

# Mexican Epidemiological Semaphore Color Prediction by Means of Mutual Information Features

Alejandra Romero-Cantón<sup>1</sup>, Ramon Aranda<sup>2,\*</sup>, Angel Diaz-Pacheco<sup>2</sup> and Juan Pablo Ramírez-Silva<sup>3</sup>

<sup>1</sup>Tecnológico Nacional de México, Campus Mérida, 97118, Yucatán, México

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

<sup>3</sup>Universidad Autónoma de Nayarit, Unidad Académica de Turismo; Tepic, Mexico

## Abstract

In this paper is presented a proposed solution to predict the Mexican Epidemiological Semaphore (MES) color from a set of online news. This problem was presented in Rest-Mex 2022: Recommendation System, Sentiment Analysis and Covid Semaphore Prediction for Mexican Tourist Texts. This task consists of determining the MES color through the COVID-19 news in Mexico until 8 weeks in advance. The MES system is crucial because it indicates which kind of activities are allowed to the population, for example, tourism activities. Thus, our approach is based on the Mutual Information (MI) measure. In the training stage, by using the training data, our approach first clusters every word from every news by the respective class. Then, for each word in each class, we compute its MI value. In this way, the set of words (*trained* words) with their normalized MI value is used as class features. In the classification stage, when a new instance is given, each word is intersected with the *trained* words for each class, and the corresponding MI values of the intersected words are summed. The predicted class is assigned to the class with the highest sum value. The final ranking value on the testing data was 0.175716016. We think that the obtained results are because the data has many noise words (tokens), and our approach does not deal with that issue.

## Keywords

Mutual Information, Mexican Epidemiological Semaphore, COVID, Mexican tourist texts.

## 1. Introduction


In the *Travel & Tourism Competitiveness Report (TTCR)* 2019 edition [1], the World Economic Forum published that growth in the Travel & Tourism (T&T) sector was achieving new records [2]. According to the World Tourism Organization (UNWTO), international tourist arrivals worldwide reached 1.4 billion in 2018, two years ahead of predictions. Though the TTCR findings warned of a potential tipping point at which the endless pursuit of growth and competitiveness in the sector might serve to undermine the very assets on which it is built and depends. Two years later, the T&T sector looks very different. Demand in this sector was one of the hardest hit by the COVID-19 pandemic, leaving not only companies but also tourism-driven national economies severely affected by shutdowns, travel restrictions and the disappearance of international travel.


---

*IberLEF 2022, September 2022, A Coruña, Spain*

✉ aranda@cicese.edu.mx (R. Aranda); diazpacheco@cicese.edu.mx (A. Diaz-Pacheco)

ORCID 0000-0001-8269-3944 (R. Aranda); 0000-0002-5978-0377 (A. Diaz-Pacheco)

 © 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)

Fortunately, there are positive signals, and analysis shows that recovery has started, albeit not at the same pace across the globe or in the same market segments. New factors, such as the war in Ukraine, are also adding to the complexity of this uneven recovery [3]. Thus, the sector and its customers have probably changed permanently. Travellers have become more discerning, not least about the health and hygiene conditions in potential destinations; they are also cautious about the impact of future COVID variants or challenges in the form of government policies, border closures and travel disruptions. Moreover, the halt in international travel gave some travellers, both leisure and business, a pause to consider the impact of their choices on climate and environment. Governments and T&T businesses have had to reassess where they invest, how they mitigate risk and an increase in the volatility of demand, and how they respond to the changing expectations of their customers [4].

Additional to the COVID-19 pandemic impacts, in the last decade tourism has also been influenced by numerous technological advances and tools such as digitization, information and communication technology, machine learning, robotics, and artificial intelligence (AI) [5, 6, 7, 8]. Thus, most of international travelers plan their trips by digital means, and a big part of their decisions rely on the online information [9]. Currently, one source of online information are the news.

One task of the *Rest-Mex 2022: Recommendation System, Sentiment Analysis and Covid Semaphore Prediction for Mexican Tourist Texts*[10] is the prediction of the epidemiological semaphore color (red, orange, yellow or green) from a set of online news. To be able to summarize large amounts of online news (or any type of written information), it is essential to use algorithms from the Artificial Intelligence field, specifically the area of the Natural Language Processing (NLP) to achieve human-like processing capabilities of the language for diverse scopes [11, 12]. NLP intersects artificial intelligence and linguistics [13] and covers a wide range of methods to analyze and represent naturally occurring text at one or more linguistic examination levels, for example see [14, 15, 16, 17]. Thus, in this work, we propose a method to predict the epidemiological semaphore color based on the Mutual Information measure [18].

This work is organized as follows: Section 2 describes the task to solve; Section 3 shows in details the proposal followed in this work; In section 4 the results are presented; and finally, section 5 presents the conclusions and limitations of our proposal.

## 2. Task Description: Epidemiological semaphore prediction

This task consists of determining the Mexican Epidemiological Semaphore System through the COVID-19 news in Mexico. The epidemiological semaphore system indicates which kind activities are allowed to the population; for example tourism activities. It is a system of four ordered colors:

- red: Only essential economic activities are allowed (tourism is not included), and people are allowed to go for a walk around their homes during the day.
- orange: In addition to essential economic activities, companies in non-essential economic activities are allowed to work with 30% of the staff for their operation, always taking into account the forms of maximum care for people with the highest risk of presenting

- a severe COVID-19 illness. Open public spaces (parks and green areas) with a reduced capacity (number of people) are opened.
- yellow: All work activities are allowed, taking care of people with a higher risk of presenting COVID-19. Open public spaces are open regularly, and closed public spaces can be opened with reduced capacity. These activities must be carried out with basic prevention measures and utmost care for people with a higher risk of presenting COVID-19.
- green: All activities are allowed.

The color of the semaphore changes weekly and is independent in each state of the Mexican Republic. This color is determined by the government taking into account various factors such as the capacity of hospitals, the contagion and death curve, and available ventilators, among others. Coincidentally, these factors are shared through the news of each state. If it is possible to predict the color of the epidemiological semaphore in advance, this will allow sellers and tourism service providers to take preventive measures. This task aims to generate classification models that help determine the color of the epidemiological semaphore system and thus be better prepared for the different changes in the evolution of the pandemic:

- *"Given the news related to COVID of a Mexican region (state), the goal is to determine the semaphore color of the weeks 0, 2, 4 and 8 in the future".*

For example:

- News: "Coronavirus en México: ¿qué hay detrás del súbito aumento de muertes por la pandemia de covid-19? Se registra la cifra más alta de muertes por covid-19. La noticia alarmó a muchos en México: de un día a otro el registro de personas muertas en la pandemia de coronavirus se duplicó."
- Region (State): Estado de México
- Labels:
  - Week 0: Red
  - Week 2: Red
  - Week 4: Orange
  - Week 8: Yellow

## 2.1. Data set

The corpus consists of 94,540 news related to COVID-19 collected from June 1, 2020, in the 32 regions of Mexico. The news were grouped into 1,912 instances. Each instance (around 50 news) represents a week. Also, each instance is labeled with the epidemiological semaphore color (red, orange, yellow or green) to be predicted depending on the time in the future to be evaluated (currently week ( $w_0$ ); two weeks in the future ( $w_2$ ); four weeks in the future ( $w_4$ ); and eight weeks in the future ( $w_8$ )). The distribution of the epidemiological semaphore colors for each  $w_i$ 's is unequal: most of the instances are in orange color, the red color is not even half of the instances of the orange color.

---

<sup>0</sup><https://coronavirus.gob.mx/semaforo/>

The most collected information comes from 16 sites, where [www.milenio.com](http://www.milenio.com), [eluniversal.com.mx](http://eluniversal.com.mx); [www.infobae.com](http://www.infobae.com); [www.elfinanciero.com](http://www.elfinanciero.com); [www.elfinanciero.com.mx](http://www.elfinanciero.com.mx); [www.jornada.com.mx](http://www.jornada.com.mx); [www.eleconomisata.com.mx](http://www.eleconomisata.com.mx); [mexico.as.com](http://mexico.as.com); [www.liderempresarial.com](http://www.liderempresarial.com); [www.animalpolitico.com](http://www.animalpolitico.com); [www.forber.com.mx](http://www.forber.com.mx). For more details see [19].

### 3. Proposed Approach

Our proposal consists in three main stages: preprocessing, training and classification. For the preprocessing stage, we applied to the text next steps:

- Uppercase was converted to lowercase.
- Stop-words were removed.
- Punctuation marks were removed.
- The digits were replaced by the letter 'd'.
- Stemming was applied to the tokens in the texts.
- Removed tokens that appear less than 50 times in the data set.

Training and classification stages are described below.

#### 3.1. Training stage

In this stage, we use the using the training data to extract features of each class (epidemiological semaphore color). Thus, to analyze the information from the dataset, similar to [20], we propose to use the well-known Mutual Information (MI) measure.

The MI measure was applied to all training data to extract the features for each epidemiological color (red, orange, yellow and green) [21]. This measure basically computes the mutual dependence between two variables  $X$  and  $Y$  (information that  $X$  and  $Y$  share). MI is computed by the following equation:

$$MI(X, Y) = P(X, Y) \text{Log}(P(X, Y)/P(X)P(Y)), \quad (1)$$

where  $P(X, Y)$  is the joint probability between the variables  $X$  and  $Y$ . For example, if  $X$  and  $Y$  are independent, then  $X$  is not important and does not exert any influence over  $Y$  and vice versa; then MI would be close to zero. Conversely, if  $X$  is describe in terms of  $Y$  (or  $Y$  is in terms of  $X$ ), then all information conveyed by  $X$  is shared with  $Y$  [22]. In our case, MI measures the influence of a word  $X = b$  with  $b \in B = \{\text{all the words in the collections}\}$  in a class  $Y = c$  with  $c \in C = \{\text{red, orange, yellow, green}\}$ :

- If a word  $b$  appears in all classes, then it is not relevant in any way, resulting in  $MI(b, c) \approx 0$ . The intuitive idea is that such word  $b$  does not help to discriminate among different classes (epidemiological colors).
- If the word  $b$  is almost exclusive to a class  $c$ , then this word is considered valuable for  $c$ , and the expected result would be  $MI(b, c) > 0$ . The intuition is that the higher the MI score, the more representative the word is to the class (epidemiological colors)..

**Table 1**

Official results of the proposal for the different weeks in the future.

	Accuracy	Macro F-measure	Macro Recall	Macro Precision
<b>Week 0</b>	23.655914	0.18872675	0.18347572	0.2247901
<b>Week 2</b>	31.0483871	0.18157364	0.14633243	0.23969697
<b>Week 4</b>	29.7043011	0.18882698	0.16504652	0.23863053
<b>Week 8</b>	34.0053763	0.16606979	0.30045077	0.25441233

- If a word appears repeatedly in other classes but not in class  $c$ , the result would be  $MI(b, c) < 0$ . The idea is that the lower the MI score, the less useful is the word to represent the class.

MI potentially reveals representative words for each class. Thus, it is possible to detect exclusive words describing the news on each class [23]. Thus, we call to the result set of words and and MI measures for class  $c$ , *trained* feature set  $\Omega_c$ . The  $i$ -th element,  $\omega_{i,c} \in \Omega_c$  is a tuple of values,  $(\omega_{i,c}^w, \omega_{i,c}^{MI})$ , where  $\omega_{i,c}^w$  represents the  $i$ -th word and  $\omega_{i,c}^{MI}$  represents the normalized MI measure for  $\omega_{i,c}^w$ .

### 3.2. classification stage

In the classification stage, when a new instance is given, first the preprocessing steps are applied. Then the resulting set of words for the instance is called  $\Theta$ . After,  $\Theta$  is intersected with the words in set  $\Omega_c^w$  (set of *trained* words,  $\omega_{i,c}^w$ , for class  $c$ ). Then, we compute the sum of the values  $\omega_{k,c}^{MI}$  for  $k \in \Theta \cap \Omega_c^w$ . This can be represented by equation 2:

$$S_c = \sum_{k \in \Omega_c^w \cap \Theta} \omega_{k,c}^{MI} \quad (2)$$

Thus, the predicted class for a instance  $\Theta$ ,  $C(\Theta)$ , is assigned to the class with the most high similarity value  $S_c$ :

$$C(\Theta) = \arg \max_c \{S_c\} \quad (3)$$

with  $c \in C = \{\text{red, orange, yellow, green}\}$ .

## 4. Results

Table 1 shows the official results for our proposal. In this sense our approach obtained the last place in the task, with a final ranking value of 0.175716016. our proposal barely managed to overcome the baseline (0.1290361261). Although, our proposal uses a simple idea, the best accuracy was obtained for predicting 8 weeks in the feature.

## 5. Conclusions

In this work, we presented a simple solution based on the Mutual Information measure to predict the Mexican Epidemiological Semaphore Color from a set of online news. This problem

was presented in Rest-Mex 2022: Recommendation System, Sentiment Analysis and Covid Semaphore Prediction for Mexican Tourist Texts. Although, our proposal is based in a simple idea it showed potential. The most significant disadvantage of our approach was the unbalance training data set. Additionally, we note that many words without meaning (e.g. *negativosqueretar*, *metrocdmx*, etc.) have a high MI value, but these words are noise coming from the data set. Thus, our proposal could be improved by removing those words without meaning.

## References

- [1] L. U. Calderwood, M. Soshkin, The travel and tourism competitiveness report 2019, 2019.
- [2] 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>.
- [3] Travel & tourism development index 2021, rebuilding for a sustainable and resilient future, 2022.
- [4] M. Á. 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>.
- [5] R. T. Qiu, J. Park, S. Li, H. Song, Social costs of tourism during the covid-19 pandemic, *Annals of Tourism Research* 84 (2020) 102994. URL: <https://www.sciencedirect.com/science/article/pii/S0160738320301389>. doi:<https://doi.org/10.1016/j.annals.2020.102994>.
- [6] S. Gossling, D. Scott, C. M. Hall, Pandemics, tourism and global change: a rapid assessment of covid-19, *Journal of Sustainable Tourism* 29 (2021) 1–20. URL: <https://doi.org/10.1080/09669582.2020.1758708>. doi:10.1080/09669582.2020.1758708. arXiv:<https://doi.org/10.1080/09669582.2020.1758708>.
- [7] J. Guerra-Montenegro, J. Sanchez-Medina, I. Lana, D. Sanchez-Rodriguez, I. Alonso-Gonzalez, J. Del Ser, Computational intelligence in the hospitality industry: A systematic literature review and a prospect of challenges, *Applied Soft Computing* 102 (2021) 107082. URL: <https://www.sciencedirect.com/science/article/pii/S1568494621000053>. doi:<https://doi.org/10.1016/j.asoc.2021.107082>.
- [8] D. Buhalis, Technology in tourism-from information communication technologies to eTourism and smart tourism towards ambient intelligence tourism: a perspective article, *Tourism Review* 75 (2020) 267–272. URL: <https://doi.org/10.1108/TR-06-2019-0258>. doi:10.1108/TR-06-2019-0258, publisher: Emerald Publishing Limited.
- [9] F. A. C. Calderón, M. V. V. Blanco, Impacto de internet en el sector turístico, *Revista UNIANDES Episteme* 4 (2017) 477–490.
- [10] 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).
- [11] T. Cai, A. A. Giannopoulos, S. Yu, T. Kelil, B. Ripley, K. K. Kumamaru, F. J. Rybicki,

- D. Mitsouras, Natural language processing technologies in radiology research and clinical applications, *Radiographics* 36 (2016) 176–191.
- [12] G. G. Chowdhury, Natural language processing, *Annual review of information science and technology* 37 (2003) 51–89.
- [13] P. M. Nadkarni, L. Ohno-Machado, W. W. Chapman, Natural language processing: an introduction, *Journal of the American Medical Informatics Association* 18 (2011) 544–551.
- [14] M. A. Álvarez-Carmona, A. P. López-Monroy, M. Montes-y Gómez, L. Villaseñor-Pineda, H. Jair-Escalante, Inaoe’s participation at pan’15: Author profiling task, *Working Notes Papers of the CLEF* 103 (2015).
- [15] 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.
- [16] 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>.
- [17] M. Á. Álvarez-Carmona, E. Villatoro-Tello, L. Villaseñor-Pineda, M. Montes-y Gómez, Classifying the social media author profile through a multimodal representation, in: *Intelligent Technologies: Concepts, Applications, and Future Directions*, Springer, 2022, pp. 57–81.
- [18] C. E. Shannon, A mathematical theory of communication, *The Bell System Technical Journal* 27 (1948) 379–423. doi:[10.1002/j.1538-7305.1948.tb01338.x](https://doi.org/10.1002/j.1538-7305.1948.tb01338.x).
- [19] M. Á. Álvarez-Carmona, R. Aranda, A. Y. Rodríguez-González, L. Pellegrin, C. Hugo, Classifying the mexican epidemiological semaphore colour from the covid-19 text spanish news, *Journal of Information Science* (2022). doi:<https://doi.org/10.1177/01655515221100952>.
- [20] 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>.
- [21] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, Y. Bengio, Learning deep representations by mutual information estimation and maximization, *arXiv preprint arXiv:1808.06670* (2018).
- [22] M. Ravanelli, Y. Bengio, Learning speaker representations with mutual information, *arXiv preprint arXiv:1812.00271* (2018).
- [23] M. Á. Álvarez-Carmona, M. Franco-Salvador, E. Villatoro-Tello, M. Montes-y Gómez, P. Rosso, L. Villaseñor-Pineda, Semantically-informed distance and similarity measures for paraphrase plagiarism identification, *Journal of Intelligent & Fuzzy Systems* 34 (2018) 2983–2990. doi:[10.3233/JIFS-169483](https://doi.org/10.3233/JIFS-169483), publisher: IOS Press.