A Mexico's Covid Traffic Light Color Prediction System Based on Mexican News*

Jorge Ramos-Zavaleta[1], Adrian Rodríguez[2]

[1]*Monterrey Institute of Technology and Higher Education (ITESM)*
Av. Eugenio Garza Sada 2501 Sur, Tecnológico, 64849 Monterrey, Mexico
[2]*Center for Research in Mathematics (CIMAT)*
De Jalisco s/n, Valenciana, 36023 Guanajuato, Mexico

Abstract

The COVID-19 pandemic has brought social life to a near standstill as many countries imposed very strict restrictions on movement to halt the spread of the virus. In Mexico, a traffic light system was implemented to indicate the crisis level to inform the society of the restrictions for each of the color stages of the system. The present work is an attempt to predict the traffic light color at the current week, and also perform a prediction for 2, 4, and even 8 weeks ahead by using Mexican news. For this work, we consider two approaches, one based on features extracted directly from the news and the other applying transfer learning.

Keywords

Natural Language Processing, Covid, Prediction, Mexican news

1. Introduction

COVID-19 was characterized as a pandemic by World Health Organization (WHO) on March 11, 2020, since then, it has spread to 224 countries and territories. Even though Mexico did not experience a large number of confirmed cases like the United States, it still got around 1.43 million confirmed cases accumulated by the end of 2020, creating and issue for the available hospital system around that time.

The Mexican government implemented a traffic light system to deal with the virus spread. It was a system of four colors to regulate the restrictions of all states. The four colors are: green, yellow, orange, and red, with green being low risk. The measures and restrictions were more flexible as the color changed [1]. For instance, if the system indicates an orange color, hotels could only provide service at 50% of their maximum capacity at most activities like concerts are suspended by complete.

This traffic light system [2] was generated by using different variables like hospital occupancy, mortality by every 100,000 habitants, and rate of effective reproduction of the virus. This system was executed by every state so two states could present different colors at the same time.

The precision and delay of the indicators that form the system could carry issues to the calculation of the color predominant for the week and makes predictions for weeks ahead less

IberLEF 2022, September 2022, A Coruña, Spain

precise. So an alternative system that is not dependent on these indicators can help to provide aid to establish the color of the system and predict future scenarios.

To handle this problem we propose a system based on Mexican news collected from different sources that can help to predict the color for the system not only for the current week but also for 2, 4, and even 8 weeks ahead [3, 4]. This system is based on machine learning techniques that take advantage of the text in this news to create features that help to get the correct prediction for the coming weeks when the news were published.

2. The data

The dataset 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 one of the 32 states of Mexico.

An important remark about this dataset is that even when the instances are separated, their indexes does not carry information about the chronological structure. For this reason, it is not possible to consider this structure to improve the predictions. Also, because this same issue is not feasible to match the data from neighboring states to consider a spatial structure, whereby we cannot consider a hypothesis of mobility of citizens between states. In addition to the lack of historical and spatial structure we also deal with some instances the emptyness.

A first exploration of the dataset showed an unequal distribution of classes. The imbalanced classes issue negatively impacts the model's performance, and it has to be taken into account. In figure 1, we can see that most of the instances for week 0 are set to the orange color while the red color is not even half of the instances of the orange color. The distribution for weeks 2, 4, and 8 are similar to the distribution of week 0.

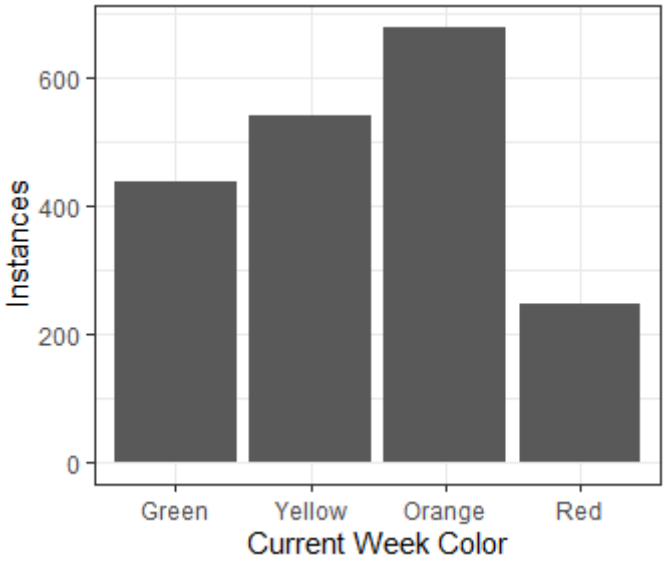

Figure 1: Distribution of colors for Week 0. The distribution indicates an unbalanced dataset that has to be considered.

3. Approaches

We consider two approaches and compare their performance against a naive baseline built by predicting the most-frequent class.

The first approach is more lexical-oriented as it considers features extracted directly from the news items. We focused on extracting some patterns in the text to determine the relevance of the news. For example, we worked with some features like the topic[5], the current virus variant, and the increase or decrease in cases, among others [6].

The second approach is a Context-Aware Machine Learning-based Approach. We took advantage of a pre-trained BERT model [7], particularly the model used in [8] was considered, and a transfer-learning model was trained directly on the news instead of the instances resulting in a larger corpus to train the model.

Both models have their own merits. From the results obtained, the transfer learning model is a better model than the first one in terms of accuracy. However, it consumed a huge computation time and more computational resources to reach those results. The first approach is quite fast to train and does not require as many computational resources, just than a regular pc. Additionally, his features are easier to explain.

3.1. Preprocessing

Preprocessing for both approaches includes the following points:

First, from the raw files, we extracted the news and place them in a tabular way, one row for each new. Then a small cleaning process was performed mainly by removing shorter lines of text that usually can be identified as an advertisement in the news. The final step for the shared preprocessing was to filter relevant news, for this task we looked inside of every news item if it contained some words that could be linked with the virus, like *COVID, SARS, hospitalized, infected, vaccine, vaccination* and some other words.

A simpler schema of this process is detailed in the first 4 steps of figure 2. The rest of the steps in the figure belong to the system built for the first approach.

3.2. First approach

For this approach, we calculate different features from the news texts based on regular expressions and topics discovery. The features that we created based on regular expressions were first used for filter the relevant news, but we believed that they would provide more information, and we considered them as another feature. For each instance, we consider the umber of *relevant words* in each news item.

Besides this feature, we created another three variables based on regular expressions. One that indicates if the texts of the news talked about infected (to have an idea of the number of cases), and another two variables indicate raising or decreasing. These two last variables were applied together with the first one to indicate if the news item was referring to an increase in infected or a decrease. With these two dummy variables, we created an indicator that shows a 1 if there was an increase (decrease) in the infected or a 0 in case there is no indication of an increase or decrease in the infected.

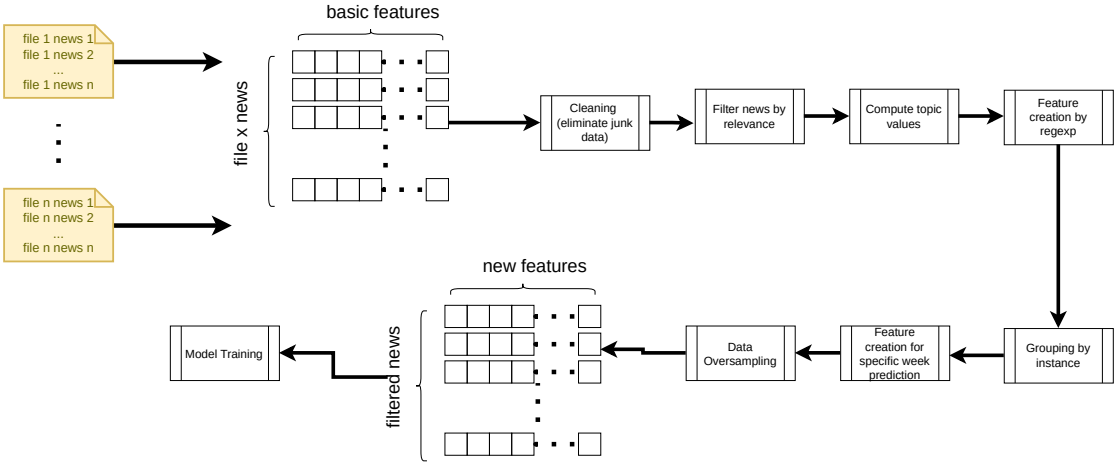

Figure 2: Preprocessing of the data and schema for the system built for the first approach.

The method to consider an increase or decrease of infected explained before was applied because the real number of infected reported in the news has different formats and contexts so an attempt to extract the number of infected would only add noise to the model. For example, in one text the number could refer to the daily infected, in another to weekly infected in another to accumulated infected to date, and also another problem to consider for this extraction is that the news item could be misplaced giving a number from a different state.

One last set of variables created using regular expression was to find mentions of the other variants of the virus. These set of variables were grouped for each instance by taking the maximum value of mentions in the news of such instance.

The topics variables created are discussed in the next subsection to give a little more detail about them and the method employed to obtain them.

3.2.1. CorEx topics

CorEx[6] is a novel approach to find possible correlations in high-dimensional data by using the concept of entropy. For this, is constructed a measure called *Total Correlation* defined by $TC(X_G) = \sum_{i \epsilon G} H(X_i) - H(X_G)$ where $H(X)$ denotes the entropy for the discrete random variable X. This expression can be written in terms of the Kullback-Leibler divergence as $TC(X_G) = D_{KL}\left(p(X_G) || \prod_{i \epsilon G} p(x_i)\right)$ so this allows to interpret the Total Correlation formula as a way to measure how different are the distributions of one word or group of words ($p(X_G)$) against another group of words ($\prod_{i \epsilon G} p(x_i)$).

A direct application of CorEx for topics discovery in texts is explained in detail in [5]. This method allows using CorEx as a semi-supervised method to find topics in the news. In our case we were looking for two specific topics, one related to Covid and another one related to economic issues, the resulting topics are

$$CovidTopic = defunciones * 0.139 + casos * 0.135 + activos * 0.119 +$$
$$+ muertes * 0.107 + confirmados * 0.103 + acumulados * 0.102 + decesos * 0.100 +$$
$$+ baja * 0.098 + camas * 0.092$$

$$EconomicsTopic = tiempo * 0.166 + presidente * 0.144 + economia * 0.120 +$$
$$+ seguridad * 0.106 + desarrollo * 0.101 + equipo * 0.095 + empleo * 0.094 +$$
$$+ gobierno * 0.088 + paises * 0.080$$

After we obtained the topics we calculate the percentage of representation of each topic for every news item and finally, we grouped each instance by taking the median values of each topic.

Figure 3 is shown that these topics provide some help to differentiate between colors. Based on their median value a larger value for the economics topic is indicative of a better color (green or yellow) while a larger value for the covid's topic indicates a worse color (orange or red).

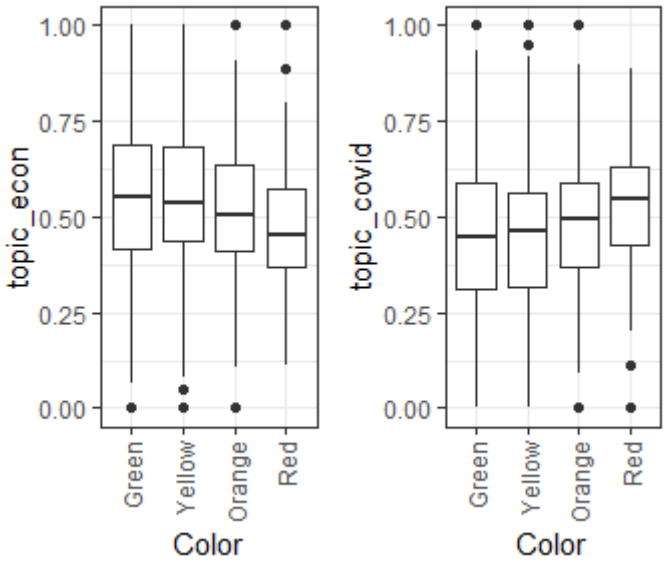

Figure 3: Distribution of colors for Week 0 for different values in the two topics.

3.2.2. R0 week variable

We needed to create unique variables for the models for each week of prediction, for this we consider a cuasi-Recursive approach. This means, the variable used for week 2 depends from the results from the variable in week 0, and the variable for week 4 depends on the values from week 2 and week 0 and so on.

For this task we used the variables created to identify variants of COVID-19 in the news to create an R0 variable for week 0. In case of a mention in the news of the variant omicron then R0 would take the value 10, 7 for delta variant and 2.5 otherwise [9]. Particularly, for week 0, another variable was included by looking for mentions of color in the news and taking the most mentioned color as our value for the variable.

For week 2, the R0 week variable was defined as the code of color obtained in the color variable developed for week 0 time the R0 for week 0. For weeks 4 and 8, we indicate different weights depending if the value of the covid topic was larger than the economics topic and also if there was a rise in the number of infected or not detoded by the variables created by using regular expressions. We place a larger value. The weights are detailed in the table 1.

Covid>Economics	Raise>Decrease	Weight
Yes	Yes	2
Yes	No	1
No	Yes	1
No	No	0.5

Table 1
Weights used to develop the R0 variables for weeks 4 and 8.

With these weights the R0 variable for week 4 is defined as $R0\_4 = Weight * R0\_2$ and R0 variable for week 8 is defined as $R0\_8 = Weight * (0.7 * R0\_4 + 0.3 * R0\_2)$.

Using all these variables we performed a bootstrap oversampling on the training data to account for the categories unbalanced and finally trained an XGBoost model.

3.3. Second Approach

For this approach, a BERT Model was considered, and using its weights to train the last layer of an MLP model using the raw news after the filter of relevance.

We developed four models each for every week that has to be predicted. For that we trained 2 epochs with a batch size of 32 and a learning rate of 1×10^{-5} with Adam optimizer and we configure the model to keep the best model objetivizing to the F1 metric.

Because the prediction is delivered for instance and the model is predicted by each news item, then we take the majority class of the news predicted in every instance as the final prediction.

An schematic way to look at this system is presented in figure 4.

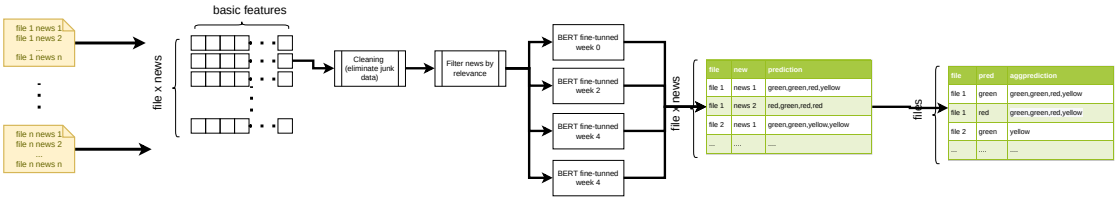

Figure 4: System developed for the second approach.

4. Results and discussion

We introduced two different approaches to model this task, each with its advantages and disadvantages. The results for both systems were up to the mark by considering a better performance than the baseline built with only the majority class. However, there is still room for improvement in both approaches.

The results of both systems and the baseline of the contest are shown in the following tables. In Table 2 the results for accuracy and F-Measure for the first 2 weeks are presented and in table 3 the results for week 4 and 8 are shown. Even when the transfer learning approach shows a better performance is important to notice that the Macro F-Measure have a minimal variation for the first approach concluding that is more consisting of their classifications.

	Week 0		Week 2	
Approach	Accuracy	Macro F-Measure	Accuracy	Macro F-Measure
Features+XGBoost	37.231	0.339	37.5	0.344
Transfer Learning	61.022	0.56	67.769	0.524
Baseline	36.696	0.134	36.962	0.135

Table 2
Accuracy and Macro F-Measure for weeks 0 and 2.

	Week 4		Week 8	
Approach	Accuracy	Macro F-Measure	Accuracy	Macro F-Measure
Features+XGBoost	37.5	0.328	37.9	0.324
Transfer Learning	60.484	0.464	62.769	0.486
Baseline	35.081	0.13	33.871	0.127

Table 3
Accuracy and Macro F-Measure for weeks 4 and 8.

It is important to note that the first approach is more interpretable and faster to train than approach two. The first approach takes around 6 minutes to train all four models (one for each week) and test with a regular computer system (i5 8th- generation processor and 8 GB of RAM). While the transfer learning (excluding the training of the original BERT Model) takes around 12 hours to train each week-model and requires a heavier computing system (Tesla P100 16 RAM and 24 GB of RAM.

Particularly for the first approach, the historical structure is needed to provide better features. Chronological information can help us to have information about the traffic light color in past weeks or construct an indicator of dates where a major spread could be presented, like holidays or some festivities. These features could be combined with the presented approach.

References

[1] 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 Science (2022). doi:https://doi.org/10.1177/01655515221100952.

[2] SSA, Lineamiento para la estimación de riesgos del semáforo por regiones covid-19, https://coronavirus.gob.mx/wp-content/uploads/2020/10/SemaforoCOVID_Metodo.pdf, 2020. Accessed: 2022-04-10.

[3] 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.

[4] 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).

[5] R. J. Gallagher, K. Reing, D. Kale, G. Ver Steeg, Anchored correlation explanation: Topic modeling with minimal domain knowledge, Transactions of the Association for Computational Linguistics 5 (2017) 529–542. doi:https://doi.org/10.1162/tacl_a_00078.

[6] G. Ver Steeg, A. Galstyan, Discovering structure in high-dimensional data through correlation explanation, Advances in Neural Information Processing Systems 27 (2014).

[7] S. Alaparthi, M. Mishra, Bert: A sentiment analysis odyssey, Journal of Marketing Analytics 9 (2021) 118–126. doi:https://doi.org/10.1057/s41270-021-00109-8.

[8] J. Arreola, L. Garcia, J. Ramos-Zavaleta, A. Rodrıguez, An embeddings based recommendation system for mexican tourism. submission to the rest-mex shared task at iberlef 2021 (2021).

[9] BBC, Covid: por qué están comparando a la variante ómicron del coronavirus con el sarampión, https://www.bbc.com/mundo/noticias-59741569, 2021. Accessed: 2022-04-08.