Pattern Recognition in Road Safety: Uncovering the Latent Causes of Accidents on Mexico's Federal Highways

Diana Zepeda-Martínez[1], Angélica Guzmán-Ponce[2], R. María Valdovinos-Rosas[1(✉)], and David Joaquín Delgado-Hernández[1]

[1] Faculty of Engineering, Autonomous University of the State of Mexico, Cerro de Coatepec, 50100 Toluca, State of Mexico, Mexico
{dzepedam001,rmvaldovinosr,david.delgado}@uaemex.mx
[2] Institute of New Imaging Technologies, Department of Computer Languages and Systems, Universitat Jaume I, Av. Vicent Sos Baynat, 12071 Castelló de la Plana, Spain
aguzman@uji.es

Abstract. Land transportation in Mexico plays a crucial role in ensuring connectivity and facilitating the mobility of both people and commodity. Nevertheless, this sector confronts substantial challenges, predominantly related with road accidents. Understanding the factors that contribute to these accidents is essential to developing and implementing effective safety strategies to reduce their frequency and severity. This research uses two unsupervised methods: latent Dirichlet allocation analysis (LDA) and the K-means algorithm, to identify the underlying factors responsible for road accidents in Mexico. LDA uncovers latent thematic structures in accident reports, revealing patterns in textual descriptions, and K-means identifies groups of accidents that share common attributes. The study period is from the years 2015 and 2019. The results suggest that traffic accidents are significantly influenced by a combination of factors such as driver behavior, road conditions, weather conditions and weather patterns.

Keywords: Road Accidents · Highways · Latent Dirichlet Allocation · Latent Topics · Clustering

1 Introduction

In Mexico, in 2022 847,716 deaths were reported, 90% were due to some disease or its consequences, in the remaining 10%, traffic accidents are among the ten main causes of death.[1]. In 2020, 13,630 people died from injuries in traffic accidents,

[1] https://www.inegi.org.mx/contenidos/saladeprensa/boletines/2023/EDR/EDR2022-Dft.pdf.

E. Mezura-Montes et al. (Eds.): MCPR 2024, LNCS 14755, pp. 13–22, 2024.
https://doi.org/10.1007/978-3-031-62836-8_2

corresponding to a mortality rate of 11 per 100,000 inhabitants[2]. About it, in 2021, an analysis of urban traffic incidents reported 340,415 accidents, where 3,849 of them had at least one death and 60,584 with injuries[3].

The analysis of traffic accidents reveals an inherent complexity in their causes and consequences. In this context, the use of machine learning (ML) and pattern recognition techniques emerges as a promising approach, with the potential to transform our understanding and response to traffic-related issues, such as the proposed by Yassin & Poooja [9], focused on accident severity in developing countries, examining factors like road conditions, lighting, weather, and casualty data (age and type). Using data from Addis Ababa's federal traffic police, including categorical and numerical data, their study implemented a hybrid method combining K-means clustering and Random Forest to identify critical variables affecting accident severity.

In Mexico, Saldana et al. [7] predict urban traffic congestion using a semi-supervised ML model. Focusing on Mexico City traffic dynamics, the study combines Support Vector Machine regression with data from Twitter, to forecast traffic events. The methodology integrates geospatial models, allowing traffic congestion to be visualized and upcoming traffic events to be predicted.

Sosa [8] applied Artificial Neural Networks (ANN) to predict car accidents in Monterrey, Mexico, analyzing variables like weather, location, time, date, type, and cause of accidents. The study used various data types across different study area sizes, including textual, categorical, time series, and images like diagrams. Using historical data from 2017, the ANN model effectively predicted accidents, though its efficiency was notably influenced by the variables, particularly the accident cause.

Daniel et al. [6] analysed the timing and occurrence of traffic accidents in Mexico City between 2014 and 2019, using data from the C5 open data portal[4]. Employing rule-based decision trees on a hexagonal grid framework, they generated rules associating points of interest with accident rates. Key findings include a high concentration of accidents in five Mexico City delegations, significant temporal patterns, and the critical role of intersections, traffic lights, and other relevant points in accident occurrences.

Hernández [4] developed a model using the information of the National Institute of Statistics and Geography on Traffic Accidents in urban and suburban areas during 2020 in Mexico[5]. The goal was to predict the severity of car accidents through data cleaning, balancing and feature selection processes. Random Forest, instance-based algorithms like KNN and ANN were implemented. The results indicate that Random Forest outperformed other models.

[2] https://www.gob.mx/cms/uploads/attachment/file/818181/Informe_SV_2021_HD2_compressed.pdf.

[3] https://www.inegi.org.mx/contenidos/saladeprensa/boletines/2021/accidentes/ACCIDENTES_2021.pdf.

[4] https://datos.cdmx.gob.mx/explore/dataset/incidentes-viales-c5/information/.

[5] https://www.inegi.org.mx/contenidos/saladeprensa/boletines/2021/accidentes/ACCIDENTES_2021.pdf.

From the efforts made by the scientific community, few studies focus on finding hidden patterns that allow identifying the impact that car accidents have on society in general. Derived from this, this study focuses on identifying hidden factors that precipitate the occurrence of traffic accidents on Mexican roads during the period from 2015 to 2019. For this, two unsupervised machine learning techniques were implemented: LDA and K-means algorithms.

2 Unsupervised Methods

2.1 Latent Dirichlet Allocation

The Latent Dirichlet Allocation (LDA) algorithm represents an unsupervised statistical approach employed to elucidate sets of observations that are not directly observed (latent). Let a dataset of M samples and N_d attributes, the essence of the LDA method is to model each sample as a mixture of several features (Algorithm 1). These features are represented by a multinomial distribution, which in turn is derived from a predefined Dirichlet distribution [3].

Algorithm 1. Latent Dirichlet Allocation

Require: M, N_d, α, T, ξ
Ensure: A distribution of topics and characteristics for each sample in the collection.
Initialize β: Matrix $T \times N$ for conditional distribution
for each sample $m = 1$ to M **do**
  $N_d \sim Poisson(\xi)$ ▷ Number of sample characteristics
  $\theta \sim Dirichlet(\alpha)$ ▷ Vector of topic mixture parameters
  for each feature $n = 1$ to N_d **do**
    $z_n \sim Multinomial(\theta)$ ▷ Select a topic
    $w_n \sim p(w_n|z_n, \beta)$ ▷ Select a word from the conditional distribution
  end for
end for

In the Algorithm 1: α sets the initial probability for the Dirichlet distribution. θ indicates how different topics are combined in a sample. z_n links the features in the samples to the underlying topics. β is a vector composed of N terms from the feature vocabulary. w is the specific features that appear in the samples.

The algorithm's outputs are: the patterns distribution or topics (β), the distribution of samples represented in the topic space (θ). To ensure the quality of the topics obtained, two metrics are used [1]: *perplexity* and *coherence*.

Perplexity. Perplexity determines how efficiently a model can handle previously unseen data [1]. In LDA, identify the optimal number of topics according to Eq. 1. Generally, there is an assumed inverse relationship between perplexity and model accuracy: lower perplexity implies higher accuracy [1]. This principle underlies the use of perplexity as a key indicator in LDA optimization, aiding

in the selection of an appropriate number of topics to enhance the ability of the model to classify and understand the data.

$$Perplexity = \exp\left(\frac{\sum_{d=1}^{M} \log P(w_{dn})}{\sum_{d=1}^{M} N_d}\right) \tag{1}$$

where, N_d represents the number of features in a sample, M denotes the total number of samples in the dataset. $P(w_{dn})$ is the probability of observing a specific word w in the n-th position of the sample d, conditional on the assigned topic z_n an the distribution β for that topic.

Coherence. Coherence focuses on assessing the clarity and consistency of themes from the human view, that is to say, C_v evaluates to what degree the induced themes of an LDA model are correlated with each other, based on conditional probability (Eq. 2). The higher coherence value, greater will be the probability of obtaining greater precision from that model [1].

$$C_v = \sum_{i<j} log \frac{D(w_i, w_j) + 1}{D(w_i) + 1} \tag{2}$$

where, $D(w_i)$ is the frequency of samples containing the word w_i, and $D(w_i, w_j)$ the frequency of samples containing both words.

2.2 K-Means Algorithm

The K-means clustering algorithm is a unsupervised learning method that partitioned a dataset into K distinct non-overlapping subsets or clusters. This partitioning is achieved by assigning each data point to the cluster with the nearest mean, serving as a cluster prototype [9]. The goal of the K-means algorithm is to minimize the within-cluster sum of squares (WCSS), also referred to as inertia, represented as follow Eq. 3:

$$\sum_{i=1}^{M} \min_{\mu_j \in C} \left(||x_i - \mu_j||^2\right) \tag{3}$$

where x_i denotes the i-th sample with N_d attributes, and μ_j is the centroid of the j-th cluster. The centroid is conceptualized as the mean position of all points in the cluster representing the centre of the cluster. The set C encompasses all clusters, where each cluster j is characterized by its centroid μ_j. The goal is to determine a division where the total sum of squared distances from each sample to the mean of its assigned cluster is minimized (Algorithm 2).

In Algorithm 2, the iterative process involves reassigning data samples among various clusters until a stable clustering configuration is achieved or when a predefined stopping criterion is met, such as a specified number of iterations [9].

Jambu's Elbow method is commonly used to find the optimal value of K [2]; this method identifies the *elbow* point in a plot of the sum of squared distances

Algorithm 2. K-Means

Require: M: samples, K: Number of clusters
Ensure: Cluster assignment for each instance in the dataset.
Initialize *centroids*: List of K initial centroids
Initialize *clusters*: List of K empty sets to store instances in each cluster
Randomly assign instances to the initial centroids
repeat
    for each sample x in M **do**
        Calculate the distance between x and each centroid at *centroids*
        Assign x to the cluster of the nearest centroid
    end for
    for each cluster C en *clusters* **do**
        Calculate the new centroid as the average of the instances in C
    end for
until Convergence or maximum number of iterations achieved

from points to their cluster centroids against different K values. This point indicates that increasing the number of clusters does not significantly improve intra-cluster homogeneity.

3 Methodology

3.1 Information Acquisition

The data used were extracted from the official website of the Mexican Government[6]. It includes 66,008 samples from traffic accident reports between 2015 and 2019, detailing locations, road conditions, times, dates, and specifics about vehicles and victims. The data are in textual and tabular formats with numerical and categorical components. 15,038 samples were excluded due to missing data.

Each sample contains 120 attributes Also, information regarding accident victims, encompassing those injured and deceased and pedestrians, was amalgamated into a singular *victims* column. Finally, various columns about the classification of accidents and the circumstances that contributed to them were categorized. The final dataset is composed by 50,970 samples and 15 attributes.

3.2 Data Pre-processing

A cleaning phase was applied, in which less relevant attributes were excluded under the guidance of an expert in the study area, in order to preserve only relevant features. Likewise, the data were adapted to the requirements of the LDA and K-means algorithms.

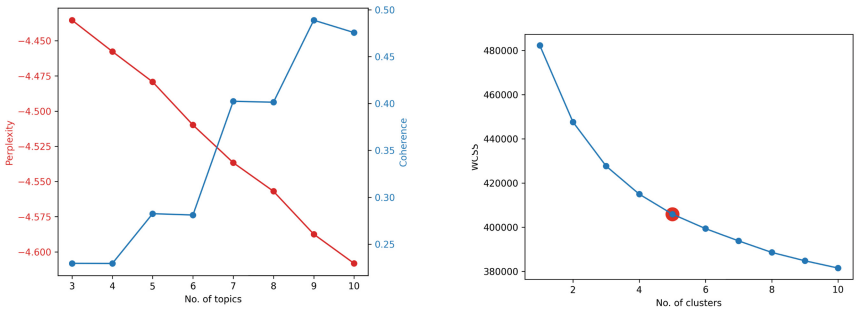
For the LDA algorithm, the attributes were transformed into textual format. While for the K-means algorithm, they were kept numerical and one-hot coding was used to adjust them. This technique changes categorical variables to a binary

[6] https://datos.gob.mx/busca/dataset/policia-federal.

format, representing the presence or absence of a category with a numerical value. Upon completing the coding process, 195 attributes were obtained.

3.3 Uncovering Latent Patterns

The implementation of the algorithms was carried out using Python, with the main library for topic modelling, Gensim[7], and for K-means sci-kit-learn[8].



(a) Perplexity and coherence metrics for various numbers of topics.

(b) Jambu's Elbow method for optimal cluster determination.

Fig. 1. Comparative visual analysis for optimizing topic and cluster numbers using perplexity, coherence, and Jambu's Elbow method.

The effectiveness of the LDA model was evaluated using the perplexity and consistency metrics, varying from 3 to 10 topics to determine the optimal number. The lowest perplexity, indicating higher accuracy, was achieved with 10 topics, despite 9 topics showing the highest consistency. The choice of a 10-topic model was driven by the significant improvement in model predictions through minimizing perplexity (See Fig. 1a). Jambu's elbow method determined the optimal number of clusters for the K-means algorithm, identifying 5 clusters as the point with the lowest inertia decline rate (See Fig. 1b).

3.4 Assessment Metric

The silhouette coefficient (SC) was used to validate the clustering quality in this analysis (Eq. 4). From -1 to 1, it measures the similarity of a sample to its group versus others. A score near 1 indicates robust matching, -1 implies poor matching, and around 0 suggests equal distance from clusters [5]. In K-means, SC assesses the similarity of a sample to others in its group (cohesion) versus neighbouring groups (separation), calculating both within-group average

[7] https://radimrehurek.com/gensim/index.html.
[8] https://scikit-learn.org/stable/.

similarity ($a(i)$) and with the closest different group ($b(i)$) to measure internal coherence.

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{4}$$

For the LDA algorithm, the (SC) assess the coherence of topics extracted from a dataset, i.e. SC measures how distinctly topics are defined and how well samples map to them. A high (SC) value suggests that the samples have a solid membership to the identified topics, which indicates good topic separation and definition in the LDA model.

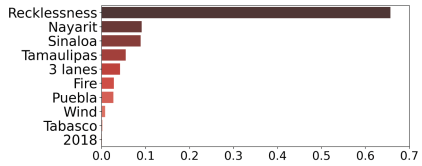
4 Results

4.1 Latent Topics in Accidents

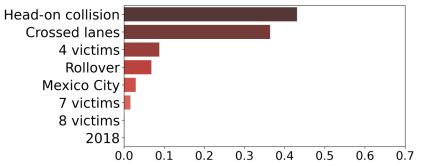
In Fig. 2 reveals that in states like Nayarit, Puebla, Sinaloa, Tabasco, and Tamaulipas, morning accidents on three-lane roads are prevalent, often attributed to reckless driving and adverse weather conditions such as wind and fires. Conversely, Mexico City experiences accidents primarily from lane invasions during overtaking, frequently leading to head-on collisions and vehicle rollovers, with four to nine victims per incident (Fig. 2b). In Fig. 2c shows that in Aguascalientes, Nuevo León, Tabasco, Jalisco, and Morelos, accidents occur mainly on Sunday mornings, typically due to failures in yielding the right of way, averaging five victims. In Yucatán, Durango, and Sonora, the peak accident periods are in April and August, on Saturdays and Wednesdays, involving two victims per incident (Fig. 2d). The State of Mexico records frequent accidents on five-lane roads caused by absent road signs, unsafe loads, and fog, with an average of one victim (Fig. 2e). In Durango and San Luis Potosí (Fig. 2f), September Tuesdays see accidents mainly from driver errors, like dozing or improper parking, and vehicle malfunctions, resulting in two to six victims. Michoacán, Hidalgo, and Sinaloa report afternoon rollovers and brake failures, affecting up to six individuals. Seasonal traffic trends in October and November indicate a rise in accidents during twilight in Guanajuato, Chiapas, and Tlaxcala, possibly linked to the Day of the Dead celebrations. Lastly, July in Querétaro, Quintana Roo, Coahuila, Baja California, Campeche, and Guerrero sees the most accidents on single-lane roads due to road conditions, driver behaviour, and weather (Fig. 2j). December features night accidents on two- and four-lane roads caused by weather and driver recklessness, correlated with the Christmas and New Year festivities.

4.2 K-Means in Accidents

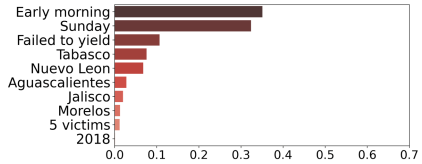
The K-means algorithm analysis shows that road accidents in Guanajuato, Jalisco, San Luis Potosí, Chiapas, and the State of Mexico predominantly occur during the night and early morning of Sundays in December 2015, involving 0 to 4 victims. Common causes include falling objects, driver recklessness, impacts with



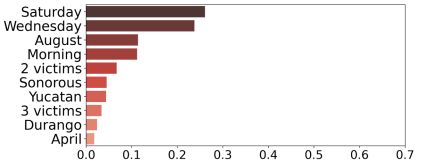
(a) Reckless Driving and Locations



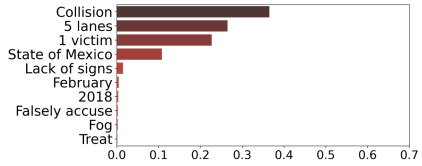
(b) Types of Collisions and Victims



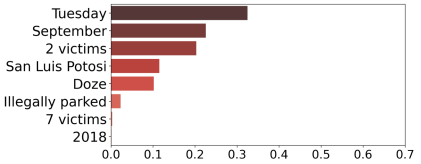
(c) Morning Accidents and Regions



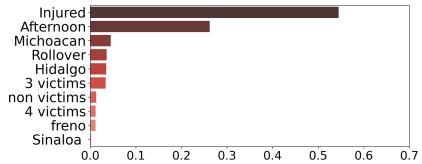
(d) Weekend Traffic Incidents



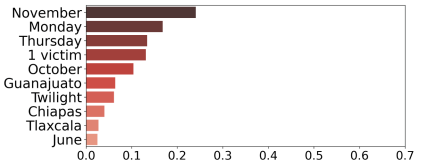
(e) Highway Collisions and Road Conditions



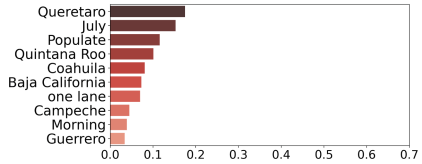
(f) Tuesday Accidents in September



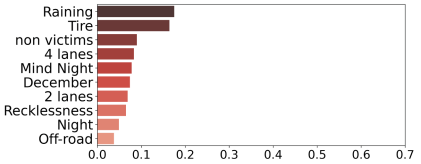
(g) Injuries and Regional Trends



(h) Monthly Traffic Trends



(i) State-Specific Incidents



(j) Rainy Weather and Road Conditions

Fig. 2. Topics on traffic accidents on Mexican highways from 2015 to 2019

stationary vehicles, and lane invasions, suggesting a link with weekend trips and holiday festivities during Christmas and New Year. Similarly, in December 2018, Veracruz, State of Mexico, and Jalisco experienced frequent accidents on 4-lane roads, mainly involving 1 to 2 victims from driver deviations, pedestrian collisions, and rear-end impacts, correlating with the Christmas season. In June and July 2015, incidents on four-lane highways during weekend nights resulted in 0 to 3 victims due to excessive speed and adverse weather. The trend continued in 2018

with accidents in Morelos and Guanajuato on Sunday afternoons, particularly in July and May, on two-lane roads involving excessive speed and rain. Additionally, in March and April 2015, evening accidents were frequent in Veracruz, Chiapas, and Guanajuato during weekends, causing 0 to 3 victims from driver recklessness, leading to head-on collisions on 4-lane highways.

4.3 Comparative Evaluation of Models

The results obtained are summarized in Fig. 3a. The word cloud reveals an emphasis on the terms 'victim,' 'Morning,' and 'Recklessness,' as well as several urban locations.

(a) Word cloud from the topics (b) Word cloud from the clusters

Fig. 3. Word clouds

Additionally, the (SC) was employed to assesses the coherence and separation within the clustering, providing a quantitative evaluation of the clustering quality, determining the cohesion and separation of the structures identified by the LDA and K-means algorithms. Although K-means is inherently designed to optimize such a metric, using it with LDA allow to validate the clarity and robustness of the topics generated.K-means exhibited a substantially higher (SC) (0.5781585923153798) compared to LDA (0.0252322471173679). This finding suggests that K-means formed more cohesive and separate clusters in the dataset.

5 Conclusions

Traffic accidents in Mexico are a significant public health issue and a leading cause of mortality and injuries. This research employed LDA and K-means to analyse road accident patterns across Mexican states. LDA identified ten latent topics linked to reckless driving, collisions, and environmental influences, revealing complex accident scenarios shaped by regional traits and weather conditions. K-means categorised incidents by timing, location, and impact, highlighting prevalent trends such as night and early morning accidents during holidays, influenced by environmental conditions like rain. These methods provided a detailed

understanding of the multifaceted nature of traffic accidents, emphasising the roles of driver behaviour, road conditions, and temporal factors. Recommendations include developing targeted safety campaigns and enhancing enforcement during high-risk periods. Future studies should analyse long-term trends, integrate socioeconomic and demographic data, create predictive models, and assess the efficacy of existing road safety policies.

References

1. Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., Hassan, A.: Topic modeling algorithms and applications: a survey. Inf. Syst. **112**, 102131 (2023). https://doi.org/10.1016/j.is.2022.102131
2. Amorim, B.D.S.P., Firmino, A.A., Baptista, C.D.S., Júnior, G.B., Paiva, A.C.D., Júnior, F.E.D.A.: A machine learning approach for classifying road accident hotspots. ISPRS Int. J. Geo-Inf. **12**(6), 227 (2023). https://doi.org/10.3390/ijgi12060227
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**(Jan), 993–1022 (2003)
4. Hernández, J.J.O.: Aplicación y valoración de algoritmos de Machine Learning para la predicción de gravedad en accidentes automovilísticos. Ph.D. thesis, Benemérita Universidad Autónoma de Puebla (2023)
5. Lenssen, L., Schubert, E.: Medoid silhouette clustering with automatic cluster number selection. Inf. Syst. **120**, 102290 (2024). https://doi.org/10.1016/j.is.2023.102290
6. Otero, D.E.R.: Descubrimiento y representación de patrones de accidentes de tránsito en la Ciudad de México usando técnicas geoestadísticas y aprendizaje máquina. Ph.D. thesis, INFOTEC Centro de Investigación e Innovación en Tecnologías de la Información y Comunicación (2021)
7. Saldana-Perez, M., Torres-Ruiz, M., Moreno-Ibarra, M.: Geospatial modeling of road traffic using a semi-supervised regression algorithm. IEEE Access **7**, 177376–177386 (2019). https://doi.org/10.1109/ACCESS.2019.2942586
8. Sosa, E.E.C.: Análisis y predicción de accidentes automovilísticos mediante la aplicación de la red neuronal artificial de máxima sensibilidad y un prototipo de sistema web para la visualización de la información. Ph.D. thesis, Universidad Autónoma de Nuevo León (2019)
9. Yassin, S.S.: Pooja: road accident prediction and model interpretation using a hybrid k-means and random forest algorithm approach. SN Appl. Sci. **2**(9), 1576 (2020). https://doi.org/10.1007/s42452-020-3125-1