gend.* **2019**, *6*, 105–114. [CrossRef]
16. Khatua, A.; Cambria, E.; Khatua, A. Sounds of Silence Breakers: Exploring Sexual Violence on Twitter. 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. 397–400. [CrossRef]
17. Al-garadi, M.A.; Varathan, K.D.; Ravana, S.D. Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. *Comput. Hum. Behav.* **2016**, *63*, 433–443. [CrossRef]
18. Watanabe, H.; Bouazizi, M.; Ohtsuki, T. Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection. *IEEE Access* **2018**, *6*, 13825–13835. [CrossRef]
19. Mubarak, H.; Rashed, A.; Darwish, K.; Samih, Y.; Abdelali, A. Arabic Offensive Language on Twitter: Analysis and Experiments. *arXiv* **2020**, arXiv:2004.02192.
20. Frenda, S.; Banerjee, S. Deep Analysis in Aggressive Mexican Tweets. In Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) Co-Located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, 18 September 2018; Volume 2150, pp. 108–113.
21. Reyes-Menendez, A.; Saura, J.R.; Ferrõ, F. Marketing challenges in the #MeToo era: Gaining business insights using an exploratory sentiment analysis. *Heliyon* **2020**, *6*, e03626. [CrossRef] [PubMed]
22. 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. [CrossRef]
23. Plaza-Del-Arco, F.M.; Molina-González, M.D.; Ureña López, L.A.; Martín-Valdivia, M.T. Detecting Misogyny and Xenophobia in Spanish Tweets Using Language Technologies. *ACM Trans. Internet Technol.* **2020**, *20*. [CrossRef]
24. Garreta, R.; Moncecchi, G. *Learning Scikit-Learn: Machine Learning in Python*; Packt Publishing: Birmingham, UK, 2013.

25. Aragón, M.E.; Alvarez, M.A.; Montes-y-Gómez, M.; Escalante, H.J.; Villaseñor, L.; Moctezuma, D. Overview of MEX-A3T at IberLEF 2019: Authorship and Aggressiveness Analysis in Mexican Spanish Tweets. In Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao, Spain, 24 September 2019; Volume 2421, pp. 478–494.
26. Krig, S. Feature Learning and Deep Learning Architecture Survey. In *Computer Vision Metrics*; Springer International Publishing: Cham, Switzerland, 2016; pp. 375–514. [CrossRef]
27. Haykin, S. *Neural Networks. A Comprehensive Foundation*, 2nd ed.; Pretince Hall: Upper Saddle River, NJ, USA, 1999.
28. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
29. Zaharia, M.; Xin, R.S.; Wendell, P.; Das, T.; Armbrust, M.; Dave, A.; Meng, X.; Rosen, J.; Venkataraman, S.; Franklin, M.J.; et al. Apache Spark: A Unified Engine for Big Data Processing. *Commun. ACM* **2016**, *59*, 56–65. [CrossRef]
30. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. TensorFlow: A System for Large-scale Machine Learning. In *OSDI'16: Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*; USENIX Association: Berkeley, CA, USA, 2016; pp. 265–283.
31. Guo, Y.; Liu, Y.; Oerlemans, A.; Lao, S.; Wu, S.; Lew, M.S. Deep learning for visual understanding: A review. *Neurocomputing* **2016**, *187*, 27–48. [CrossRef]
32. Reyes-Nava, A.; Sánchez, J.; Alejo, R.; Flores-Fuentes, A.; Rendón-Lara, E. Performance Analysis of Deep Neural Networks for Classification of Gene-Expression microarrays. In *MCPR 2018: Pattern Recognition—10th Mexican Conference*; Springer: Cham, Switzerland, 2018; Volume 10880, pp. 105–115. [CrossRef]
33. Li, D.; Huang, F.; Yan, L.; Cao, Z.; Chen, J.; Ye, Z. Landslide Susceptibility Prediction Using Particle-Swarm-Optimized Multilayer Perceptron: Comparisons with Multilayer-Perceptron-Only, BP Neural Network, and Information Value Models. *Appl. Sci.* **2019**, *9*, 3664. [CrossRef]
34. Pacheco-Sánchez, J.; Alejo, R.; Cruz-Reyes, H.; Álvarez-Ramírez, F. Neural networks to fit potential energy curves from asphaltene-asphaltene interaction data. *Fuel* **2019**, *236*, 1117–1127. [CrossRef]
35. Looney, C.G. *Pattern Recognition Using Neural Networks*, 1st ed.; Oxford University Press: Oxford, UK, 1997.
36. Ruder, S. An overview of gradient descent optimization algorithms. *arXiv* **2016**, arXiv:1609.04747.
37. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
38. Rao, J.S. A Survey on Sentiment Analysis and Opinion Mining. In *Proceedings of the International Conference on Advances in Information Communication Technology and Computing*; Association for Computing Machinery: New York, NY, USA, 2016; [CrossRef]
39. Rajput, A. Chapter 3—Natural Language Processing, Sentiment Analysis, and Clinical Analytics. In *Innovation in Health Informatics*; Lytras, M.D., Sarirete, A., Eds.; Next Gen Tech Driven Personalized Med and Smart Healthcare, Academic Press: New York, NY, USA, 2020; pp. 79–97. [CrossRef]
40. Devika, M.D.; Sunitha, C.; Ganesh, A. Sentiment Analysis: A Comparative Study on Different Approaches. *Procedia Comput. Sci.* **2016**, *87*, 44–49. [CrossRef]
41. Dashtipour, K.; Ieracitano, C.; Morabito, F.C.; Raza, A.; Hussain, A. An Ensemble Based Classification Approach for Persian Sentiment Analysis. In *Progresses in Artificial Intelligence and Neural Systems*; Springer: Singapore, 2021; pp. 207–215. [CrossRef]
42. Al-Bayati, A.; Al-Araji, A.; Ameen, S. Arabic Sentiment Analysis (ASA) Using Deep Learning Approach. *J. Eng.* **2020**, *26*, 85–93. [CrossRef]
43. Mantyla, M.; Graziotin, D.; Kuutila, M. The Evolution of Sentiment Analysi—A Review of Research Topics, Venues, and Top Cited Papers. *Comput. Sci. Rev.* **2016**, *27*, 16–32. [CrossRef]
44. Mishev, K.; Gjorgjevikj, A.; Vodenska, I.; Chitkushev, L.T.; Trajanov, D. Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers. *IEEE Access* **2020**, *8*, 131662–131682. [CrossRef]
45. Lin, P.; Luo, X.; Fan, Y. A Survey of Sentiment Analysis Based on Deep Learning. *Int. J. Comput. Inf. Eng.* **2020**, *14*, 473–485.
46. Kapil, P.; Ekbal, A.; Das, D. Investigating Deep Learning Approaches for Hate Speech Detection in Social Media. *arXiv* **2020**, arXiv:2005.14690.
47. Liang, H.; Sun, X.; Sun, Y. Text feature extraction based on deep learning: A review. *J. Wirel. Commun. Netw.* **2017**, *2017*, 211. [CrossRef]
48. Eshan, S.C.; Hasan, M.S. An application of machine learning to detect abusive Bengali text. In Proceedings of the 2017 20th International Conference of Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 22–24 December 2017; pp. 1–6.
49. Hasan, M.; Islam, I.; Hasan, K.M.A. Sentiment Analysis Using Out of Core Learning. In Proceedings of the 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox'sBazar, Bangladesh, 7–9 February 2019; pp. 1–6.
50. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
51. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; Le, Q.V. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv* **2019**, arXiv:1906.08237.
52. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv* **2019**, arXiv:1909.11942.
53. Clark, K.; Luong, M.T.; Le, Q.V.; Manning, C.D. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *arXiv* **2020**, arXiv:2003.10555.

54. Uğuz, H. A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. *Knowl.-Based Syst.* **2011**, *24*, 1024–1032. [CrossRef]
55. Abdi, L.; Hashemi, S. To Combat Multi-class Imbalanced Problems by Means of Over-sampling Techniques. *IEEE Trans. Knowl. Data Eng.* **2016**, *28*, 1041–4347. [CrossRef]
56. Fernandez, A.; Garcia, S.; Herrera, F.; Chawla, N.V. SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *J. Artif. Intell. Res.* **2018**, *61*, 863–905. [CrossRef]
57. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
58. Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. [CrossRef]