Understanding Violence Against Women in Digital Space from a Data Science Perspective

I. INTRODUCTION

Mexico is one of the most violent countries in the world, its recent history has been marked by generalized violence. One of its most underestimated practices is Violence Against Women, which can end up in murder. This latter has been an issue even before this millennium. Thus, when hundreds of women in Ciudad Juárez went missing and found murdered, Lourdes Portillo filmed a documentary, Señorita Extraviada, giving an excellent overview of what was happening, [14].

Violence Against Women has its most extreme and lethal expression in Feminicides, which are a consequence of the systemic Human Rights violations in public and private sectors, and are conformed by the set of misogynistic behaviors that can carry Social and State impunity. Also, this can culminate in homicide and other violent deaths of women, [5].

Since then, subsequent investigations showed that this is a pervasive problem throughout the country. Numbers from the Secretariado Ejecutivo del Sistema Nacional de Seguridad Pública have shown an increase of 143.1 feminicides per year in a linear relation for a $R^2$ of 0.986, an increase of 1118.7 rapes per year in a linear relation for a $R^2$ of 0.917, and more increments in other crimes related to family violence, emergency calls, and sexual assaults, [6].

Until a few years ago, Violence Against Women took place both in public spaces, such as streets, and private spaces, such as households. However, due to advances in Technology and Communications, there is a new interaction space where it takes place, the one given by the internet platforms, social networks, e-mails, and any digital communications: the Digital Space.

Although it is not the most extreme and lethal expression of violence, Digital Space reproduces Violence Against Women in other more symbolic expressions, such as speech and discourse, maintaining the structural foundations of violence.

The proposition of this article is that the application of Data Science approaches such as Text Mining, Sentiment Analysis, Pattern Recognition, and Unsupervised Learning, can produce pattern finding and classification of the Digital Violence Against Women phenomena and its violent social discourses in several ways. First, enabling semi-automated consolidation of relevant data from a wide range of heterogeneous, structured and unstructured data. Second, offering insights on temporal trends and frequent patterns. Finally, by discovering associations via analytic models learned from the data, and by involving Machine Learning in the learning algorithms that infer associations between variables using the gathered data.

In this sense, we propose two research questions:

- What is an appropriate framework to work with the Digital Violence Against Women phenomena?
- Does a Data Science approach can contribute in the understanding of the Digital Violence Against Women phenomena?
II. CONNECTING THE SOCIAL PROBLEM WITH DATA SCIENCE

This article seeks to understand the big social problem of Digital Violence Against Women. We try to understand it by analyzing language, speech, and trends in social networks. In order to address this problem, we appeal to social data science approaches that will allow us to properly understand it and to propose possible solution paths for public policies, social projects and efforts.

In the following subsections of the Problem Definition, we will first set the framework to understand Digital Violence Against Women phenomena in order to adequately attend it through Data Science approaches. Then, we present the understanding of the problem from a technical point of view, so that the Data Science approach and tools can be clearly stated.

A. Understanding Violence

Digital Violence Against Women is still a new concept that needs to be understood from different angles. Thus, we first present the concepts of Symbolic Violence and Hate Speech, based on Bourdieu, [4], and Richardson-Self, [16]; then we go through the research of Rita Segato regarding Violence Against Women, [17]; and finally we look at the recent legally defined definition for Digital Violence Against Women in Mexican Laws.

1) Symbolic Violence and Hate Speech: It is important to state that violence is conformed by not just physical and visible acts, but also by behaviors, speech, thoughts and structures that are not tangible and cannot be seen. Consequently, we approach Symbolic Violence following Pierre Bourdieu in his book The Male Domination, [4], as mental schemes that are the product of the assimilation of power relations and that are explained by the imposition of a universal objectivity of social structures. Thus, it works as the matrices of perceptions, thoughts and actions of all society members.

In this article we will seek to understand Digital Violence Against Women from a Symbolic Violence perspective and, specifically, through language and speech. However, it is worth noting that, when we work with Symbolic Violence we are not minimizing the role of physical violence, nor its implications in the lives of women. Moreover, we do not pretend to excuse the aggressors or naturalize structural violence and power relations.

In addition, misogynistic speech will be treated as a specific type of hate speech according to Louise Richardson-Self in her article Woman-Hating: On Misogyny, Sexism, and Hate Speech, [16]. This is different from cyberbullying, which is bullying that takes place over digital devices targeting specific individuals. Hate speech is often characteristically hostile, and does certain things, such as silence, malign, disparage, humiliate, intimidate, incite violence, discriminate, vilify, degrade, persecute, threaten, and the like. Likewise, hate speech is typically understood to be an expressive conduct that targets not individuals, but real or imagined group traits, such as race, religion, sexual orientation, disability, gender status, and gender identity, to mention some.

Furthermore, hate speech is taken to express hostility to, and about, historically and contemporary oppressed groups. Such focus is by no means accidental. This speech constitutes an oppressive act, therefore it is not simply that it picks out an oppressed group, but that it enacts, reinforces and perpetuates oppression. Hence, this research will use language and speech as a dimension of violence that actively and effectively reinforces oppression and attacks individuals as members of oppressed groups.

2) Violence Against Women and its particularities: As we stated, Violence Against Women is a phenomena that has been increasing over the last decades. Even though there has been important legislation that addresses this problem, it has not been adequately attended due to the growing numbers of violence against women in all of its forms.

In the book The war against women, Rita Segato explains that all the crimes against women are contained by a great gender symbolic structure: patriarchy. As a result, every crime has the gender scheme in its insides, [17]. Overall, Rita Segato focuses her arguments on the lethal crimes against women. However, since the patriarchal structure penetrates every place and interaction, non-lethal crimes such as the Digital Violence Against Women are also loaded with the great gender symbol.

In lethal crimes, this gender symbolic can be read in woman’s body appropriation, but in non-lethal crimes, such as symbolic violence, gender symbolic can be read in Woman’s image and concept appropriation. This gives importance to the analysis of language and speech in social networks and to the violence rhetoric that is constructed, which affects both women as individuals and women as a social group.

Then, it is important to understand the distinct crimes against women in its particularities. If there are not specific protocols to attend them, then those crimes cannot be adequately addressed, although they are visible. In the case of digital violence, cyberbullying and digital violence against women cannot be investigated with the same methodology. As a result, judiciary authorities cannot analyze them with the same comprehension patterns. Hence, it is worth addressing Digital Violence Against Women with Data Science approaches in order to contribute to the understanding of its particularities so that there can be specific protocols to attend it.

3) Digital Violence Against Women: To set a ground for the Digital Violence Against Women phenomena guiding the arguments of this article, we will appeal to the definition found in the Mexican Laws which, in terms of Women Human Rights, has become a reality after the efforts of Organized Civil Society.

The last 26th of November of 2019, our Chamber of Federal Elected Legislators approved the addition of Digital Violence Against Women as fraction VI to article 6 of the General Law on Women’s Access to a Life Free of Violence at the federal level. Hence, Digital Violence Against Women is legally defined in Mexico as:
Data Science is a modern approach that seeks to make use of data to comprehend any kind of problems. The idea is to benefit from data while combining computer science, math, statistics, and domain knowledge in order to find patterns or relations that may be useful for proposing new approaches to attend the world problematic.

So, to show which is the Data Science approach defined to address the chosen Digital Violence Against Women problem we first expose the Data Sources and type of information that will be used in this project. Next, we state the technical scope and hence the Data Science techniques that we are going to use. Finally, we present the specific tools which we will work with along the project.

1) Existence of Data Sources: There are different types of data such as text, integer, date, and many more, as well as different types of data sources such as databases, flat files, live measurements, and some others. In our case, the proposed scope to address Digital Violence Against Women guide us to specifically use unstructured text and integer types of data from the social network databases.

The social network from which the dataset will be obtained is Twitter since it is a microblogging social network that is used to express thoughts as well as to have close interactions between people. In addition, Twitter has a high number of monthly active users which provides an ideal environment for analyzing violent social discourses.

Besides, Twitter offers access to their public conversational data for academic research and puts within reach several tools that can be used to read, analyze and visualize their data.

Furthermore, in an effort to provide the space for an analysis that can be deep and complete, we establish the following initial series of variables or features to disaggregate data:

- date
- text
- length of text
- contributors
- coordinates
- place
- hashtags
- symbols
- URLs
- user mentions
- media
- favorite count
- retweet count
- possibly sensitive
- geo
- source
- is reply
- is quote
- user creation date
- user description
- user favourites count
- user followers count
- user friends count
- user listed count
- user statuses count

Most of these features refer to text data, which will be fundamental to this research. Depending on how it is represented, text data can be viewed as either a string or as multidimensional data. Here we are going to use a vector-space representation for text data since using string representation makes it difficult to use ordering between words in a large scale application, and due to the ease of use of vector-space representation in frequencies of the words analysis [1]. However, in order to counteract data sparsity we will use a bag-of-words representation.

2) Data Science scopes to obtain knowledge: The proposed Data Science techniques to tackle Digital Violence Against Women through unstructured text and integer types of data are: (1) Text Mining, (2) Sentiment Analysis, (3) Pattern Recognition, and (4) Unsupervised Learning.

The utilization of Text Mining comes from the type of data that contain social networks, in our particular case, Twitter. For the purpose of setting a common language, we will take the concepts used in the book Data Mining: the text book, written by Chargu Aggarwal, [1]. The set of features or dimensions of text is going to be referred as lexicon. A collection of documents or tweets, in our case, is going to be referred as corpus. The text in the documents or tweets, which is a discrete sequence of words, is going to be referred as string but will be represented as multidimensional data in the form of frequency-annotated bag-of-words. Finally, the words in texts are going to be referred as terms.

The second Data Science technique that we are going to use is Sentiment Analysis, which Ronen Feldman defines in his article Techniques and Applications for Sentiment Analysis, [7], as the task of finding opinions of authors to specific entities. We have to take into consideration that decision-making is affected by the general thoughts of people, particularly in social networks, including Twitter, where there are waves of
discourses. There are usually two main sentence classifications, objective sentences that contain factual information and subjective sentences that contain explicit opinions, beliefs and views. There are three different levels of sentiment analysis that we are going to use in this research, the document-level, the sentence-level, and the aspect-level.

Pattern recognition is the third Data Science technique mentioned to tackle Digital Violence Against Women and is applied under the field of Machine Learning. According to Christopher Bishop in his book Pattern Recognition and Machine Learning, [3], pattern recognition works to discover regularities in data through the use of computer algorithms in order to classify data into different categories. Machine Learning can be divided into two different types of pattern recognition, Supervised learning which refers to the applications where the training data has its corresponding classification, and Unsupervised learning, which is the fourth Data Science stated technique and refers to applications where the training data does not have a corresponding classification.

In the book An Introduction to machine learning, Miroslav Kubat explains that Unsupervised learning seeks to discover useful properties in available data where these properties are mainly grouped in clusters. The text data from social network Twitter will be unstructured text with no classifications, therefore Unsupervised learning becomes an adequate approach to find useful patterns and properties in data.

### B. Framework

In this section we present the Framework for this research, which shows the theoretical concepts on the field that will be the basis of our proposed model.

There is a wide variety for the theoretical framework that can be utilized. There is the research by Gutiérrez-Esparza et. al., [8], this research classifies situations of cyber-aggression on social networks, specifically for Spanish-language users of Mexico. They applied Random Forest, Variable Importance Measures (VIMs), and OneR to support the classification of offensive comments in three particular cases of cyber-aggression: racism, violence based on sexual orientation, and violence against women. Their results reach more than 90% of accuracy. They utilized a Random Forest classification algorithm, which is a set of decision trees that outputs a predictive value and is robust against over-fitting. Their algorithm has two parameters, mintry, which represents the number of input variables chosen at random in each division and, ntree, which represents the number of trees. The function of the Random Forest classification algorithm is the following.

<table>
<thead>
<tr>
<th>Algorithm 1: Random Forest for classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Let (N) be the number of training cases, and (M) the number of variables.</td>
</tr>
<tr>
<td>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).</td>
</tr>
<tr>
<td>3 Vary and choose (N) for the tree (completing all training cases) and use the rest of the test cases to estimate the error.</td>
</tr>
<tr>
<td>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</td>
</tr>
</tbody>
</table>

Moreover, they utilized a Mean Decrease Impurity importance (MDI) with a Gini index to classify comments in cases of cyber-aggression, which is defined by:

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

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.

Another relevant work to take in consideration is the one done by Liebovitch et. al., [13], which uses data science methods to measure quantitative values of the peace factors from structured and unstructured (social media) data. They also developed a graphical user interface for the mathematical model so that social scientists or policy makers, can by themselves, explore the effects of changing the variables and parameters in these systems. This gives another dimension that can be explored, the one of the digital violence in social media data. Likewise, it can set the precedents for presenting interfaces to the public in general as of our objective of using an open data platform.

There are also two 2019 PhD Thesis that contribute to these themes. The first one is the one of Claudia Volpetti, [18], which develops hybrid techniques to detect hate speeches through deep learning models, and implements a bias mitigation treatment to reduce the unintended bias in online microblogging platforms, such as Twitter. Moreover, she proposes dynamic representations of words as a suitable deep learning tool to study the evolution of users roles and their sentiments across the plot of a narrative text or an online discourse. The second PhD is the one of Elaheh Raisi, [15], which proposes a general framework that trains an ensemble of two learners in which each learner looks at the problem from a different perspective. One learner identifies bullying incidents by examining the language content in the message, the second learner considers the social structure to discover bullying. They train the models by optimizing an objective function that balances a co-training loss with a weak-supervision loss, and evaluate their experiments with data from Twitter.

Another research done under the process of a PhD thesis is the paper from Mengfan Yao, [19], which identifies the three

266
key practical challenges still unresolved at cyberbullying: detection timeliness, scalability to the staggering rates of online social networks, and reliance on high quality annotations from human experts. She proposes a formulation of the classification problem as sequential hypothesis that seeks to drastically reduce the number of features used while maintaining high classification accuracy with semi-supervised methods.

Regarding text mining and unsupervised learning, there is the work of Nina Janasik et. al., [10], that introduces to the Self-Organized Map (SOM) method used in qualitative data. SOM is a versatile quantitative method commonly used to analyze large dataset, which has a map as an outcome where entities are positioned according to their similarity. Moreover, text mining using SOM can be particularly effective in improving inference quality within qualitative research and quality concepts and categories used in the analysis. Thus, SOM can be considered in the implementation of this project as to maintain the best possible quality in the concepts and categories used.

There is also the research of Rahul R. Iyer et. al., [9], which aims to understand and detect persuasion through speech. Instead of using lexical features for detecting persuasive tactics in text, which do not take advantage of the structures inherent in the tactics used, they formulate the task as a multi-class classification problem and propose an unsupervised, domain-independent machine learning framework. This exploits the inherent sentence structure present in the persuasion tactics, and, in the case of our project, can help to better address language and speech since Violence Against Women in Digital Space may have an specific sentence structure as well.

Their approach has two steps, a simple argument extractor model, to extract arguments from a given piece of text, and the output extractor into the persuasion detection model, which classifies the arguments into different tactic classes. This process is showed in Figure 1.

![Fig. 1. The outline of the problem considered based on the persuasive detection model. Based on the one at Iyer et. al., [9].](image)

Moreover, there is also a research of Amir Karami, [11], which uses unsupervised learning to reduce dimensionality of text data. Their approach uses fuzzy clustering over unsupervised feature transformation (UFT) as to explore a new approach UFT-based to create a lower-dimensional representation of documents. Their research shows performance of fuzzy clustering to exceed principal component analysis (PCA) and singular value decomposition. Hence, this work opens the possibility of using other techniques than just the traditional ones for text data dimensionality reduction in our project, as dimensionality reduction may be key in the clustering formation.

C. Modeling

As part of the modeling, the specific techniques and tools that we are going to use to address Digital Violence Against Women from the mentioned Data Science techniques are the following: (1) Representative-Based Algorithms, (2) Probabilistic Text Clustering Algorithms, (3) Co-Clustering, and (4) Probabilistic Latent Semantic Analysis. All these tools reside in the combination of the four approaches and will be used to understand violent social discourses and structural violence in digital space.

Representative-Based Algorithms belong to the type of algorithms such as k-means, k-modes, and k-median algorithms. However, they need to be adapted with two modifications, [1], to effectively tackle text data. The first modification is to use the cosine similarity function instead of the Euclidean distance. The second alteration is that low-frequency words in the cluster need to be projected out in order to provide a representative set of topical words for the cluster, this is called cluster digest. Moreover, we are going to use the Scatter Gather Clustering approach, which uses a combination of k-means with hierarchical partitioning. This combination prevents the seed choice sensitivity of a k-means approach and the large $\Omega(n^2)$ run-times of the hierarchical partitioning.

Now, according to Charu Aggarwal in his book, [1], Probabilistic Text Clustering Algorithms are considered an unsupervised version of the naive Bayes classification method. Its basic idea is based in two iterative steps: (1) Select one of the clusters, and (2) Generate its term distribution based on the generative model. For our research we will use the Bernoulli model.

The Co-Clustering approach refers to simultaneously discover word clusters and document clusters. This concept begins with the distinction of high-dimensional clustering methods to best characterize in terms of both columns and rows simultaneously. However, in the text domain this approach uses the topical words of a cluster that provide insights of that cluster, giving a natural interpretation of text data [1]. In addition, this approach works more effectively with matrices sparsely populated. Its idea is to rearrange the rows and columns in the data matrix so that most of the nonzero entries become arranged into blocks.

Probabilistic Latent Semantic Analysis (PLSA) is also known as Topic Modeling and provides an alternative method for performing dimensionality reduction and has several advantages over its traditional LSA. This approach differs from others because the underlying generative process is optimized to discover the correlation structure of the words rather than the clustering structure of the documents.

These four specific tools lie between the four techniques defined in this article to attend Digital Violence Against Women, Text Mining, Sentiment Analysis, Pattern Recognition, and Unsupervised Learning. Moreover, the four tools have different
characteristics that complement with each other when they are globally applied to tackle a problem.

IV. RESULTS AND DISCUSSION

A. First Stage Results

We used four different datasets for our first stage results exploration. Three of those datasets belong to the SemEval 2019 Task 5, [2], which is part of a competition on detecting hate speech against immigrants and women in Twitter. The last dataset belongs to the data used by Lara Quijano-Sanchez et. al., [12], which was used to construct HaterNet, a system for detecting and analyzing hate speech in Twitter.

All of the datasets contain tweets in the Spanish language. The tweets that belong to [2] were mainly collected in the time span from July to September 2018, where different approaches were employed to collect them, such as monitoring potential victims of hate accounts, downloading the history of identified haters and filtering Twitter streams with keywords, i.e. words, hashtags and stems. The three datasets consist of 6,600 tweets. On the other hand, the dataset that belongs to [12] contains a corpus of 6,000 tweets collected randomly between February 2017 and December 2017, and labeled as hate containing or not.

For this datasets exploration, then we used the leading Natural Language Processing (NLP) library spaCy. This library relies on specific models that are language-specific and come in different sizes. In this case, we loaded an Spanish model, the es_core_news_sm, which is a Spanish multi-task CNN trained on UD Spanish AnCora and WikiNER. This model assigns context-specific token vectors, POS tags, dependency parses and named entities.

Moreover, for the exploration we realized a text classification task in order to detect hate speech, according to the classification that the datasets already contain.

The results we obtained can be seen in Table I and correspond to an accuracy of the hatespeech detection of 64.0%, 68.8%, and 78.9% for the three Semeval datasets [2], and of 78.3% for the Zenodo dataset [12]. The applied model was the simplest one, which means that there are still some improvements that can be done to have a better classification of texts. However, these results also reflect an initial difficulty to classify hate speech regarding its diverse forms in language.

The Digital Violence Against Women phenomena is a different phenomenon from just plain cyberbullying, having different kind of logic and dynamics. The fact that it is part of the symbolic violence gives the language and the discourse around it a great importance. Therefore, as we try to understand the phenomena, the Text Mining and NLP techniques from Data Science become really relevant in the search for counteracting it.

Moreover, the first stage results shown are themselves an insight for looking at how the sentiment of hat against women can be identified in the text datasets. Thus, the Machine Learning Models can be trained to identify patterns in order to classify these texts as Digital Violence Against Women or not. Likewise, it opens up to consider Unsupervised Learning models as another form of technique that can be used to decipher the phenomena and understand it.

V. CONCLUSIONS

Digital Violence Against Women is a phenomena that has still too much to be explored. The main lines of research focus on detecting violence or hatespeech. However, this project seeks to specifically contribute in understanding the structural violence that is contained in Violence Against Women in Digital Space. Therefore, we will analyze the symbolic violence that is contained in the language of violent social discourses like insults, threats and hate messages. Moreover, we are seeking as well to apply a gender perspective to Data Science and Unsupervised Learning.

The first stage datasets exploration showed that there is indeed some language that can be detected through those tweets that contain digital violence against women. Thus, this gives us the framework to work under the assumption that analyzing language in violent social discourses will show results and insights towards the understanding of symbolic and structural violence. However, the fact that the accuracy detection was under 80%, means that those relations will be difficult to unravel.

VI. ACKNOWLEDGEMENTS

The authors thank Tecnologico de Monterrey, CONACYT-PNPC (student grant number 737330), and CONACYT-SNI (researcher grant number 9804) for the financial support to conduct this study. We also want to thank the Intelligent Systems and Corporalities and Gender groups for their support.

REFERENCES


<table>
<thead>
<tr>
<th>Dataset</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semeval (1/3)</td>
<td>64.0%</td>
</tr>
<tr>
<td>Semeval (2/3)</td>
<td>68.8%</td>
</tr>
<tr>
<td>Semeval (3/3)</td>
<td>78.9%</td>
</tr>
<tr>
<td>Zenodo</td>
<td>78.3%</td>
</tr>
</tbody>
</table>

We present the accuracies obtained by the model in each dataset.