

# Combining Linear Regressions to Determine the Future of the Covid in Mexico from the News

Gabriela Carmona-Sánchez<sup>1</sup>, Ángel Carmona<sup>1</sup> and Miguel Ángel Álvarez-Carmona<sup>2,\*</sup>

<sup>1</sup>Benemérita Universidad Autónoma de Puebla (BUAP), 72000, Puebla, Mexico

<sup>2</sup>Centro de Investigación Científica y de Educación Superior de Ensenada, Unidad de Transferencia Tecnológica Tepic (CICESE-UT3), 63173, Nayarit, Mexico

## Abstract

This paper presents an approach to determine the Semaphore Covid in Mexico from the news to participate in the Rest-Mex 2022 evaluation forum. The purpose of the task is to determine the covid semaphore color (red, orange, yellow, and green) in different time spaces. The proposed approach consists of two main steps. First, to generate a list of topics of the news, and second, to implement several linear regressions methods in order to these results serve to feed a deep neural network. For the first step, the LDA algorithm was implemented, and for the second, well-known methods such as Lasso, Ridge, Lars, among others, were utilized. With this approach, a weighted average of 0.48 was obtained, which is considerably higher than the baseline proposed by the organizers, which is 0.12. The best result to classify the semaphore was two weeks in the future with 0.56 of F-measure.

## Keywords

Semaphore Covid, Mexican Spanish, LDA, Linear Regressions.

## 1. Introduction


The covid-19 pandemic generated economic havoc in various areas. Tourism was one of the sectors most affected by the restrictions applied to minimize the impact of the pandemic on public health [1][2][3]. From this, the Mexican government implements the epidemiological semaphore of covid-19, which is a system of four ordered colors [4]. These four colors are defined in an order that establishes the capacity of activities that will be allowed [5]:


- Red color: Only essential economic activities will be allowed, and people will also be allowed to go for a walk around their homes during the day.
- Orange color: In addition to essential economic activities, companies in non-essential economic activities will be allowed to work with 30 % of the staff for their operation, always considering the forms of maximum care for people with the highest risk of presenting a severe covid-19 illness. Open public spaces with a reduced capacity (number of people) will be opened.
- Yellow color: All work activities are allowed, taking care of people with a higher risk of presenting covid-19. Open public spaces are available regularly, and closed public spaces

IberLEF 2022, September 2022, A Coruña, Spain

✉ malvarez@cicese.edu.mx (M. Álvarez-Carmona)

ORCID 0000-0003-4421-5575 (M. Álvarez-Carmona)

 © 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

can be opened with reduced capacity. As in other ES colors, these activities must be carried out with basic prevention measures and utmost care for people with a higher risk of presenting covid-19.

- Green color: All activities are allowed, including school zones.

Suppose the evolution of the pandemic is known. In that case, it is possible to determine the color of the semaphore in the future and take measures to prevent certain inconveniences, especially in the tourist industry. One of the ways to know the evolution of the pandemic is through the news published on the internet, where they publish data related to covid, such as infections, hospitals, vaccines, and other essential statistics.

In this way, the organizers of Rest-Mex 2022 propose the task of determining the epidemiological semaphore through the news [6][7]. The Rest-Mex is an evaluation forum that emerged in 2021 [8]. For the 2022 edition, they took on the task of collecting a corpus of news regarding covid-19 for all the states of the Mexican Republic. The task aims to predict the semaphore's color for weeks 0, 2, 4, and 8 after a set of weekly news is released.

To establish the semaphore color of each week, the Mexican government considers several variables, such as the contagion curve, the number of deaths, the number of people recovered, and the capacity and saturation of hospitals, among others. Since this data is disclosed in the news, it is possible to take advantage of this information to determine the semaphore color

One of the main tools that can help extract and represent the essential ideas of the news is topic extraction [9]. This work hypothesizes that finding the hidden topics in the collection of news related to covid helps represent and predict the covid semaphore in the future.

In addition, to help the thematic representation, it is proposed to add a set of regressions that take as input the representation of topics of each instance and try to determine the relationship between that vector and its respective semaphore color. Although the semaphore labels are discrete, this representation should work since they are also ordinal.

The rest of the document is organized as follows; In section 2, the methodology followed in this work is described. In section 3, the results and their analysis are presented. Finally, section 4 presents the conclusions of this work.

## 2. Methodology

This section presents the database with which it experimented and the proposed methodology to represent the covid news. Also, the methodology is presented.

### 2.1. Covid News Collection

The training collection used for this task was built and shared by the Rest-Mex 2022 organizers [6] and consists of a total of 94,540 news items grouped in 1,912 instances. Each instance represents a week of news, mainly regarding the covid topic in each of the 32 states of Mexico.

Each instance has four labels:  $W_0$ ,  $W_2$ ,  $W_4$ ,  $W_8$ . For each label,  $W_i$ ,  $i$  means the number of weeks in the future after the news was published, and the possible value of each  $W_i$  could be: Red, Orange, Yellow, or Green.

The dataset showed an unequal distribution of classes. Most of the instances are sent to the orange color, while the red color is not even half of the instances of the orange color. The distributions for weeks 2, 4, and 8 are similar to week 0.

## 2.2. Pre-processing

In order to extract the most important characteristics of the texts, a pre-processing phase was applied. The transformations that were carried out are:

1. Uppercase was converted to lowercase.
2. Stop-words were removed.
3. Punctuation marks were removed.
4. The digits were replaced by the letter 'd'.
5. To avoid influencing the dates, the words for the months of the year were also removed.
6. Stemming was applied to the tokens in the texts.
7. Removed tokens that appear less than 50 times in the entire collection.

## 2.3. Topic Modeling on the Covid News

For this task, the idea is to represent the news in order to classify the color semaphore in the future. For this work, we propose to extract the hidden topics to build a vector because, in the context of text modeling [10], the topic probabilities provide an explicit representation of a document [11], and this could be a good representation for the semaphore prediction task.

For this propose, we use the well-known Latent Dirichlet Allocation (LDA) method. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities [11].

The first step for the representation proposed is to extract the LDA topics from the news. For this, we use the implementation of Gensim library in Python 3.0. We set the number of topics (k) in 500<sup>1</sup>.

## 2.4. Set of Linear Regressions on the Topic Modeling

For this work, we propose applying a set of 14 regressions methods. In particular, the methods are: Linear regression, Lasso, Ridge, SGD, ElasticNet, Lars, Orthogonal, ARD, Bayesian, Huber, TheilSen, Poisson, Tweedie, and PassiveAggressive [12]. Each of these methods computes a regression model differently. For these tasks, the features would be the 500 topics extracted from LDA, while the target that the models will seek to approximate would be the traffic light labels of the week that is being classified.

Once the regression methods' values are obtained, it is proposed to use a simple deep neural network to decide the semaphore color. In particular, a neural network with ten hidden layers ensures that the best relationship is found between the outputs of the participating systems and the real class of each instance. Table 1 summarizes the principal characteristics of the Deep

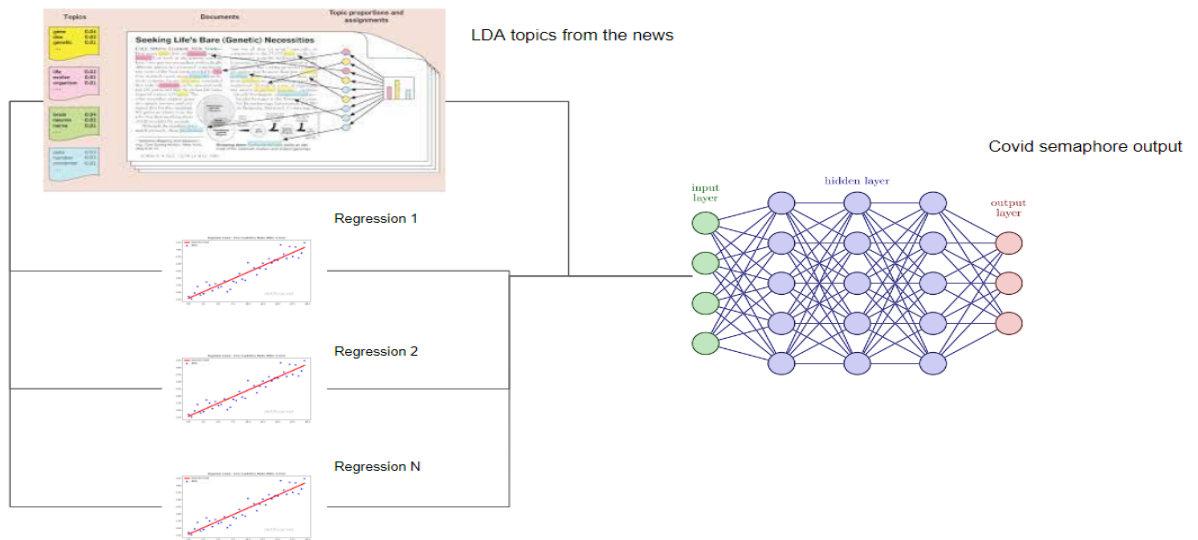
---

<sup>1</sup>Empirically selected

**Table 1**

Characteristics of the applied Deep Learning algorithm

Hidden Layers	10
Neurons per layer	1000
Activation function	Relu
Neurons of the final layer	4
Final layer	Softmax
Loss function	Categorical Cross Entropy
Optimizer	Adam
Epochs	50



**Figure 1:** Proposal to extract topics, calculate regression methods and feed the neural network

Learning algorithm. This architecture was chosen due to works as in [1] have achieved good results with this proposal.

Figure 1 shows the proposal of this work to determine the covid semaphore color.

### 3. Experiments and Results

In this Section, The obtained results are presented. First, the results derived from the training corpus are displayed. Subsequently, the official data published by the organizers of Rest-Mex 2022 are presented.

#### 3.1. Training Data Set Results

10-fold cross validation was performed for the training corpus to obtain the results. In addition, the final model was generated with all the training data to send the results and participate in

**Table 2**  
Training data set results

Week	Acc	Macro-F	F Red	F Orange	F Yellow	F Green
0	52.77	0.51	0.40	0.57	0.43	0.62
2	55.54	0.53	0.39	0.59	0.49	0.64
4	53.24	0.50	0.34	0.56	0.46	0.63
8	52.66	0.46	0.20	0.57	0.45	0.61
Measure	53.20	0.48	0.27	0.57	0.45	0.62

the Rest-Mex 2022 forum.

Table 2 shows the results obtained for the training set for all the target weeks, and this is for weeks 0, 2, 4, and 8. This table shows each class’s accuracy, macro F-measure, and F-measure (Red, Orange, Yellow, and Green).

For this edition, the authors propose a measure that gives more weight to well-ranked future weeks to obtain a final result. This measure is defined as shown in the equation 1.

$$measure = \frac{F_{w_0} + 2 * F_{w_2} + 4 * F_{w_4} + 8 * F_{w_8}}{15} \quad (1)$$

Also the Table 2 shows the global result using the equation 1. Although the organizers only considered the result of the macro F-measure average, the same equation was applied to all measurements.

From these results, it is possible to observe that, in general, the method works better for the green class, which is an unexpected result since it is not the majority class, while the worst-ranked class is the red one.

The best F-measure macro result is obtained for week 2. However, although the results decrease for weeks 4 and 8, having results above 0.50 is a good result considering that it is a problem of 4 classes and unbalanced.

### 3.2. Test Data Set Results

Table 3 shows the results for the test corpus for the proposal of this work. These results are also shown on the official Rest-Mex site <sup>2</sup>. In these results, it is possible to observe that for weeks 2 and 4, the highest values of the F-measure are found. These results are mostly no different than the results of the train. However, for week 0, a significant change can be noted from 0.51 of F.measure to 0.33, so clear overfitting is noticeable here.

Finally, Table 4 shows the result obtained by our proposal compared to the rest of the participants of the 2022 edition of Rest-Mex. It is possible to see that our proposal obtained second place in the competition. This result is interesting, considering that, although a deep neural network was used for the final classification, it has highly comparable results with more complex architectures such as Transformers.

<sup>2</sup><https://sites.google.com/cicese.edu.mx/rest-mex-2022/results>

**Table 3**

Test data set results

Week	Acc	Macro-F	F Red	F Orange	F Yellow	F Green
0	43.14	0.33	0.35	0.56	0.33	0.06
2	60.34	0.56	0.37	0.64	0.52	0.71
4	52.55	0.51	0.39	0.48	0.45	0.70
8	54.83	0.46	0.20	0.61	0.44	0.61

**Table 4**

Final Rest-Mex Rank

Team	Measure
MCE-Team-Run2	0.4899842235
Arandanito (Our proposal)	0.4835859170
MCE-Team-Run1	0.3287472507
ML-Team-Run2	0.2489134415
ML-Team-Run1	0.2270582096
The last	0.1757160160
Majority Class (Baseline)	0.1290361261

## 4. Conclusions

This work addressed the problem of predicting the covid semaphore with Mexican news data for the entire country.

This work was presented within the Rest-Mex 2022, where a corpus with more than 100 thousand news items is proposed.

A representation based on topics extracted from the news was powered with the LDA algorithm. Once the topics were obtained, a set of linear regressions was proposed, which served to generate models that will approach the final classification result and served as input to a simple architecture of a deep neural network.

The results obtained are competitive compared to the rest of the participants, obtaining second place in the competition and obtaining a performance similar to that of a method based on Transformers.

## References

- [1] M. A. Álvarez-Carmona, R. Aranda, R. Guerrero-Rodríguez, A. Y. Rodríguez-González, A. P. López-Monroy, A combination of sentiment analysis systems for the study of online travel reviews: Many heads are better than one, *Computación y Sistemas* 26 (2022). doi:<https://doi.org/10.13053/CyS-26-2-4055>.
- [2] S. Arce-Cardenas, D. Fajardo-Delgado, M. Á. Álvarez-Carmona, J. P. Ramírez-Silva, A tourist recommendation system: A study case in mexico, in: *Mexican International Conference on Artificial Intelligence*, Springer, 2021, pp. 184–195. doi:[https://doi.org/10.1007/978-3-030-89820-5\\_15](https://doi.org/10.1007/978-3-030-89820-5_15).

- [3] R. Guerrero-Rodríguez, M. Á. Álvarez-Carmona, R. Aranda, A. P. López-Monroy, Studying online travel reviews related to tourist attractions using nlp methods: the case of guanajuato, mexico, *Current Issues in Tourism* (2021) 1–16. doi:<https://doi.org/10.1080/13683500.2021.2007227>.
- [4] M. Á. Álvarez-Carmona, R. Aranda, Determinación automática del color del semáforo mexicano del covid-19 a partir de las noticias (2022). doi:<https://doi.org/10.1590/SciELOPreprints.3834>.
- [5] M. A. Álvarez-Carmona, R. Aranda, A. Y. Rodríguez-González, L. Pellegrin, H. Carlos, Classifying the mexican epidemiological semaphore colour from the covid-19 text spanish news, *Journal of Information Sciences* (2022). doi:<https://doi.org/10.1177/01655515221100952>.
- [6] M. Á. Álvarez-Carmona, Á. Díaz-Pacheco, R. Aranda, A. Y. Rodríguez-González, D. Fajardo-Delgado, R. Guerrero-Rodríguez, L. Bustio-Martínez, Overview of rest-mex at iberlef 2022: Recommendation system, sentiment analysis and covid semaphore prediction for mexican tourist texts, *Procesamiento del Lenguaje Natural* 69 (2022).
- [7] M. E. Aragón, M. A. Álvarez-Carmona, M. Montes-y Gómez, H. J. Escalante, L. V. Pineda, D. Moctezuma, Overview of mex-a3t at iberlef 2019: Authorship and aggressiveness analysis in mexican spanish tweets., in: *IberLEF@ SEPLN, 2019*, pp. 478–494.
- [8] M. Á. Álvarez-Carmona, R. Aranda, S. Arce-Cárdenas, D. Fajardo-Delgado, R. Guerrero-Rodríguez, A. P. López-Monroy, J. Martínez-Miranda, H. Pérez-Espinosa, A. Rodríguez-González, Overview of rest-mex at iberlef 2021: Recommendation system for text mexican tourism, *Procesamiento del Lenguaje Natural* 67 (2021). doi:<https://doi.org/10.26342/2021-67-14>.
- [9] M. Á. Álvarez Carmona, E. Villatoro Tello, M. Montes y Gómez, L. Vilaseñor Pineda, Author profiling in social media with multimodal information, *Computación y Sistemas* 24 (2020) 1289–1304. doi:<https://doi.org/10.13053/cys-24-3-3488>.
- [10] M. A. Alvarez-Carmona, A. P. López-Monroy, M. Montes-y Gómez, L. Villasenor-Pineda, H. Jair-Escalante, Inaoe’s participation at pan’15: Author profiling task, *Working Notes Papers of the CLEF* (2015) 103.
- [11] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, *Journal of machine Learning research* 3 (2003) 993–1022.
- [12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research* 12 (2011) 2825–2830.