International Journal of Combinatorial Optimization Problems and Informatics, 16(3), May-Aug 2025, 195-215. ISSN: 2007-1558.
https://doi.org/10.61467/2007.1558.2025.v16i3.602

www.editada.org

Uncovering Patterns of Violence in Mexican Digital News Articles Through Data Science Methods

Jonathan Zárate-Cartas[1], Alejandro Molina-Villegas[1]
[1] SECIHTI - Centro de Investigación en Ciencias de Información Geoespacial
jzarate@centrogeo.edu.mx, amolina@centrogeo.edu.mx

Abstract. Violence against women is one of the most common human rights violations, with its most extreme form being femicide. In this context, we considered it relevant to demonstrate how artificial intelligence tools and geospatial analysis techniques can contribute to a better and faster analysis of these crimes. In this study, we analysed femicides that occurred in Mexico between 2014 and 2022. Our data source comprised digital news articles from leading Mexican newspapers. The study begins with the preprocessing of texts and the detection of those mentioning femicide. Subsequently, using unsupervised learning models, we grouped the texts according to their semantic similarity. We then employed deep learning models to classify each crime according to its specific characteristics. Finally, we used spatial analysis tools to detect geographic patterns in the occurrence of these crimes in the metropolitan area of the Valley of Mexico, analysing the automatically detected characteristics as variables.

Keywords: Document classification, Femicides, Geographical analysis, Journalistic notes, Spatial autocorrelation, Text mining.

Article Info
Received January 16, 2025
Accepted March 21, 2025

1 Introduction

In Mexico, the "Ley General de Acceso de las Mujeres a una Vida Libre de Violencia", enacted in February 2007, defines femicide violence as *"the extreme form of gender violence against women resulting from the violation of their human rights, in both public and private spheres; such violations consist of a set of behaviors that involve misogyny, impunity, and social tolerance, and that may culminate in homicide and other forms of violent death of women"* (Cámara de Diputados del H. Congreso de la Unión, 2007).

The tracking of femicide figures was formalized in 2006 with the "Encuesta Nacional sobre la Dinámica de la Relaciones en los Hogares" (ENDIREH) (INEGI, n.d.); this survey allows us to know deeper the types of violence suffered by women over 15 years of age and considers aspects such as the context, domestic violence, penal accusations, the most vulnerable groups, as well as perceptions about gender roles and stereotypes between men and women. In 2011, *Estado de México* was the first state to classify femicide as a crime, and it was until 2017 that all states had classified femicide as a crime (INMUJERES, n.d.).

It is worth mentioning that the *Secretariado Ejecutivo del Sistema Nacional de Seguridad Pública* (SESNSP) periodically publishes information on the incidence of femicides at the national level, but this information only shows statistics of femicides by states (structured data) (Secretariado Ejecutivo del Sistema Nacional de Seguridad Pública, 2019).

The objective of this research goes beyond analyzing the counts of femicides by state in the Republic. As will be seen later, violence phenomenon follows various patterns that can only be studied from unstructured data. In this sense, news articles

represent a valuable source of information that can be exploited through data science techniques. As far as we know, there is no automatic text analysis tool specialized in femicide violence. Therefore, one of the main contributions of this work is the development of a software tool specifically designed to extract information from these texts, using artificial intelligence techniques such as: supervised learning, unsupervised learning, text coding with embeddings and classification neural networks, among others.

In the present work we analyzed almost 10,000 texts from the most important newspapers in Mexico. Different researchers (including the authors) and activists collected such digital news articles during the period from 2014 to 2022.

2 Related works

Our literature review examined numerous works that addressed Natural Language Processing (NLP) and spatial analysis. However, the works presented below were considered the most relevant and current for developing this work.

2.1 Automatic detection of high-risk areas for criminal events through news reports.

The first research considered is from 2020; the authors (Laureano de Jesús et al., 2020) used convolutional (CNN) and short-term memory (LSTM) neural networks to classify news of five types of criminal events (homicide, kidnapping, assault, suicide, and rape) and detect high-risk areas in Mexico. The training data were 8,714 Twitter news headlines, and the test data were 179 news stories from the Milenio newspaper. The experimental results showed that in the classification phase, the accuracy achieved was 98%, and in the analysis phase, the highest risk areas were the states of Sonora and Chihuahua.

It is important to note that the labels used to classify newspaper articles and detect the risk areas do not show details about these events, such as the weapons used, age of the victim, profile of the aggressor, among others.

2.2 Event detection from geotagged tweets considering spatial autocorrelation and heterogeneity

Following this line, in 2021, researchers from K. N. Toosi University of Technology, Tehran, Iran (Ghaemi & Farnaghi, 2023) proposed a method called Varied Density-based Spatial Clustering for Autocorrelated Twitter Data (VDCAT) to *extract the latent spatial pattern in geo-referenced tweets;* it should be noted that VDCAT is an extension of previous methods such as VDCT (Ghaemi & Farnaghi, 2019) and VDBSCAN (Liu, Zhou & Wu, 2007). The relevance of VDCAT is that it considers spatial autocorrelation as well as spatial heterogeneity while revealing the underlying pattern in georeferenced tweets. The method was applied and evaluated during Hurricane Dorian on the east coast of the United States, specifically in the state of Florida, and the results demonstrated its superiority over the Moran's Index and VDBSCAN algorithms in *extracting clusters with various densities*. The authors found a Moran's Global Index of 0.085912 with a p-value less than 0.05, proving that the pattern expressed by tweets related to hurricane Dorian formed clusters according to the clustering quality measures used, such as Davies-Bouldin, Dunn Index, and Silhouette coefficient. *VDCAT showed a better performance by 10% with respect to VDBSCAN.*

2.3 Data Against Feminicide: Data Highlighter.

Subsequently, in 2022, students from the Massachusetts Institute of Technology (MIT) developed a plug-in (Google LLC, n.d.) for the Chrome browser whose function is to highlight in the text of a web page words that are useful for logging, in particular names, places, dates, and words selected by each user. The data highlighter is a tool for activists, journalists, human rights defenders, and others who collect data manually by reviewing news articles (e.g., reading news stories to record cases of femicide and cases of police violence). This tool works for English, Spanish, and Portuguese texts.

It is interesting how the authors manage to highlight the wealth of patterns that the text may contain. Thus, modern NLP methods can be particularly useful for the extraction of characteristics of phenomena such as violence. Likewise, by means of the methods that we will explain later, texts converted into numerical patterns can serve as variables in spatial autocorrelation methods.

2.4 Analysis of journalistic articles on violence against women through data management strategies from information design.

In the case of Mexico, in 2023, researchers from the Universidad de Guanajuato (Mata-Santel, Luna-Gijón & Ronquillo-Bolaños, 2023) conducted, through a mixed approach, an analysis of journalistic notes reported by three digital newspapers on violence against women in Puebla, Mexico, during the month of January 2022. Their conclusions showed that the referred media do not present sufficient data to identify the profile of the aggressor and that there is a tendency to objectify those who have suffered violence, the consequence of which is that the reader is unable to dimension the situation.

Their contribution lay in the valuation of the design of information as a tool that, through its management processes, reveals patterns that would otherwise remain unnoticed. The authors concluded that newspapers perpetuate patterns that conceal violence and suggested *the need for a proposed digital application that presents information through data visualization in an interface that makes it accessible to the public*.

3 Methodology and Models

One of the biggest challenges facing experts in violence in Mexico is the exhaustive manual review of news articles to extract patterns and analyze them. This raises the need to generate technologies capable of carrying out some of these processes in an automated way. As part of our contribution to this problem, we have developed a tool that automatically processes the news articles to extract patterns and perform analyses and visualizations that help experts carry out their work effectively. Our proposed global methodology to develop such a tool consists of four modules, which can be seen in a general way in Figure 1.

The first module preprocesses, vectorizes texts and, through a classifier created by us, detects and discards texts that do not refer to femicides. The second module uses some of the most common unsupervised learning models to group texts according to their semantic similarities. Subsequently, the third module classifies, through classifiers created by us, the texts according to labels or characteristics that experts on the subject consider necessary to study this phenomenon in greater depth. The fourth module considers the characteristics of interest automatically detected in the third module and detects spatial patterns of occurrence of certain types of femicides in an area of interest. In this case, we choose the metropolitan area of the valley of Mexico as a study case.

Finally, the tool reports with the summary of each module. The tool can give us the probability of a note being a femicide, it can also group similar notes, it can label relevant aspects mentioned in the texts, such as the age of the victim, the type of weapon and other aspects of interest and it can generate visualizations such as autocorrelation maps. In the rest of the section each of the modules will be described in detail.

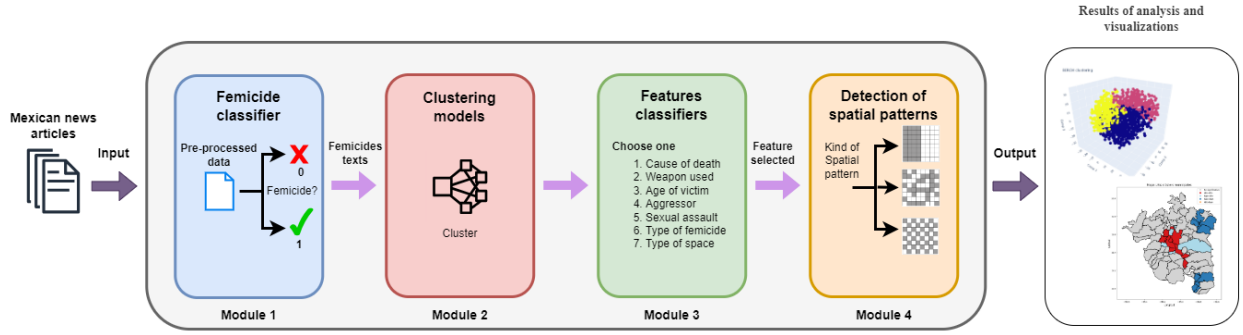

Figure 1.- Overview of the general methodology: The first module preprocesses, vectorizes, and filters out texts unrelated to femicide using a classifier. The second applies unsupervised learning methods to group texts by semantic similarity. Then, texts are categorized based on expert-defined labels for deeper analysis. Finally, key detected features undergo spatial autocorrelation analysis.

3.1 Data and Preprocessing

The corpus used for this work consists of approximately 10,000 digital news articles downloaded from the websites of major newspapers across the country, for a period spanning from 2014 to 2022. It is worth mentioning that we include dozens of local

newspapers because they contain details that can help to extract specific patterns. We divided the dataset into three parts as shown in Figure 2.

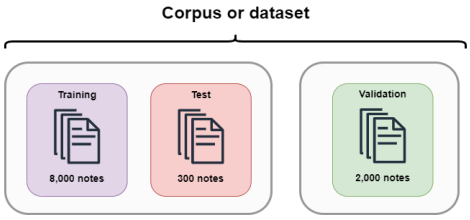

Figure 2.- Dataset used in this research.

The validation set corresponds to violent events located only in the metropolitan area of the valley of Mexico during the period 2016-2020.

The preprocessing of digital news articles included the following steps: 1) Decodification of special characters, 2) Standardization of word formatting, 3) Conversion of text to lowercase, 4) Deletion of punctuation marks, 5) Stopwords suppression, and 6) Tokenization.

3.2 Text vectorization

After tokenization the texts were vectorized using *embeddings;* these are dense vector (numerical) representations of data in which the linear distances of such vectors are related to the structure of the original data (Arize AI, n.d.). We created such vectors using a *Large Language Model (LLM)* called BETO (Cañete et al., 2020). BETO is a BERT model trained on a big Spanish corpus; this model converts each text document into dense vectors with 768 positional entries. Figure 3 shows the vectorization process.

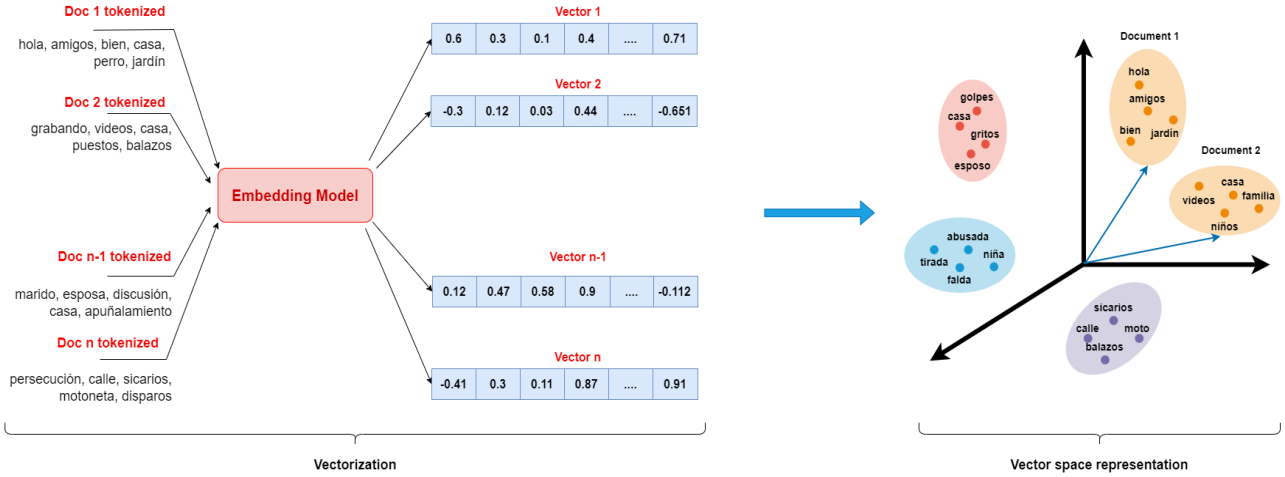

Figure 3.- Texts vectorization and representation of document embeddings.

3.3 Module 1: Femicide classifier

Module 1 is designed to automatically distinguish news articles that refer to a femicide. To do this, the texts are encoded into dense numerical vectors and the result is determined by a binary neural network. This configuration is known as binary classification. The neural network was trained using vectorized texts (embeddings) as input and a label as output as illustrated in Figure 4. The classifier will label with a '1' if the text corresponds to the description of a femicide and with a '0' otherwise. *We used texts only labeled with a '1'.*

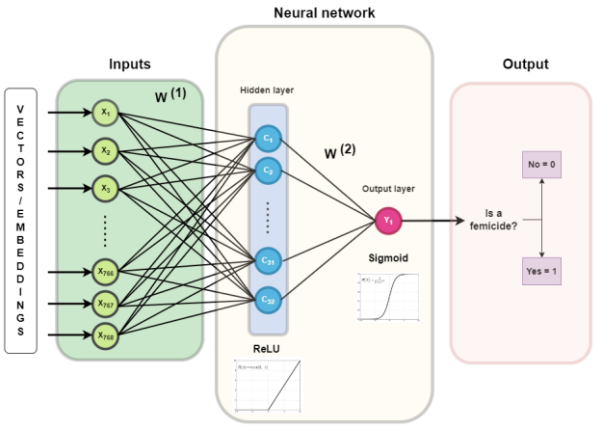

Figure 4.- Femicide classifier.

The classifier is a Recurrent Neural Network (RNN) with 24,641 trainable parameters. It is composed of 768 inputs, 32 neurons in one hidden layer with ReLu activation function, and 1 neuron output layer with a sigmoid activation function. The output is binary, calculating the probability that the label corresponds to the value 'true' (1). The model needed 20 epochs of training.

3.4 Module 2: Clustering models

Module 2 is designed for grouping texts according to their semantic similarities or characteristics that they share. This feature is useful when texts present certain types of crimes for which there is little information. For example, transphobic homicides are examples of this. Thus, the clustering module can detect when there are new events that emerge in the press but that look like previous events. Clustering is a process divided into three steps: 1) dimensionality reduction, 2) clustering algorithms, and 3) metrics.

3.4.1 Dimensionality reduction algorithms

Before performing any clustering process, it is necessary to reduce the dimension of each vector. By reducing the dimensionality of such vectors, it will enable clustering algorithms to have better metrics, reduce storage space, and allow us to visualize data in 2D and 3D. The following is a brief description of the dimensionality reduction methods available in this module.

- *Principal Component Analysis (PCA)*.- The objective of this method is to obtain a new set of variables, without linear correlation between them, called principal components (PCA, n.d.). This new set will have the same dimension n as the original data, except that only the first m ($m < n$) components may explain a large part of the problem with little loss of information.

- *t-Distributed Stochastic Neighbor Embedding (t-SNE)*.- Is a nonlinear technique developed by Georey Hinton and Laurens van der Maaten in 2008 (van der Maaten & Hinton, 2008) which in turn is an improvement of a previous technique known as *Distributed Stochastic Neighbor Embedding* (*SNE*) presented by Hinton and Roweis in 2002 (Hinton & Roweis, 2002), t-SNE improves SNE minimizing the divergence of the *Kullback-Leibler* cost function.

The first step of the algorithm consists of converting the quadratic Euclidean distance $d^2(i, j)$ existing between two points x_i and x_j that belong to a high-dimensional space, in a conditional probability $p_{j|i}$ that represents this similarity of separation, this means that for close points, $p_{j|i}$ is higher, otherwise $p_{j|i}$ closer to zero (for reasonable values of the Gaussian variance σ_j). The second step also consists of converting distances into probabilities, unlike step 1, in this step the distances between the low-dimensional points y_i and y_j are converted to a conditional probability $q_{j|i}$ using a *t-Student* probability distribution with a degree of freedom that has heavier tails than the Gaussian distribution. This is to eliminate the agglomeration problems generated due to the Gaussian kernel used in the Euclidean distance and p_{ij} converge asymptotically to a constant.

- *Uniform Manifold Approximation and Projection (UMAP)*.- This method is a variant of t-SNE, although it has marked differences that make it a more computationally optimal algorithm than t-SNE. The theoretical foundations of UMAP are based on *Variety Theory* and *Topological Data Analysis*. In a high-dimensional space, UMAP uses local variety approximations and

joins its representations of local fuzzy simplicial fuzzy sets to construct a high-dimensional topological representation of the data. Given a low-dimensional representation of the data, a similar process can serve to construct an equivalent topological representation. UMAP then optimizes the layout of the data representation in the low-dimensional space to minimize cross-entropy between the two topological representations (McInnes & Healy, 2018).

3.4.2 Clustering algorithms

- *K-means*.- This algorithm solves an optimization problem, where the optimization function is the sum of the square of the distances of the object (sample) to the centroid of its cluster. By real vectors of d dimensions ($x_1, x_2, \ldots, x_n$), that represent such objects, *k-means* algorithm builds k groups $S = (S_1, S_2, \ldots, S_k)$ where the distance of the objects within each group to its centroid is the shortest possible.

- *Hierarchical*.- This algorithm groups the objects ($x_1, x_2, \ldots, x_n$) hierarchizing them according to a specific order or criterion. This hierarchization works in two ways:

a) *Agglomerative*.- This method starts by considering a point x_i as an independent cluster and then iteratively combines it with another closer cluster until reaching a stopping criterion.

b) *Divisive*.- This is an inverse approach to agglomerative hierarchization. Initially considering all points as a single cluster and recursively dividing them until reaching a stop criterion (a predetermined number of k clusters or a single point becomes a cluster).

- *Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH)*.- This algorithm was proposed by Zhang et al. (Zhang, Ramakrishnan & Livny, 1996), and according to the authors, it is especially suitable for clustering large volumes of data to generate a small and compact summary that retains as much information as possible with an efficient use of memory.

- *Spectral*.- This algorithm was popularized in 2000 by Jianbo Shi and Jitendra Malik (Shi & Malik, 2000). It is based on the idea of constructing partitions of a graph from the eigenvectors of the adjacency matrix, although W. E. Donath and A. J. Homan proposed this idea in 1973 (Donath & Hoffman, 1973).

3.4.3 Clustering metrics

It is especially important to evaluate the effectiveness of the unsupervised learning algorithms used to determine how well they have performed. The clustering metrics consider two factors to evaluate the quality of the clusters formed: *separation* (inter-cluster) and *cohesiveness* (intra-cluster). Such metrics will help us to choose the best dimensionality reduction and clustering methods. We used the following quality metrics in this work:

- *Silhouette score*.- This metric evaluates the quality of a clustering method by considering the cohesion of elements within a cluster and their separation from all other clusters (Rousseeuw, 1987). The value of the measure varies between -1 and +1, where a value close to +1 indicates that the element belongs to the cluster assigned, while a value close to -1 indicates that the element could belong to another cluster. This can be seen in Equation (1).

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{1}$$

where *a(i)* is the mean distance of the object *i* to all other objects within the cluster (intra-cluster), and *b(i)* is the mean distance of the object *i* to all objects that belong to the nearest cluster (inter-cluster).

- *Davies-Bouldin index (DB)*.- In 1979, David L. Davies and Donald W. Bouldin introduced this index (Davies & Bouldin, 1979) to quantify the degree of cohesion within clusters and the distance between them. It is based on the relationship between the dispersion within clusters and the distance between cluster centroids. *A lower value indicates better separation and cohesiveness between clusters*. This can be seen in Equation (2).

$$DB = \frac{1}{n}\sum_{i=1}^{n} \max\left(\frac{\sigma_i + \sigma_j}{d(\sigma_i, \sigma_j)}\right) \tag{2}$$

where c_i is the center of the i^{th} cluster, σ_i is the mean distance between the objects within the cluster i to the center of that cluster, $d(c_i, c_j)$ represents the distance between cluster c_i and c_j and n is the number of clusters.

- *Calinski-Harabasz index (CH)*.- In 1974, Tadeusz Calinski and Jerzy Harabasz introduced this index (Caliński & Harabasz, 1974). Its function is to maximize the ratio of intra-cluster variance and inter-cluster variance; this means that *a high value denotes a better separation between clusters, while a low value denotes close clusters*. This can be seen in Equation (3).

$$CH = \frac{\sum_{i=1}^{k} n_i \, \|c_i - c\|^2/(k-1)}{\sum_{i=1}^{k} \sum_{x \in C_i} \|x - c_i\|^2/(n-k)} \tag{3}$$

where n is the number of objects, k represents the number of clusters, n_i is the number of objects in the cluster c_i and c represents the centroid of all objects. $\sum^{k}_{i=1} n_i \, ||c_i - c||^2$ is the weighted sum of the squared Euclidean distances between the centroid of each group (mean) and the general data centroid (mean), and $\sum^{k}_{i=1} \sum_{x \epsilon Ci} ||x - c_i||^2$ is the sum of the squared Euclidean distances between the objects and their respective centroids within the group.

3.5 Module 3: Features classifiers

Module 3 is designed to extract from the texts certain characteristics of interest related to the femicides described in them. In collaboration with experts on femicide violence, they have provided insights into the specific details they typically examine in news texts. These characteristics are the following: (1) *Cause of death*, (2) *type of weapon used*, (3) *age of the victim*, (4) *aggressor's identity*, (5) *sexual assault*, (6) *type of femicide*, and (7) *type of space where the aggression occurred*. It is important to mention that manually doing this task requires many hours and many people are dedicated to reading thousands of texts to find such features, so it is necessary to have a module that integrates this function.

In this module each feature is detected and extracted individually by a neural network, so the module is composed of a total of 7 neural networks. It is important to note that each of these characteristics has their own set of labels, and *the output layer of each network's neurons represents the likelihood of that label appearing in the text*. Below, we present a description of each of these neural networks along with the labels that match each characteristic of interest.

In general, all feature classifiers from (1) to (7) are models trained using supervised learning. Specifically, recurrent neural networks were used in all cases, with the only difference being that the output layer can be binary (5) and (7) or multiclass (1, 2, 3, 4, 6). Figure 5 generically illustrates the feature classifiers. The input layer receives a vector representation of the texts; then, the intermediate layers transmit numerical messages through the neurons; finally, the output layer determines the probability that the corresponding text contains the label of interest, $\text{Prob\_label}_{(n)}$.

3.5.1 Label "cause of death"

This characteristic explains how the victim died and the type of aggression that she suffered. The labels used were the following: undefined cause (0), gunshot wounds (1), head trauma or blows to the body (2), stabbing or lacerations (3), calcination or burns (4), mutilation or decapitation (5), torture (6), mechanical asphyxia (7), run over by a motor vehicle (8), and poisoning or overdose (9).

The classifier is a Recurrent Neural Network (RNN) with 12,474 parameters. It is composed of 768 inputs, 16 neurons in one hidden layer with a ReLU activation function, and a 10-neuron output layer with a softmax activation function. This means that the output is multi-class. The model needed 100 epochs of training.

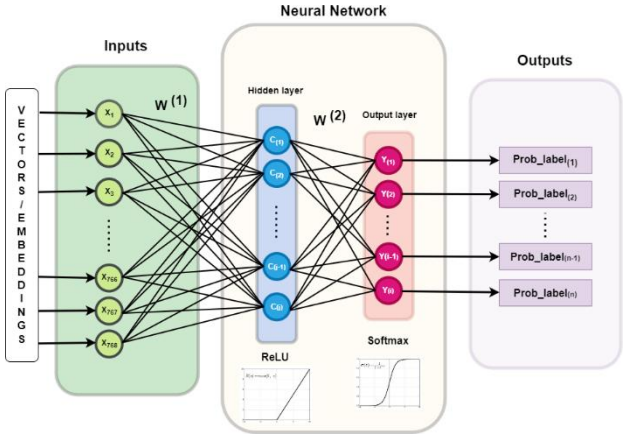

Figure 5.- Generic illustration of the feature classifiers for the models (1) *Cause of death*, (2) *type of weapon used*, (3) *age of the victim*, (4) *aggressor's identity*, (5) *sexual assault*, (6) *type of femicide*, and (7) *type of space where the aggression occurred*.

3.5.2 Label "weapon used"

This label identifies the kind of weapon used in the crime; it is significant because it provides information about the victim's attack environment and could serve to profile the aggressor. The labels used were the following: unidentified (0), firearm (1), knife (2), blunt object (3), blows with hands (4), fire or chemical agent (5), constrictor element (6), airway obstructing element (7), automobile (8) and drugs (9).

The classifier is a Recurrent Neural Network (RNN) with 7,800 parameters. It is composed of 768 inputs, 10 neurons in one hidden layer with ReLu activation function and 10 neurons output layer with softmax activation function, this means that the output is multi-class. The model needed 100 epochs of training.

3.5.3 Label "victim's age"

This feature is important since it would allow the stage of life at which a woman is most susceptible to suffering an aggression. *We considered the WHO standards to determine each period of life*. The labels used were the following: Adult (0), Young Adult (1), Old Age (2), Adolescent (3), Early Childhood (4), and Infancy (5). It is important to note that the life cycle is divided into 6 stages according to WHO; these are early childhood (0-5 years), infancy (6-11 years), adolescence (12-18 years), youth (14-26 years), adulthood (27-59 years), and old age (60 years and over).

The classifier is a Recurrent Neural Network (RNN) with 7,756 parameters. It is composed of 768 inputs, 10 neurons in one hidden layer with ReLu activation function and 6 neurons output layer with softmax activation function, this means that the output is multi-class. This model needed 100 epochs of training.

3.5.4 Label "aggressor (identity)"

This characteristic is important because it lets us know the identity of the aggressor, and especially if the victim and the aggressor had a relationship, this would allow for a better classification of the type of femicide committed. The labels used to describe the aggressor were the following: Unknown (0), Partner or ex-partner (1), Acquaintance (2), Parent or guardian (3), and Relative (4).

The classifier is a Recurrent Neural Network (RNN) with 7,745 parameters. It is composed of 768 inputs, 10 neurons in one hidden layer with ReLu activation function and 5 neurons output layer with softmax activation function, this means that the output is multi-class. This model needed 100 epochs of training.

3.5.5 Label "sexual assault"

This label verifies whether the victim suffered any sexual assault or humiliation; such acts would suggest misogyny or hostility against her. The labels are binary: there was not sexual assault (0) and there was sexual assault (1). Figure 6 shows this classification model.

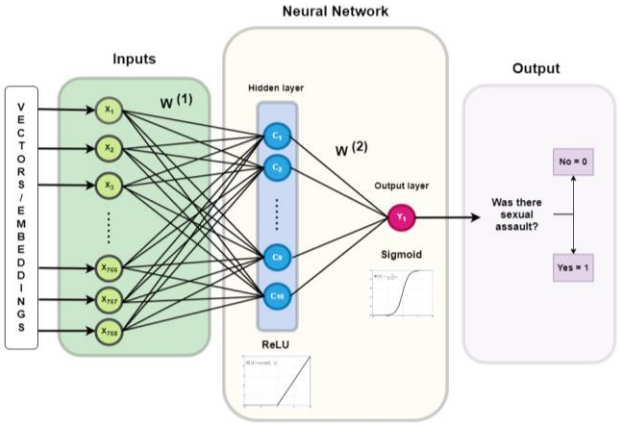

Figure 6.- "Sexual assault" classifier.

The classifier is a Recurrent Neural Network (RNN) with 7,701 parameters. It is composed of 768 inputs, 10 neurons in one hidden layer with ReLu activation function and 1 neuron output layer with sigmoid activation function that represents the probability that the text analyzed mentions a sexual assault. This model needed 20 epochs of training.

3.5.6 Label "type of femicide"

This label refers to the distinct types of femicides according to the classifications mentioned in the specialized literature (Monárrez, 2010). In addition to these labels, we decided to add one, label (0), which refers to crimes that occurred in the context of the insecurity experienced in the country. The labels were the following: violent death (0), intimate (1), non-intimate (2), child (3), family (4), connection (5), disorganized systemic sexual (6), organized systemic sexual (7), prostitution or stigmatized occupations (8), human trafficking (9), trafficking (10), transphobic (11) and lesbophobic (12).

The classifier is a Recurrent Neural Network (RNN) with 25,037 parameters. It is composed of 768 inputs, 32 neurons in one hidden layer with ReLu activation function and 13 neurons in its output layer with softmax activation function, this means that the output is multi-class. This model needed 100 epochs of training.

3.5.7 Label "type of space"

This label identifies whether the victim was attacked or found in public or private space. In public space (street or a place the victim did not know of) attacks are usually committed by someone unknown to the victim and in private space (usually at home) attacks are usually committed by someone known to the victim. The labels used were the following: public (0) and private (1). Figure 7 shows this classification model.

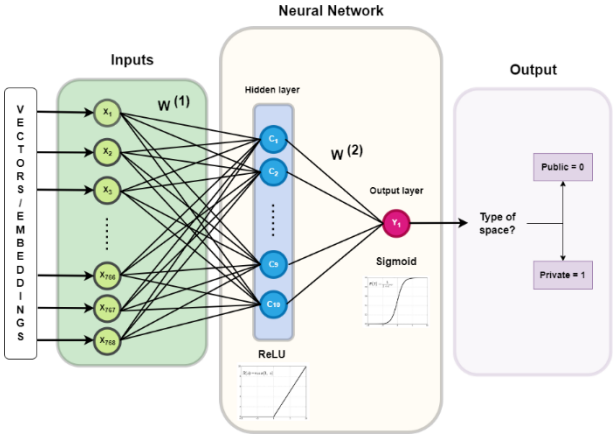

Figure 7.- "Type of space" classifier.

The classifier is a Recurrent Neural Network (RNN) with 7,701 parameters. It is composed of 768 inputs, 10 neurons in one hidden layer with ReLu activation function and 1 neuron in its output layer with sigmoid activation function that represents the probability that the text analyzed mentions a crime occurred in private space. This model needed 20 epochs of training.

3.5.8 Classification metrics

To measure the quality of the classifiers, we use the following metrics to know their performances, but first we will define the 4 possible outcomes for a binary classifier; this is known as the confusion matrix:

Table 1. Confusion matrix

	Actual positive	*Actual negative*
Predicted positive	*True Positive (TP).-* Is the number of correctly classified positive samples.	*False Positive (FP).-* Is the number of samples incorrectly classified as positive.
Predicted negative	*False Negative (FN).-* Is the number of samples incorrectly classified as negative.	*True Negative (TN).-* Is the number of correctly classified negative samples.

- *Accuracy*.- Used to measure model performance. It is the ratio of the number of correct predictions to the total number of predictions. This can be seen in Equation (4).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

- *Precision*.- This is a measure of the accuracy of the positive predictions of a model. It is defined as the ratio of correct positive predictions to the total number of correct and incorrect positive predictions made by the model. This can be seen in Equation (5).

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

- *Recall*.- Used to measure the effectiveness of a classification model identifying all relevant instances of a data set. It is defined as the ratio of correct positive predictions to the total number (positive and negative) of correct predictions made by the model. This can be seen in Equation (6).

$$Recall = \frac{TP}{TP + TN} \tag{6}$$

- *F1 Score*.- Used to evaluate the overall performance of a classification model. It is the harmonic average of accuracy and recall. This can be seen in Equation (7).

$$F1\ Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{7}$$

3.6 Module 4: Detection of spatial Patterns

Module 4 performs the search and detection of spatial patterns of the analyzed texts by means of spatial autocorrelation; the variables under analysis are the characteristics of interest automatically detected and labeled in module 3. These characteristics are known as *spatial variables*. In this work the geographical area chosen to analyze was the metropolitan area of the valley of Mexico and the femicides occurred in the period from 2016 to 2020.

3.6.1 Spatial units of study

Spatial variables are rarely studied in a punctual way; they are commonly studied for a certain area or spatial units (Olaya, 2020). Defining spatial units is an *arbitrary task*, i.e., it does not follow scientific criteria; an example of this could be the political divisions of countries, regions, states, cities, municipalities, etc. Consequently, we decided to evaluate distinct types of spatial units, previously defined and studied, to see how this could alter the study of spatial patterns of these crimes and even address them at the public policy level. Spatial units used in this work were three:

- *Municipality*. - This is a political division considered a second level. It is a political-administrative cell that, in the case of Mexico, is the basis for the organization of the federative entities.

- *Geohash*. - It is a system, created in 2008 by Gustavo Niemeyer, which encodes locations (pairs of geographic coordinates) in a string of alphanumeric characters in base 32^6, each string representing a rectangle that covers a certain area. The length of this string determines the accuracy of the Geohash: the longer the string, the higher the precision. Geohash has 12 levels of precision.

- *H3 (Hexagonal)*. - It is a geospatial indexing system developed by *Uber* that divides the world into a discrete global grid consisting of hierarchically organized multi-precision hexagonal tiles. The H3 mesh starts from an icosahedron that recursively creates meshes of hexagons of increasing precision until reaching the desired resolution. H3 has 15 levels of resolution.
It is important to mention that Geohash and H3 have distinct levels of accuracy or resolution, this means that *the area of each spatial unit generated can vary*. We decided to prove 2 levels of resolution for our research.

3.6.2 Global Spatial Autocorrelation

Andrew Cli and Keith Ord coined this term (Cliff & Ord, 1969) during the annual conference of the *Regional Science Association*. It is the procedure by which we can measure the distribution and variability of a phenomenon across geographic space (the existence of a correlation of the phenomenon with itself). Tobler's First Geographic Law (Tobler, 1970) supports the concept of spatial autocorrelation: *everything is related to everything, but things close to each other are more related than distant things*.

One of the most widely used statistics for analyzing variations of spatial autocorrelation with nearest neighbor values is the *Moran's Index*. Developed by Alfred Pierce Moran in the late 1940s and early 1950s, this index is based on a measure of covariance, where the similarity measure between the values of x (variable under analysis) in spatial unit i is defined by this interaction with its neighbors in spatial units j. This can be seen in Equation (8).

$$I = \frac{N}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}} \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(x_i - \overline{x})(x_j - \overline{x})}{\sum_{i=1}^{N}(x_i - \overline{x})^2}, \forall i \neq j \tag{8}$$

The values of I range from +1 to -1, where +1 indicates that the values are clustered (have similar values close to each other), -1 indicates that the values are isolated (have different values close to each other), and 0 indicates the absence of spatial patterns. It

is worth mentioning that *I* fit the *z-statistical significance test*, assuming a normal distribution. This is a decision rule that validates the research hypothesis, where:

H_0: No spatial autocorrelation.
H_1: There is spatial autocorrelation.

3.6.3 Local Indicators of Spatial Association (LISA)

This Index decomposes the Moran's Global Index to evaluate the contribution of each spatial unit (I_i) to the global value; it represents those locations with significant values in statistical indicators of local spatial association (Anselin, 1995). This can be seen in Equation (9).

$$I_i = (x_i - \overline{x}) \sum_{j=1}^{N} w_{ij}(x_j - \overline{x}), \forall i \neq j \tag{9}$$

The graphical tool that allows us to observe the dispersion and behavior of each spatial unit was launched in 1993 by Luc Anselin and is known as the Moran Scatterplot. This graph is divided into four quadrants numbered counterclockwise. Quadrant (I) shows the units with values greater than the mean, and which have neighbors with high values, a situation commonly called *High-High* or "hot spots." In quadrant (III), the opposite occurs; here the situation is *Low-Low* or "cold spots." In quadrant (II), it is *High-Low*, with values of spatial units higher than the mean (high) with neighbors having low values (lower than the mean). In quadrant (IV), the situation is *Low-High*.

4 Experiment design

In this section we described the operation of each step shown in Figure 1 and how we used the corpus shown in Figure 2 in our experiments. The data were downloaded from the main Mexican newspapers as briefly described in Table 2. We created a database with the following structure: *Date* (Month-Day-Year when the events occurred), *Title* (Main headline of the newspaper article), *Note* (Description of events), *Link* (Download web page) and *Coordinates* (Location coordinates of the events).

Table 2. Sources of information and a brief description of corpus dataset

Source	*Description*
Debate	800 articles from "mexico" section
Milenio	1000 articles from "estados" section
La Prensa	800 articles from "policia" section
Afondoedomex	800 articles from "feminicidio" section
Reforma	900 articles from "articulo" section
Excélsior	1000 articles from "comunidad" section
El Universal	1000 articles from "articulo" section
Proceso	900 articles from "nacional" section
El Sol de México	1100 articles from "local" section
Quadratín	1000 articles from "regiones" section
El Gráfico	1000 articles from "la roja" section

The experiments carried out consisted of using a large portion of the data, described in Figure 2, to *train* machine learning (ML) models. Each experiment consisted of performing a training batch with certain parameters depending on the type of ML model being trained. The comparison between the resulting models was performed by validating the results with a portion of the data called *test* as described in Section 3.1. Each model classified the test data according to its results (confusion matrix) measured with the following quality metrics: Accuracy (4), Precision (5), Recall (6) and F1 Score (7). In the ML area, the results of the quality metrics are meaningful from a certain number of available data. Typically, in the order of thousands, as in our case.

These classification models were used to label a dataset called *validation*, such models allowed to detect and select texts referring to femicides and features of interest mentioned in such texts. These characteristics of interest were used for the hypothesis test corresponding to the Moran's I. This test verifies whether the femicides analyzed follow some kind of spatial pattern according to their characteristics.

For this research we used Python language version 3.9 and libraries as scikit-learn version 1.6.1 to use clustering algorithms and Tensorflow version 2.18.0 stand out to create and test several classification models. The computer used was a msi thin B12UC equipped with a 12th Gen Intel Core i5-12450H of 2.0 Ghz processor, GeForce RTX 3050 Laptop GPU GDDR6 128-bit and 16.0 GB RAM.

In this part we describe the steps followed with the validation data to achieve the results shown in section 5.

Module 1.- Before selecting a neural network as classifier model, we evaluated other algorithms dedicated to this task. Among the best known as described in Table 3. For each technique chosen, we performed 50 training courses to obtain the best model of each type and compare them, as we show in table 4, the average processing time was in the order of 2.0 minutes.

Table 3. Models of classification and parameters used during tests

Algorithm	*Parameters*
Logistic regression	max_iter: 100, penalty: l2
Multinomial Naïve-Bayes	alpha: 1.0
Support Vector Machine (SVM)	Kernel: Linear, penalty: l2, max_iter: 100
Random Forest	n_estimators: 10, max_features: 1.0

Module 2.- In module 2 we used clustering metrics mentioned in the equations (1), (2) and (3) to evaluate all possible combinations among dimensionality reduction methods, clustering algorithms and testing them for a number proposed of 10 clusters. After all tests, we obtained the best metrics combining the UMAP method, k-means algorithm for 3 clusters. This means that the femicides that occurred in the metropolitan area of the valley of Mexico in the period 2016-2020 can be grouped into 3 types according to their semantic similarities.

Module 3.- In module 3 we considered only 4 of 7 characteristics or labels for the analysis. The first of which refers to the number of femicides (provided by module 1) and the other 3 characteristics were detected by their corresponding classifiers of this module. These variables or labels were the following: number of femicides, femicides by violent death (cause of death), intimate femicide (type of femicide) and sexual aggression.

Module 4.- Finally, each label obtained in modules 1 and 3 was used as a spatial variable or input vector to calculate the global and local spatial autocorrelation for 3 different spatial units (Municipal, Geohash and H3). The objective is to find spatial patterns of femicides occurred in the metropolitan area of the valley of Mexico with the characteristics or labels selected between 2016 and 2020.

5 Results

This section shows, by means of tables and graphs, the results obtained for each module with the *validation dataset*.

5.1 Results module 1

Table 4 shows the classification metrics obtained by different algorithms of classification of femicide texts. In this case the classifier labeled only 1,027 texts as femicides, we used only these texts.

Table 4. Metrics for different classification methods for femicide texts detection

Classifier model	*Accuracy*	*Precision*	*Recall*	*F1 Score*
Logistic regression	0.97	1.00	0.94	0.97
Multinomial Naïve-Bayes	0.91	0.86	0.98	0.92
Support Vector Machine	0.96	0.98	0.94	0.96
Random Forest	0.98	0.98	0.98	0.98
Neural Network	0.97	0.94	0.97	0.97

It is important to mention that we choose the neural network as our classifier, despite other algorithms showing slightly better metrics, because *neural networks support dense vectors or embeddings as inputs*.

An example of a text classified by the best algorithm is shown in Table 5. This table presents an excerpt from the corpus notes, for which the best classifier determined with a probability of 0.97 that the text in note (a) describes patterns related to femicide reports. As a counterexample, the text in note (b) was evaluated with a probability of 0.02 and therefore is not considered a femicide report.

Table 5. Examples of texts classified in module 1

(a) Info raxen. 26/01/2018 *"Ejecutan a transexual en la Gustavo A. Madero".* Prob(femicide) = 0.97	**(b) El Gráfico. 29/05/2018 *"Acribillan a propietario de cocina económica en Chalco".* Prob(femicide) = 0.02**
Han regresado las agresiones a nuestra comunidad transexual y la impunidad de los casos que siempre ha predominado: Activista *24 HORAS.- Una mujer transexual fue asesinada con arma de fuego luego de abordar un taxi en la delegación Gustavo A. Madero. De acuerdo a la carpeta de investigación correspondiente, Paola Carrasco se dirigía a realizar su oficio de sexoservidora a bordo del transporte, el cual circulaba por Calzada de Guadalupe a la altura de Juventino Rosas, en la colonia Peralvillo.* *Fue entonces cuando un vehículo color blanco interceptó al taxi. En el otro automóvil viajaban dos sujetos, quienes dispararon en al menos cinco ocasiones en contra de Paola y una contra el conductor.* *Al lugar llegaron unidades médicas, las cuales determinaron que la mujer transexual ya no presentaba signos vitales. En tanto, el chofer del taxi fue trasladado al hospital Rubén Leñero, donde continúa con su rehabilitación. Tras no haber indicios de que un presunto asalto fuera el móvil del homicidio, la Procuraduría General de Justicia de la Ciudad de México, investiga un posible ataque directo como causa probable.*	*Dentro de la cocina económica de la que era propietario en el municipio de Chalco, Juan Javier Torres, de 50 años, fue asesinado a balazos.* *El homicidio probablemente ocurrió durante la noche del martes, pero fue hasta ayer por la mañana cuando su cadáver fue descubierto.* *De acuerdo con los reportes preliminares, el padre de Juan Javier estuvo con él en la fonda 'La Tradicional', la tarde del martes.* *Alrededor de las 18:00 horas -contó- él se marchó a su casa dejando a su hijo en el establecimiento que está ubicado en la avenida Vicente Guerrero en el Barrio San Francisco.* *Por la noche se le hizo extraño que no llegara a su casa a dormir. Preocupado, el hombre acudió temprano a la cocina económica que encontró cerrada. Al abrir descubrió el cuerpo de su hijo tirado en el piso. De inmediato dio aviso a la policía al darse cuenta que sangraba y no respondía.* *Paramédicos confirmaron su muerte a causa de dos disparos en el tórax. El padre de Juan Javier dijo no tener sospechas de quién pudo haber asesinado a su hijo.*

5.2 Results module 2

Table 6 shows all the metrics obtained with the number of clusters, dimensionality reduction method and cluster algorithm.

Table 6. Quality metrics for UMAP + k-means algorithms

Number of clusters	*Silhouette score*	*Davies-Bouldin score*	*Calinski-Harabasz score*
2	0.454	0.847	1076.503
3	0.479	0.814	1282.317
4	0.441	0.859	1179.503
5	0.381	0.951	1157.383
6	0.332	1.088	1033.751
7	0.332	1.070	1007.715
8	0.324	1.082	1004.949
9	0.332	0.999	997.232
10	0.337	0.965	981.011

Figure 8 shows the separation and cohesiveness of the number of clusters chosen according to the results of the quality metrics.

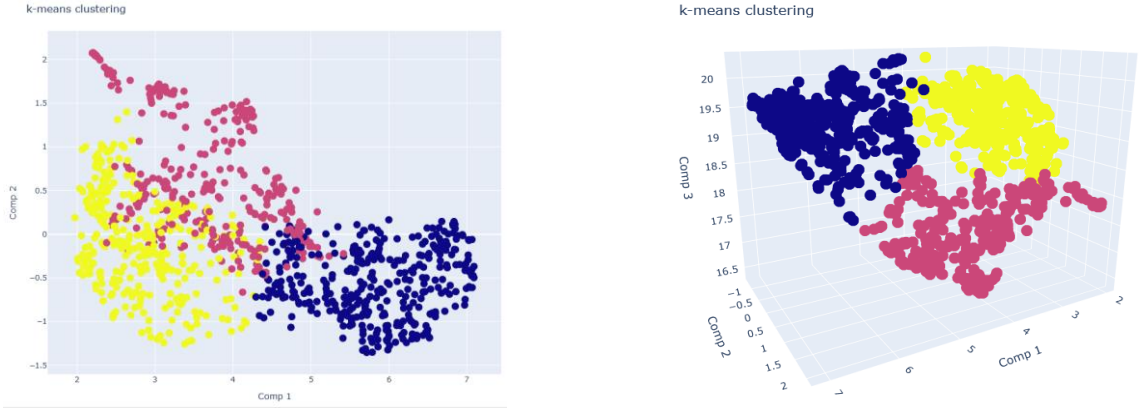

Figure 8.- Clustering visualization 2D and 3D.

The following is an interpretation of each cluster:

(Cluster 1) *Violence related to organized crime:* Here we highlight features such as executions in public streets, use of long weapons, untimely events, hired killings, i.e., this violence is closely related to the environment of insecurity in Mexico.
(Cluster 2) *Intimate violence:* This cluster is characterized by mentioning events in which aggression occurred in the private sphere and was exercised in the context of a romantic-affective relationship. The aggressions were death by asphyxiation, blows or punches and injuries with sharp weapons.
(Cluster 3) *Violence with sexual aggression:* These aggressions were committed in public, or the victim's body was found in the street, and the victims suffered humiliation, sexual violence or rape committed by a stranger. Table 7 shows the kind of notes of each cluster.

Table 7. Notes representing each cluster

Example note of cluster 1	**Example note of cluster 2**	**Example note of cluster 3**
FEMINICIDIO #179: Balacera en bautizo deja una mujer muerta y 10 heridos en Ecatepec *POR: Beda Peñaloza* *ECATEPEC, Méx.- Una mujer muerta y 10 personas lesionadas, fue el resultado de una balacera que se desató durante una riña en una fiesta.*	*Coyoacán: Asfixia a su esposa tras discutir y huye* *22/03/2016 | 10:22 | Arturo Ortiz Mayén* *A Marily José Zúñiga presuntamente la asesinó su pareja sentimental en la vivienda que compartían en la colonia Santa Úrsula Coapa, en Coyoacán.*	*Reportan feminicidio en el Centro Histórico de la CDMX* *Por Letra roja | 18 de marzo 2016* *A las 6:30 horas, policías reportaron el hallazgo del cuerpo de una mujer no identificada, de entre 40 y 45 años, que presentaba golpes y estaba semidesnudo.*

5.3 Results module 3

Table 8 shows the classification metrics obtained for the 7 characteristics sought in texts that described femicides that occurred in the metropolitan area of the valley of Mexico between 2016 and 2020.

Table 8. Metrics of each classifier

Classifier	*Accuracy*	*Precision*	*Recall*	*F1 Score*
Cause of death	0.76	0.83	0.68	0.75
Weapon used	0.71	0.81	0.63	0.71
Victim's age	0.76	0.78	0.72	0.75
Aggressor (identity)	0.85	0.85	0.84	0.84
Sexual assault	0.93	0.93	0.93	0.90
Type of femicide	0.82	0.85	0.80	0.82
Type of space	0.84	0.83	0.84	0.82

It is important to note that the classifiers that had better metrics were those that had fewer output neurons (especially binary ones), this is because the data were much better balanced.
Table 9 shows an example of the probabilities of labels detected in the note (a) shown in table 5.

Table 9. Example of characteristics detected according to their probabilities

Labels detected	**Probabilities labels**
(1) *Cause of death:* gunshot wounds	*(1) Prob = 0.7015*
(2) Type of weapon used: firearm	*(2) Prob = 0.6469*
(3) Age of the victim: Adult	*(3) Prob = 0.9173*
(4) Agressor's identity: Unknown (hitmen)	*(4) Prob = 0.9026*
(5) Sexual assault: No	*(5) Prob = 0.9877*
(6) Type of femicide: Transphobic	*(6) Prob = 0.9274*
(7) *Type of space where the aggression occurred:* Public	*(7) Prob = 0.9942*

5.4 Results module 4

In this last module, we found some spatial patterns for different characteristics or labels and with different spatial units.

5.4.1 Municipality

Table 10 shows the global values of Moran's I obtained for *the number of femicides counted* and the characteristics studied, considering the municipalities of the metropolitan area of the valley of Mexico as spatial units.

Table 10. Municipal spatial pattern values

Label	*Moran's index*	*p-value*	*z-score*
Number of femicides	0.237	0.005	3.368
Cause of death	0.233	0.003	3.505
Intimate femicide	0.113	0.115	1.196
Sexual assault	0.057	0.174	0.894

The values obtained show that for the number of femicides and violent deaths together with the p-value < 0.05 allows us to reject the null hypothesis (H_0), which means that for these characteristics there is spatial autocorrelation, i.e., there is a spatial pattern. For the other two variables, according to the p-value, we cannot reject the null hypothesis, i.e., there is no spatial pattern. Figure 9 shows the densities of femicides by municipality and the clusters for the local Moran index.

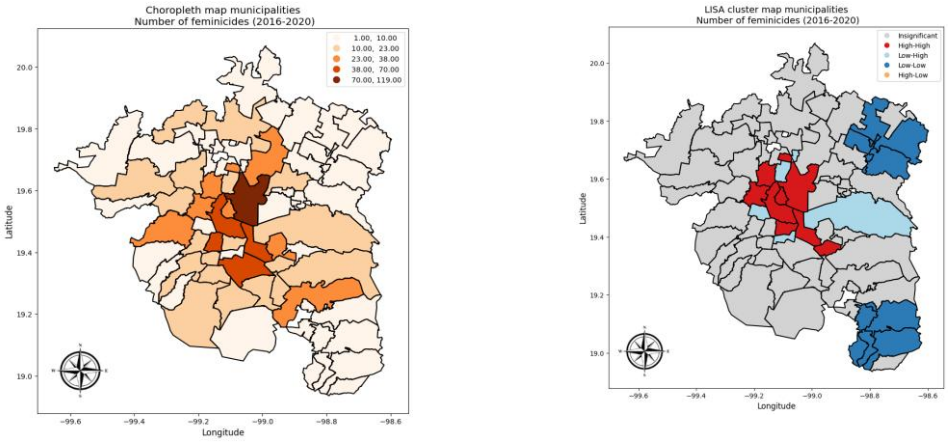

Figure 9.- Choropleth and LISA cluster maps for each municipality.

The LISA cluster map shows that municipalities with a positive correlation above the mean (High-High) tend to form patterns in the center of the valley of Mexico, these are in the north of Mexico City and share a border with the state of Mexico. Among these municipalities are Ecatepec de Morelos, Tlalnepantla de Baz, Nezahualcóyotl, La Paz and Tultitlán in the state of Mexico, in Mexico City are the municipalities of Gustavo A. Madero and Venustiano Carranza. And the municipalities with positive correlation but below average (Low-Low) are in the extreme southeast and northeast of the state of Mexico.

5.4.2 Geohash

Table 11 shows the global values of Moran's I for each of the studied features, considering the geohash mesh with *precision 4* as spatial units. It is worth mentioning that the mean area of each spatial unit = 760.686 km^2

Table 11. Spatial pattern values with Geohash accuracy 4

Label	*Moran's index*	*p-value*	*z-score*
Number of femicides	0.175	0.043	1.975
Cause of death	0.173	0.053	1.805
Intimate femicide	0.133	0.108	1.347
Sexual assault	-0.077	0.418	0.060

The values obtained show that only for the number of femicides with the p-value < 0.05 allows us to reject the null hypothesis (H_0), which means that for this characteristic there is spatial autocorrelation, i.e., there is a spatial pattern. For the other three variables, according to the p-value, we cannot reject the null hypothesis, i.e., there is no spatial pattern. Figure 10 shows the femicide densities by dividing the metropolitan area of the valley of Mexico into geohash units of the same size with precision 4 and the clusters for the local Moran's index.

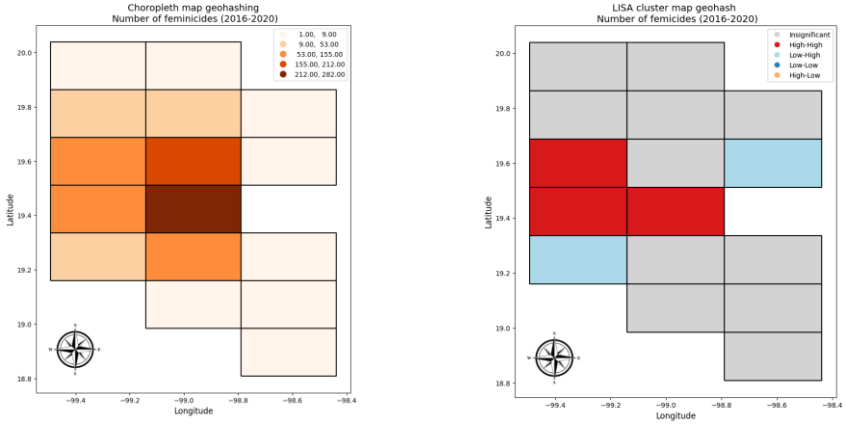

Figure 10.- Choropleth and LISA cluster maps for each precision 4 geohash unit.

This LISA cluster map shows that for precision 4 of the geohash spatial units, in the units located in the center and west of the area covered, femicides have a positive and above average spatial autocorrelation, showing a somewhat similar behavior to the municipal spatial units. Table 12 shows the global values of Moran's I for each of the studied features, considering the geohash mesh with *precision 5* as a spatial unit. It is worth mentioning that the mean area of each spatial unit = 23.756 km^2

Table 12. Spatial pattern values with Geohash accuracy 5.

Label	*Moran's index*	*p-value*	*z-score*
Number of femicides	0.417	0.001	8.784
Cause of death	0.396	0.001	7.493
Intimate femicide	0.157	0.032	2.030
Sexual assault	0.165	0.026	2.240

The values obtained show that all characteristics with the p-value < 0.05 allows us to reject the null hypothesis (H_0), which means that for these characteristics there is spatial autocorrelation, i.e., there is a spatial pattern. Figure 11 shows the femicide densities by dividing the metropolitan area of the valley of Mexico into geohash units of the same size with precision 5 and the clusters for the local Moran's index.

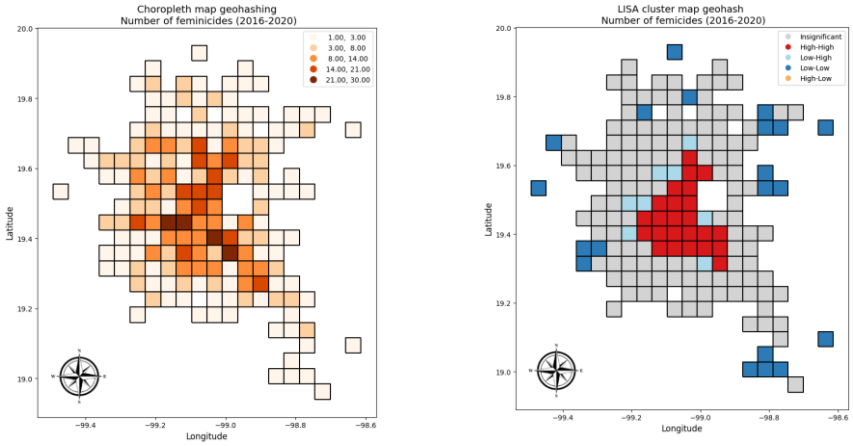

Figure 11.- Choropleth and LISA cluster maps for each precision 5 geohash unit.

This LISA cluster map shows that for precision 5 of the geohash spatial units, the behavior tends to be more like that of the municipal spatial units, especially in the center of the metropolitan area of the valley of Mexico. In this case it stands out that the units with positive spatial autocorrelation, but of low average (Low-Low) are in the periphery.
It is important to note that we did not consider the island units (isolated from the rest) for the calculations.

5.4.3 Hexagonal

Table 13 shows the global values of Moran's I for each of the studied features, considering the hexagonal mesh with *resolution 4* as the spatial unit. It is worth mentioning that the mean area of each spatial unit = 1,770.348 km^2

Table 13. Spatial pattern values with hexagonal resolution 4

Label	*Moran's index*	*p-value*	*z-score*
Number of femicides	-0.317	0.265	-0.699
Cause of death	-0.294	0.318	-0.624
Intimate femicide	-0.347	0.246	-0.808
Sexual assault	-0.084	0.224	0.498

The negative values obtained show that for the features chosen with this spatial unit, facts with such features tend to disperse and the p-value > 0.05 shows that we cannot reject the null hypothesis (H_0), i.e., there is no tendency to form spatial patterns.

Figure 12 shows the femicide densities by dividing the metropolitan area of the valley of Mexico into hexagonal units of the same size with resolution 4 and the clusters for the local Moran index.

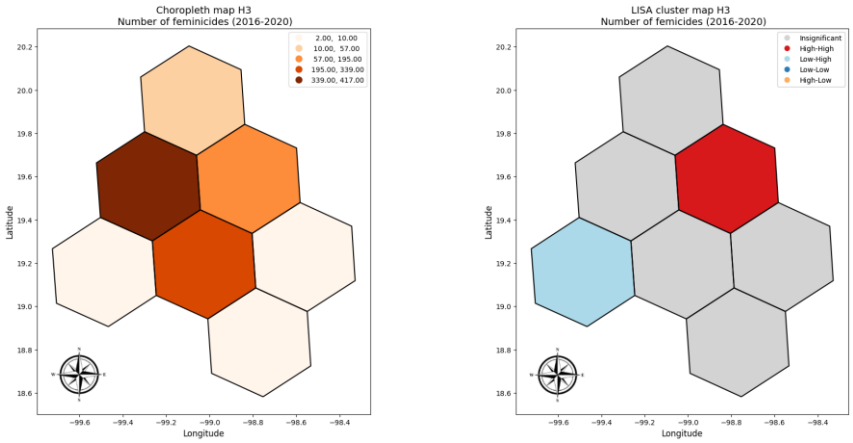

Figure 12.- Choropleth and LISA cluster maps for each resolution 4 hexagonal unit.

For this case, with hexagonal units of resolution 4, the values of positive autocorrelation above the mean (High-High) tended to cluster in the northeast of the metropolitan area of the valley of Mexico represented by the areas of these hexagons. In contrast to the municipal area units, the values in the northeast have positive autocorrelation but below the average (Low-Low). Table 14 shows the global values of Moran's I for each of the studied features, considering the hexagonal mesh with *resolution 5* as the spatial unit. It is worth mentioning that the mean area of each spatial unit = 252.904 km^2

Table 14. Spatial pattern values with hexagonal resolution 5

Label	*Moran's index*	*p-value*	*z-score*
Number of femicides	0.411	0.002	3.866
Cause of death	0.402	0.002	3.841
Intimate femicide	0.074	0.173	0.850
Sexual assault	0.145	0.097	1.330

The values obtained are like those shown in Table 5, with the difference that the spatial patterns are more notable than in the case of the municipalities. Figure 13 shows the femicide densities by dividing the metropolitan area of the valley of Mexico into hexagonal units of the same size with resolution 5 and the clusters for the local Moran index.

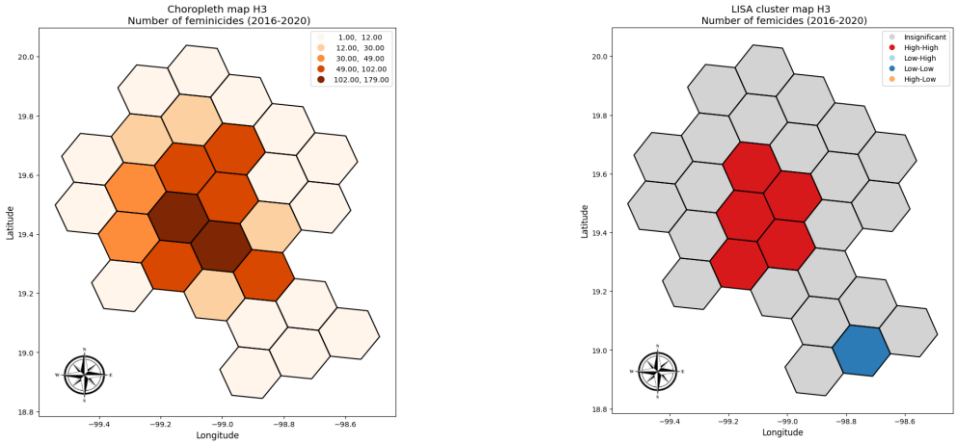

Figure 13.- Choropleth and LISA cluster maps for each resolution 5 hexagonal unit.

These spatial units, represented by hexagons of resolution 5, allow us to obtain a positive autocorrelation above the mean (High-High) like those obtained with the municipal spatial units, i.e., femicides tend to form a pattern in the center of the metropolitan area of the valley of Mexico. And femicide patterns with a value below the average (Low-Low) tend to form in the periphery.

6 Conclusions

We could verify that the artificial intelligence tools used reduce the time required to analyze texts with adequate accuracy, although some metrics can be improved by adding more texts for classifiers training. All the process with the validation data required almost 30 minutes between the input to the output according to the methodology shown in Figure 1. The experts in violence have estimated that the analysis of the 2,000 validation notes would take between 6 and 8 months by manual methods with a team of 4 people. It is important to emphasize this, since the speed of the analysis of these crimes will define the speed of the responses to address this problem.

The results after analyzing the femicides that occurred in the metropolitan area of the valley of Mexico show that these could be useful to analyze these crimes at a national level. It is important to emphasize the importance of models used in module 3, as the labels detected by these models are useful for experts in violence in their investigations, although they have mentioned the need to improve the accuracy of such models of classification and expand the labels used.

The spatial patterns obtained in module 4 show that population density does not usually influence the occurrence of femicides, and these patterns tend to repeat or resemble each other, regardless of the spatial unit used.

In any case, *results suggest that these crimes should be addressed at a lower level than the municipal one, and that some of the spatial units used should be chosen with different levels of resolution*.

References

Anselin, L. (1995). Local indicators of spatial association—LISA. *Geographical Analysis, 27*(2), 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

Arize AI. (n.d.). *Embeddings: Meaning, examples and how to compute*. Retrieved December 21, 2023, from https://arize.com/blog-course/embeddings-meaning-examples-and-how-to-compute/

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics, 3*(1), 1–27. https://doi.org/10.1080/03610927408827101

Cámara de Diputados del H. Congreso de la Unión. (2007). *Ley general de acceso de las mujeres a una vida libre de violencia* [PDF]. Retrieved April 20, 2022, from https://www.diputados.gob.mx/LeyesBiblio/pdf/LGAMVLV.pdf

Cañete, J., Chaperon, G., Fuentes, R., Ho, J., Kang, H., & Pérez, J. (2020). Spanish pre-trained BERT model and evaluation data. *arXiv preprint arXiv:2308.02976*. https://arxiv.org/abs/2308.02976

Cliff, A. D., & Ord, K. (1969). The problem of spatial autocorrelation. In A. J. Scott (Ed.), *London Papers in Regional Science* (pp. 25–55). Pion.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1*(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

Donath, W. E., & Hoffman, A. J. (1973). Lower bounds for the partitioning of graphs. *IBM Journal of Research and Development, 17*(5), 420–425. https://doi.org/10.1147/rd.175.0420

Ghaemi, Z., & Farnaghi, M. (2019). A varied density-based clustering approach for event detection from heterogeneous Twitter data. *ISPRS International Journal of Geo-Information, 8*(2), Article 82. https://doi.org/10.3390/ijgi8020082

Ghaemi, Z., & Farnaghi, M. (2023). Event detection from geotagged tweets considering spatial autocorrelation and heterogeneity. *Journal of Spatial Science, 68*(3), 353–371. https://doi.org/10.1080/14498596.2021.2002201

Google LLC. (n.d.). *Datos contra el feminicidio* [Chrome extension]. Retrieved January 12, 2024, from https://chromewebstore.google.com/detail/datos-contra-el-feminicid/nimnliogakaliffledlfadpocbmillma?pli=1

Hinton, G. E., & Roweis, S. (2002). Stochastic neighbor embedding. In *Advances in Neural Information Processing Systems* (Vol. 15).

Instituto Nacional de Estadística y Geografía (INEGI). (n.d.). *Tableros estadísticos VCMM* [Web page]. Retrieved September 10, 2023, from https://www.inegi.org.mx/tablerosestadisticos/vcmm/

Instituto Nacional de las Mujeres (INMUJERES). (n.d.). *Desigualdad en cifras* [PDF]. Retrieved December 19, 2023, from http://cedoc.inmujeres.gob.mx/documentos_download/BoletinN3_2019.pdf

Laureano de Jesús, Y., De Ita Luna, G., & Tovar Vidal, M. (2020). Detección automática de zonas de alto riesgo de eventos delictivos a través de noticias periodísticas. *Research in Computing Science, 149*(8), 213–225.

Liu, P., Zhou, D., & Wu, N. (2007). VDBSCAN: Varied density-based spatial clustering of applications with noise. In *Proceedings of the 2007 International Conference on Service Systems and Service Management* (pp. 1–4). IEEE. https://doi.org/10.1109/ICSSSM.2007.4280175

Mata-Santel, J., Luna-Gijón, G., & Ronquillo-Bolaños, A. (2023). Análisis de notas periodísticas sobre violencia contra las mujeres mediante estrategias de gestión de datos desde el diseño de información. *Acta Universitaria,*(33), 1–17. https://doi.org/10.15174/au.2023.3664

McInnes, L., & Healy, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction [Preprint]. https://doi.org/10.48550/arXiv.1802.03426

Monárrez, F. J. (2010). *Las diversas representaciones del feminicidio y los asesinatos en Ciudad Juárez, 1993–2005* (pp. 353–398). El Colegio de la Frontera Norte.

Olaya, V. (2020). *Sistema de información geográfica: Conceptos básicos para el análisis espacial* (Vol. 1, pp. 204–206). Creative Commons.

Principal component analysis and factor analysis. (n.d.). In *Springer Series in Statistics* (pp. 150–166). https://doi.org/10.1007/0-387-22440-8_7

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Secretariado Ejecutivo del Sistema Nacional de Seguridad Pública (SESNSP). (n.d.). *Información sobre violencia contra las mujeres (incidencia delictiva y llamadas de emergencia 9-1-1)* [Web page]. Retrieved September 10, 2023, from https://www.gob.mx/sesnsp/articulos/informacion-sobre-violencia-contra-las-mujeres-incidencia-delictiva-y-llamadas-de-emergencia-9-1-1-febrero-2019

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(8), 888–905. https://doi.org/10.1109/34.868688

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography, 46*, 234–240. https://doi.org/10.2307/143141

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*(86), 2579–2605.

Zhang, T., Ramakrishnan, R., & Livny, M. (1996). BIRCH: An efficient data clustering method for very large databases. *SIGMOD Record, 25*(2), 103–114. https://doi.org/10.1145/235968.233324