Determinación automática del color del semáforo Mexicano del COVID-19 a partir de las noticias

Miguel Ángel Alvarez-Carmona, Ramón Aranda

https://doi.org/10.1590/SciELOPreprints.3834

Envíado en: 2022-03-23
Postado en: 2022-03-25 (versión 1)
(AAAA-MM-DD)

Determinación automática del color del semáforo Mexicano del COVID-19 a partir de las noticias

Automatic determination of the color of the Mexican semaphore of COVID-19 from the news

Miguel Á. Álvarez-Carmona[0000−0003−4421−5575], Ramón Aranda[0000−0003−4250−2802]

Unidad de Transferencia Tecnológica Tepic
Centro de Investigación Científica y de Educación Superior de Ensenada (CICESE-UT3), Tepic 63173, México
Consejo Nacional de Ciencia y Tecnología (CONACyT), CDMX 03940 México
{malvarez, aranda}@cicese.edu.mx

Resumen Este trabajo presenta el análisis de modelos de clasificación textual para determinar automáticamente el semáforo epidemiológico regional mexicano a través de noticias de COVID. Se recolectó una base de datos con 4270 noticias referente a COVID, desde el 1 de junio de 2020 hasta el 28 de marzo de 2021. La etiqueta de cada noticia es el color del semáforo epidemiológico que el gobierno mexicano catalogó en la semana de la publicación de la noticia. Se aplicaron clasificadores como: SVM, KNN, Random Forest y Deep Learning. Los resultados muestran que es posible aprovechar la información que se publica en las noticias para determinar el color del semáforo hasta con 4 semanas de anticipación obteniendo resultados de hasta 0.74 de F-measure, el cual es un resultado competitivo tomando en cuenta el desbalance de clases de esta tarea.

Keywords: COVID-19, procesamiento de lenguaje natural, clasificación textual, semáforo epidemiológico.

Abstract This paper presents the analysis of textual classification models to automatically determine the Mexican regional epidemiological traffic light through COVID news. A database was collected with 4,270 news items referring to COVID, from June 1, 2020, to March 28, 2021. The label of each news item is the color of the epidemiological traffic light that the Mexican government cataloged in the week of publication of the news Classifiers such as SVM, KNN, Random Forest, and Deep Learning were applied. The results show that it is possible to take advantage of the information published in the news to determine the color of the traffic

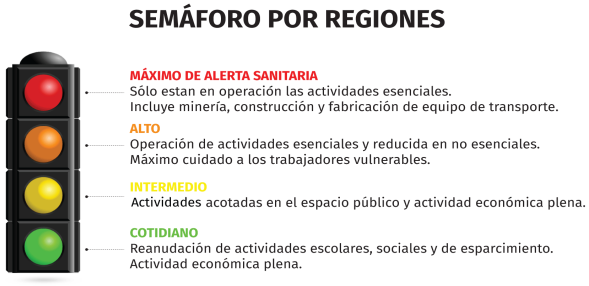


Figura 1: Semáforo epidemiológico regional en México. Fuente: https://auge-corp.com.mx/asi-funciona-el-semaforo-de-reinicio-de-actividades/

light up to 4 weeks in advance, obtaining results of up to 0.74 F-measure, which is a competitive result taking into account the imbalance of classes of this task.

Keywords: COVID-19, natural language processing, textual classification, epidemiological semaphore.

1 Introducción

En 2020, México, al igual que todos los países del mundo, se enfrentó a la pandemia generada por el COVID-19, declarada como emergencia de salud pública de importancia internacional [4]. Esta pandemia obligó a varios sectores económicos a pausar su actividad (principalmente el sector turístico), lo que provocó una pérdida económica importante en distintos niveles como en alojamiento, restaurantes, transporte, comercio, entre otros [5]. Para mitigar en mayor medida las pérdidas económicas derivadas de la pandemia, mientras se minimiza el peligro de contagio, el gobierno mexicano implementó un semáforo epidemiológico, el cual, dependiendo de su color permite ciertas actividades. De esta manera, el color del semáforo se vuelve muy importante para conocer las medidas que se deben tomar y estimar el movimiento que podría haber entre clientes y prestadores de servicios. En la Figura 1 se muestran los 4 niveles del semáforo epidemiológico. Es un sistema de cuatro colores ordenados: rojo, naranja, amarillo y verde, donde rojo es el color más restrictivo y verde el que concede mayor movimiento e interacción. El color del semáforo se actualiza de manera semanal y es independiente en cada estado del país. El color se calcula a partir de diversos factores, como la inercia de la curva epidemiológica, camas disponibles en los hospitales, ritmo de contagio entre otros [11]. Aunque el gobierno publica el color semanal unos días antes de su entrada en vigor, si se pudiera conocer con mayor antelación esta información, podrían tomarse mejores

medidas y estar mejor preparados para los cambios que involucra una variación en el semáforo.

Los datos relevantes que se consideran para calcular el color semanal del semáforo, son dados a conocer indirectamente a través de las noticias estatales, normalmente a través de sitios web de noticieros. De esta manera surge la posibilidad de tomar todos estos datos y con ayuda de Inteligencia Artificial (IA), tratar de predecir el color del semáforo epidemiológico. Esto se puede aterrizar como una tarea de clasificación, donde las instancias son las noticias que tienen que ver con el COVID–19 en algún estado de la república y la clase sería el color del semáforo. De esta manera, para el tratamiento del texto de las noticias se aplicarían métodos de Procesamiento de Lenguaje Natural (PLN).

La problemática de tratar de solucionar la predicción del semáforo, recae en la poca cantidad de información, ya que desde que hay registro del semáforo epidemiológico a la fecha, han pasado 43 semanas. Por lo que las técnicas de clasificación textual que han tenido mucho éxito en los últimos años, como los Transformers [16], no funcionarían de manera óptima en este tema. También, existe un claro desbalance de datos, ya que colores como el verde han ocurrido pocas veces, desde que inició la pandemia.

En este trabajo de investigación, se propone implementar modelos de clasificación textual a través de las noticias de COVID, planteando dar respuestas a 3 preguntas de investigación:

1. ¿Existen datos relevantes en las noticias de COVID regionales de tal manera que pueden ser aprovechados para un modelo de clasificación textual y determinar el color del semáforo epidemiológico?
2. ¿Cuáles son las mejores representaciones textuales y qué algoritmos de clasificación funcionan mejor para la tarea de clasificación del semáforo epidemiológico a través de las noticias tomando en cuenta la cantidad de datos y el desbalance de clases?
3. ¿Con cuántas semanas de anticipación se puede predecir el color del semáforo epidemiológico de tal manera que se obtenga un resultado razonable?

De todos los estados de la república mexicana, el que más cambios en el color del semáforo ha tenido es Veracruz. Este estado ha cambiado de color de una semana a otra 10 veces. Además de que es de los pocos que ha pasado por los 4 colores del semáforo e incluso ha tenido retrocesos de color en sus cambios. De esta manera, se utilizarán los datos de este estado para llevar a cabo esta investigación. Se recolectó una colección de 4270 noticias de Veracruz relacionadas con el COVID–19 (aproximadamente 99.3 noticias por semana).

El resto del documento está organizado de la siguiente manera: en la sección 2 se describen algunos trabajos de investigación sobre noticias que se han llevado a cabo para analizar el COVID-19 desde el punto de vista informativo. En la sección 3 se detalla la metodología que se llevó a cabo para determinar el color del semáforo epidemiológico de Veracruz a través de las noticias. En la sección 4 se muestran los experimentos y resultados obtenidos. Finalmente, en la sección 5 se presentan las conclusiones y el trabajo a futuro derivado de esta investigación.

2 Noticias y COVID–19

Hoy en día, el apoyo de nuevas tecnologías y aplicaciones de la IA, Internet de las Cosas (IoT), *Big Data* y *Machine Learning* contra el COVID ha sido de gran importancia debido al poder de detección, seguimiento, predicción y toma de decisiones ante los diferentes panoramas asociados a la pandemia [6, 8, 15]. Particularmente, las noticias en línea han tomado un papel importante para mantenerse informado sobre la pandemia. Y en relación a esto, han surgido recientes estudios sobre la relación entre la COVID–19 y las noticias. En [1] modelan los predictores sobre compartir noticias falsas entre los usuarios de las redes sociales, en [10], los autores crearon un sistema automatizado para verificar noticias e información sobre el COVID-19. En [12] predicen la rentabilidad de las acciones combinando noticias financieras con noticias relacionadas a salud. Los autores de [2] analizan segmentos de videos con noticias de COVID para determinar la información transmitida sobre el COVID. En [7] desarrollan nuevas técnicas para medir la detección de mentiras usando noticas falsas sobre COVID–19. Con los ejemplos mencionados, se puede ver como las noticas están jugando un papel importante en la era del COVID–19.

En este trabajo, a diferencia de los trabajos mencionados, proponemos predecir el color del semáforo epidemiológico de los estados de la república mexicana, particularmente del estado de Veracruz, mediante la implementación de modelos de clasificación textual aplicado a las noticias relacionadas con el COVID.

3 Metodología

La principal idea detrás de este trabajo es tratar de capturar las características importantes de las noticias que hablan de COVID-19 para determinar el color del semáforo epidemiológico. Para lograr esto se propone una metodología dividida en 5 etapas:

1. Recolección de datos
2. Agrupaciones de datos
3. Pre–procesamiento de datos
4. Representación de datos
5. Clasificación de los datos

En la Figura 2 se muestra una representación gráfica de la metodología propuesta. A continuación se describirán cada una de las etapas.

3.1 Recolección de datos

Para recolectar las noticias en este trabajo, se desarrolló una herramienta en lenguaje de programación *Python* y el uso de la librería *BeautifulSoup*. La herramienta generada fue configurada para utilizar el motor de búsqueda de Google Noticas de México, con noticias solo en español, usando las palabras claves "*COVID y Veracruz*". La búsqueda de las noticas se realizó por periodos semanales

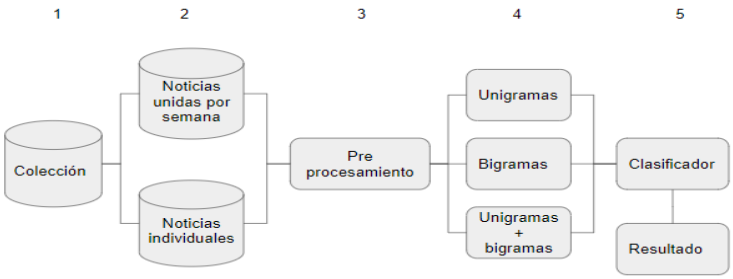


Figura 2: Metodología propuesta

a partir del 1 de junio de 2020 hasta el 28 de marzo de 2021, dando un total de 43 semanas. Se limitó a recolectar un máximo de 100 noticias por semana. Es importante mencionar que una ventaja de usar el motor de búsquedas de Google es que ordena las noticias por relevancia, lo cual nos da certeza de que la información recolectada no contenga noticias *falsas*.

Un total de 4270 noticias fueron recolectadas en todo el periodo de tiempo de 242 noticieros online. La Figura 3 muestra el número de noticias obtenidas por noticiero. Se puede observar que en 13 noticieros se concentra más del 60 % de los datos (2665 noticias), donde *elsoldeorizaba.com.mx* aporta más información con un total de 334 noticias (7.8 % del total de la información recolectada). El 37.6 % de las noticias se repartió en 229 noticieros con un promedio de 7.01 noticias y una desviación estándar de 14.03. De estos noticieros, 8 aportaron entre 100 y 50 noticias, 27 noticieros aportaron entre 50 y 10 noticias, y 193 con menos de 10 noticias (donde 114 noticieros reportaron solo 1 nota). Todos los datos recolectados se pueden descargar a través del siguiente repositorio https://github.com/moncho-arac/News-COVID-Veracruz.git

De las 43 semanas recolectadas, en 9 semanas hubo semáforo en rojo, 13 semanas el semáforo estuvo en naranja, 19 en amarillo y solo 2 en verde. Estas cifras complican el problema de clasificación por el evidente desbalance de datos.

3.2 Agrupaciones de datos

Como se mencionó en la sección 3.1 las noticias están agrupadas en semanas. Se proponen dos maneras de agrupar los datos. Primero, utilizar cada noticia individual como una instancia. Así se contarían con 4270 instancias etiquetadas. La segunda manera propuesta es juntar todas las noticias de una semana en un solo texto concatenado. De esta forma, se tienen 43 instancias etiquetadas.

3.3 Pre–procesamiento de datos

Con el fin de extraer las características más importantes de los textos, se aplicó una fase de pre–procesamiento. Las transformaciones que se llevaron a cabo son:

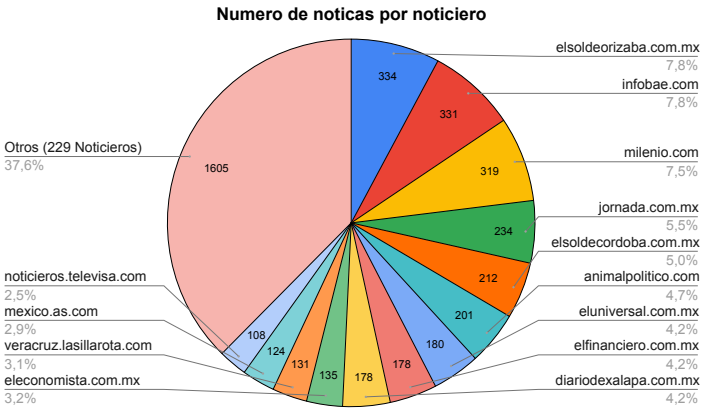


Figura 3: Cantidad de noticias de la colección divididas por noticiero.

1. Se convirtieron las mayúsculas en minúsculas.
2. Se retiraron las palabras vacías.
3. Se removieron los signos de puntuación.
4. Se reemplazaron los dígitos por la letra 'd'.
5. Para evitar que las fechas influyan, también se eliminaron las palabras de los meses del año.
6. Se aplicó *stemming* a cada una de las palabras en los textos.
7. Se removieron los tokens que aparecen menos de 15 veces[1] en toda la colección.

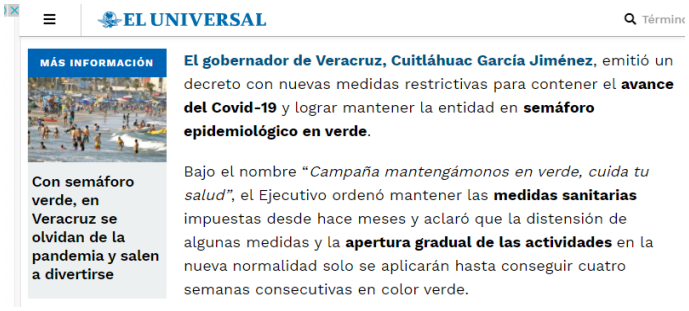
3.4 Representación de datos

Ya que lo importante del texto se centra en el contenido de la noticia y no en el estilo, se plantea trabajar a nivel de palabras. Una de las técnicas que funcionan bien para representar el contenido con pocos datos es el de N–gramas de palabras [3]. Para representar los datos se propone utilizar N = {1,2,{1,2}}. Es decir, se extraerán unigramas, bigramas y finalmente se fusionarán los dos espacios de unigramas y bigramas.

A partir del pre–procesamiento de datos descrito en la sección 3.3 y de la representación de N-gramas se propone aplicar al texto el modelo de pesado de TF-IDF (*Term frequency – Inverse document frequency*) [9]. Esta representación se basa en no solo premiar las palabras más frecuentes de un documento (en este caso, una noticia), sino que también castiga a las palabras que aparecen en muchos documentos. Así palabras que no son importantes o que aparecen a lo largo de toda la colección tendrán un peso bajo, mientras que las palabras importantes en una noticia y que no aparecen mucho en otras noticias tendrán un valor de pesado alto.

[1] Número elegido de manera empírica

(a)



(b)

Figura 4: Ejemplos de noticias durante (a) semáforo en rojo, y (b)semáforo verde. Fuente: *eluniversal.com.mx/tag/veracruz*

En la Figura 4 se presentan dos ejemplos de noticias, la primera (4a) es una noticia cuando el semáforo estaba en rojo y la segunda (4b) cuando el semáforo estaba en verde. Como es posible observar, el vocabulario que se utiliza en ambas noticias es muy distinto, lo cual permite pensar que un pesado como el de TF-IDF es ideal para esta tarea.

3.5 Clasificación de los datos

Para clasificar las noticias, se optó por aplicar una división de datos de *10–fold cross–validation* [13]. Posteriormente se aplicaron los algoritmos más populares para tareas de clasificación supervisada [14]. Estos algoritmos son:

1. *Support Vector Machine (SVM)*
2. *K–Nearest–Neighbor (KNN)*, con K = {1,3,5,7}
3. *Decision Tree (DT)*
4. *Random forests (RF)*
5. *Naive Bayes (NB)*
6. *Deep learning (DL)*

Tabla 1: Características del algoritmo de Deep Lerning aplicado.

Capas ocultas	5
Número de neuronas por capa	1000
Función de activación de las capas ocultas	Relu
Neuronas de la capa final	4
Función de activación de la capa final	Softmax
Función de pérdida	Categorical Cross Entropy
Optimizador	Adam
Épocas	50

Para el algoritmo de *Deep Learning* (DL) se hizo una implementación en *python* 3 con la paquetería de *keras* con *TensorFlow* versión 2. En la Tabla 1 se describen las características de la arquitectura del algoritmo de *Deep Learning*. Para los demás algoritmos de clasificación se utilizó la paquetería de *sklearn* de *python*. Para todos y cada uno de los algoritmos aplicados, se debe evaluar el rendimiento. Las métricas utilizadas para los resultados son *Accuracy* y *F-measure* [14].

3.6 Clasificación de semanas futuras

La razón principal de construir los modelos mencionados de clasificación es predecir el color del semáforo epidemiológico. Sin embargo, no es útil determinar dada una noticia de alguna semana, el semáforo de esa misma semana, esto debido a que ese semáforo ya se sabe. Lo interesante sería determinar los semáforos futuros. Para este trabajo proponemos inferir, dada una noticia de alguna semana S, el semáforo epidemiológico de la semana $S + m$ donde $m \in \{0, 1, 2, 3, 4\}$. Cuando $m = 0$, se estará infiriendo la semana de la noticia que se está analizando y cuando $m = 4$ se estará infiriendo el semáforo del siguiente mes (4 semanas) desde que la noticia se publicó.

4 Experimentos y resultados

En esta sección se describen los resultados obtenidos a partir de los experimentos seguidos por la metodología descrita en la sección 3.

En la Tabla 2 se muestran los mejores resultados obtenidos cuando se analizan las noticias separadas. En este caso, la colección tiene un total de 4270 instancias. La dimensión de características para unigramas fue de 6595, para bigramas fue de 12975 y la combinación de unigramas y bigramas obtiene una dimensionalidad de 19560 características. En esta tabla aparecen los mejores resultados obtenidos por cada combinación, tanto de semanas a futuro como de representación de datos. Es notable ver que en todos los resultados el mejor algoritmo de clasificación fue el de *Deep Learning* (DL). También es importante mencionar que los bigramas obtuvieron los mejores resultados respecto a los unigramas y su combinación. También, en estos resultados es posible ver que los resultados van desde 0.57 de F-measure hasta 0.59 lo que da una idea de que

Tabla 2: Mejores resultados para las noticias separadas.

m	Representación	Algoritmo	Accuracy	F-measure
0	Unigramas	DL	54.56	0.53
0	Bigramas	DL	57.23	**0.57**
0	Unigramas + Bigramas	DL	57.75	0.55
1	Unigramas	DL	53.46	0.54
1	Bigramas	DL	58.50	**0.58**
1	Unigramas + Bigramas	DL	57.35	0.57
2	Unigramas	DL	55.92	0.54
2	Bigramas	DL	58.26	**0.58**
2	Unigramas + Bigramas	DL	56.15	0.56
3	Unigramas	DL	55.48	0.54
3	Bigramas	DL	59.13	**0.57**
3	Unigramas + Bigramas	DL	59.13	0.57
4	Unigramas	DL	56.88	0.57
4	Bigramas	DL	59.92	**0.59**
4	Unigramas + Bigramas	DL	59.53	0.59

los resultados varían muy poco entre ellos. El peor resultado se obtiene cuando $m = 0$ y el mejor cuando $m = 4$.

Por otro lado, en la Tabla 3 se muestran los resultados obtenidos por el análisis cuando las noticias se unen de manera semanal en un solo documento. En esta tabla es posible ver que los resultados son más altos que los resultados obtenidos por las noticias separadas. Es importante mencionar que para este caso se analizan 43 documentos en lugar de 4270. Sin embargo parece que al unir esta información los algoritmos son capaces de capturar mejor información a pesar de la poca cantidad de instancias. También es importante notar que los algoritmos con mejores resultados son más variados, ya que aparecen otros algoritmos diferentes a *Deep Learning*. Esto se puede explicar por la misma disminución de instancias. En este caso aparecen algoritmos como DT y KNN, sin embargo, también vuelve a aparecer DL en 3 ocasiones.

En la Tabla 4 se muestra el resumen de resultados obtenidos al analizar las noticias unidas semanalmente. En esta tabla se puede ver el peor resultado se obtiene cuando $m = 4$ con 0.52 de F-measure mientras que el mejor resultado se obtiene cuando $m = 1$ con 0.74. También es importante ver que aunque los bigramas dan buena información, los unigramas son mejores cuando m=0 y la unión de unigramas y bigramas es mejor cuando $m = 4$.

4.1 Análisis de los resultados

En esta sección se pretende responder a las preguntas de investigación planteadas para este trabajo.

Pregunta 1: ¿Existen datos relevantes en las noticias de COVID regionales de tal manera que pueden ser aprovechados para un modelo de clasificación

Tabla 3: Resultados para las noticias unidas por semana.

m	Representación	Algoritmo	Accuracy	F-measure
0	Unigramas	DT	69.76	**0.54**
0	Bigramas	KNN-1	69.76	0.53
0	Unigramas + Bigramas	KNN-3	67.44	0.51
1	Unigramas	KNN-1	69.76	0.68
1	Bigramas	KNN-1	67.44	**0.74**
1	Unigramas + Bigramas	KNN-1	67.44	0.74
2	Unigramas	DT	62.79	0.47
2	Bigramas	DL	63.50	**0.63**
2	Unigramas + Bigramas	DL	62.50	0.58
3	Unigramas	DL	57.00	0.50
3	Bigramas	KNN-1	60.46	**0.63**
3	Unigramas + Bigramas	KNN-1	51.16	0.56
4	Unigramas	KNN-1	62.79	0.47
4	Bigramas	KNN-1	65.11	0.50
4	Unigramas + Bigramas	KNN-7	69.76	**0.52**

Tabla 4: Resumen de resultados para las noticias unidas.

m	Representación	Algoritmo	Accuracy	F-measure
0	Unigramas	DT	69.76	0.54
1	Bigramas	KNN-1	67.44	0.74
2	Bigramas	DL	63.50	0.63
3	Bigramas	KNN-1	60.46	0.63
4	Unigramas + Bigramas	KNN-7	69.76	0.52

textual y determinar el color del semáforo epidemiológico?

A partir de los resultados, es posible notar que los algoritmos están aprovechando información importante de las noticias para poder tomar decisiones. Para poder observar esta información se obtienen las características con mayor ganancia de información. Estas características representan los tokens que contribuyen más para clasificar entre los cuatro colores del semáforo. En la Tabla 5 se muestran las 10 características más importantes para clasificar el color del semáforo con $m = \{0, 1, 2, 3, 4\}$.

Lo primero que se observa es que cuando $m = 0$, 1 y 2, lo que más información aporta son estadísticas de contagios y fallecidos como se ve en bigramas como "confirm fallec", "ddd victim", "contagi ddd" o "baj la". Esto indica que efectivamente, esta información es compartida por las noticias y se está aprovechando. También se debe notar que en la semana 2 empiezan a aparecer temas como impacto económico o industria turística, lo cual puede dar evidencia de que el factor económico también ayuda a que el gobierno cambie el color del semáforo. Por otro lado, mientras se aleja en el futuro, las características más importantes se centran en noticias que hablen de aglomeraciones y fiestas, por ejemplo en

Tabla 5: Top 10 bigramas con ganancia de información.

Top	0	1	2	3	4
1	cambi roj	acces vacun	asciend dd	aglomer paseant	acat med
2	cas defuncion	actualiz semafor	baj la	aislamient domicili	fiest play
3	centr hospitalari	antros bar	batall covid	amarill naranj	reabr puert
4	confirm fallec	asciend dd	contagi ddd	ampli capac	reactiv activ
5	confirm millon	aument contagi	cuatr seman	buen result	realiz fiest
6	cuant fallec	confirm fallec	ddd sospech	cam diput	sospech ddd
7	ddd victim	protocol sanitari	diagnost posit	concentr person	sospech port
8	mil contagi	contagi ddd	impact econom	reactiv activ	muert contagi
9	muert acumul	transmision virus	industri turist	realiz fiest	confirm millon
10	quedat cas	ola cov	mal manej	variant sudafrican	muert acumul

bigramas como “aglomer paseant”, “realiz fiest” o “concentr person”. Esto podría indicar que es posible calcular el efecto que un evento masivo puede tener en el semáforo epidemiológico futuro a través de este tipo de técnicas. También es interesante ver que “cam diput” es un bigrama importante para calcular el color del semáforo. Finalmente, algo a notar es la importancia del bigrama “variant sudafrican” el cual se refiere a la variante sudafricana que apareció a mediados de diciembre de 2020 y que empeoró la situación de la pandemia.

Pregunta 2: ¿Cuáles son las mejores representaciones textuales y qué algoritmos de clasificación funcionan mejor para la tarea de clasificación del semáforo epidemiológico a través de las noticias tomando en cuenta la cantidad de datos y el desbalance de clases?

Cuando se trabajó con noticias separadas, las mejores combinaciones siempre fueron con bigramas de palabras. Cuando se trabajó con las noticias unidas semanalmente, cuando $m = 0$ y a 4 el mejor resultado se obtuvo con unigramas y bigramas con unigramas respectivamente. En las demás combinaciones lo mejor fue obtenido con bigramas. Esto da evidencia de que los bigramas son una mejor representación que los unigramas ya que captan de mejor manera el contenido del texto. También parece que la unión de los bigramas con los unigramas no representa una mejora importante por lo que parece que no vale la pena mezclar ambas características.

Por otro lado, cuando se comparan los algoritmos de clasificación, cuando se analizan las noticias separadas, siempre el mejor algoritmo es el de Deep Learning (DL). Esto sería posible porque existen más instancias que cuando se analizan las noticias unidas, que es cuando funcionan mejor este tipo de algoritmos. Para las noticias unidas semanalmente, el algoritmo que mejores resultados obtuvo fue KNN, apareciendo 10 veces de las 15 combinaciones posibles, como se puede ver en la Tabla 3. De estas 10, 8 veces fue utilizando k=1, y una vez k= 3 y finalmente una vez k=7. Dos combinaciones tuvieron mejores resultados con el algoritmos de *Decision Tree* (DT) y 3 combinaciones con el algoritmo de *Deep Learning*. En

Tabla 6: Promedio del ranking de los algoritmos de clasificación.

Noticias Unidas		Noticias Separadas	
Algoritmo	Promedio	Algoritmo	Promedio
DL	1.93	DL	1.00
KNN-1	2.86	DT	2.86
KNN-5	4.13	RF	3.33
KNN-3	4.53	KNN-1	3.66
KNN-7	4.80	KNN-7	6.20
NB	5.26	SVM	6.40
RF	5.40	KNN-3	6.66
DT	6.86	KNN-5	6.93
SVM	8.66	NB	7.93

la Tabla 6 se puede observar el promedio del ranking obtenido, por algoritmo, para cada una de las posibles combinaciones. Es decir, por cada experimento que se hizo, se otorgó un número entre 1 y 9 a los algoritmos de clasificación dependiendo el lugar que alcanzaron, donde 1 se le otorgaba al algoritmo con el *F-measure* más alto y 9 al más bajo. Todos estos *ranks* se sumaron y se dividieron entre 15 (5 semanas, incluyendo a la semana cero multiplicado por 3 representaciones textuales de unigramas, bigramas y su combinación) para obtener el promedio del ranking de cada algoritmo. Esto se llevó a cabo para el análisis de las noticias unidas y separadas. Sorpresivamente, para las noticias unidas, aunque el algoritmo de KNN obtuvo mejores resultados en 10 de 15 combinaciones, el algoritmo de *Deep Learning* obtuvo un mejor promedio. Sin embargo, después de este algoritmo todas las variantes de KNN vienen detrás, siendo la mejor opción utilizar $k = 1$. Es muy probable que conforme la base de datos siga creciendo, el algoritmo de *Deep Learning* empiece a tener mejores resultados que el resto de algoritmos. Por otra parte, para las noticias separadas, como ya se había dicho, el algoritmo de DL obtuvo siempre el primer lugar, sin embargo los algoritmos que le siguen en la tabla son *Decision Tree* (DT) y *Random Forest* (RF).

Pregunta 3: ¿Con cuántas semanas de anticipación se puede predecir el color del semáforo epidemiológico de tal manera que se obtenga un resultado razonable?

En la Tabla 4 Se observan los mejores resultados obtenidos para clasificar el semáforo epidemiológico. Estos resultados se obtienen analizando las noticias unidas de manera semanal. Es interesante ver que los peores resultados se obtienen cuando $m = 0$ y a 4, es decir, cuando se analiza el semáforo actual a las noticias y el semáforo del mes siguiente con un *F-measure* de 0.54 y 0.52 respectivamente. Cuando $m = 2$ y 3 el resultado es de 0.63 que, considerando el desbalance de clases y los pocos datos es un resultado competitivo. El mejor resultado se obtiene con $m = 1$, es decir, cuando se predice el semáforo de la

Tabla 7: F-measure por clase de los mejores resultados pora los valores de m.

m	Rojo	Naranja	Amarillo	Verde
0	**0.85**	0.46	0.76	0.00
1	0.70	0.58	0.71	**1.00**
2	0.80	0.41	0.67	0.66
3	0.80	0.43	0.65	0.66
4	0.71	**0.62**	**0.78**	0.00

Tabla 8: Matriz de confusión para $m = 1$ con KNN-1 y bigramas.

	Rojo	Naranja	Amarillo	Verde
Rojo	7	2	0	0
Naranja	3	9	1	0
Amarillo	1	7	11	0
Verde	0	0	0	2

semana siguiente a la noticia, en este caso se obtiene 0.74 de *F-measure*. Este resultado se obtiene con KNN-1 y su ventaja es que es capaz de capturar, en buena medida, instancias minoritarias. En la Tabla 7 se muestran los resultados de *F-measure* por cada clase para los mejores resultados en cada valor de m. En estos resultados se puede ver que los bajos resultados en $m = 0$ y 4 se deben a que no pudieron clasificar instancias de la clase verde, sin embargo obtienen buenos resultados para los otros colores. Por otro lado cuando $m = 1$ se observa que el resultado se debe a que fue capaz de clasificar correctamente las instancias verdes. En la Tabla 8 se muestra la matriz de confusión resultante, donde se puede ver que tanto con las instancias en rojo como en verde se tiene un buen rendimiento.

De esta manera es posible capturar información de los 4 colores al menos para $m = \{1, 2, 3\}$, sin embargo para $m = \{0, 4\}$, aunque no se obtienen buenos resultados para la clase verde, los demás colores tienen buen rendimiento.

5 Conclusiones y trabajo a futuro

En este trabajo de investigación se evaluaron diversas representaciones textuales y modelos de clasificación para determinar de manera automática el color del semáforo epidemiológico del estado de Veracruz. Los resultados dieron evidencia de que, de las noticias locales, es posible extraer información importante para poder alimentar clasificadores automáticos y generar modelos capaces de determinar el color del semáforo, incluso hasta con 4 semanas de antelación, siendo el mejor resultado obtenido a la semana 1 después de publicada una noticia. Los mejores resultados se obtuvieron uniendo todas las noticias de una semana en un solo documento utilizando el color del semáforo como etiqueta. También, la mejor representación para esta tarea resultó ser la de bigramas de palabras. Los mejores clasificadores fueron una arquitectura de Deep Learning y el algoritmo

de KNN. También es importante ver que mientras se vaya alimentando estos modelos con información semanal, los resultados pueden mejorar.

Como trabajo a futuro se propone extender estos modelos para los demás estados de la república mexicana. También se plantea implementar estrategias de fusión ya que muchos modelos se complementan entre sí.

6 Contribución de los autores

- **Miguel Á. Álvarez-Carmona**: Diseño de experimentos, análisis de resultados y creación y revisión del manuscrito.
- **Ramón Aranda**: Obtención de los datos y creación y revisión del manuscrito

7 Declaración de conflictos de intereses

Los autores declaran que no hay conflicto de intereses con el presente artículo.

Referencias

1. Apuke, O.D., Omar, B.: Fake news and covid-19: modelling the predictors of fake news sharing among social media users. Telematics and Informatics **56**, 101475 (2021). https://doi.org/https://doi.org/10.1016/j.tele.2020.101475, https://www.sciencedirect.com/science/article/pii/S0736585320301349
2. Basch, C.H., Hillyer, G.C., Meleo-Erwin, Z., Mohlman, J., Cosgrove, A., Quinones, N.: News coverage of the covid-19 pandemic: Missed opportunities to promote health sustaining behaviors. Infection, Disease & Health **25**(3), 205–209 (2020). https://doi.org/https://doi.org/10.1016/j.idh.2020.05.001, https://www.sciencedirect.com/science/article/pii/S2468045120300274
3. Basile, A., Dwyer, G., Medvedeva, M., Rawee, J., Haagsma, H., Nissim, M.: N-gram: New groningen author-profiling model. arXiv preprint arXiv:1707.03764 (2017)
4. Cervantes Holguín, E.: Resistir la covid-19. intersecciones en la educación de ciudad juárez, méxico. Instituto de Ciencias Sociales y Administración (2020)
5. Crick, J.M., Crick, D.: Coopetition and covid-19: Collaborative business-to-business marketing strategies in a pandemic crisis. Industrial Marketing Management **88**, 206–213 (2020)
6. Cury, R.C., Megyeri, I., Lindsey, T., Macedo, R., Batlle, J., Kim, S., Baker, B., Harris, R., Clark, R.H.: Natural language processing and machine learning for detection of respiratory illness by chest ct imaging and tracking of covid-19 pandemic in the us. Radiology: Cardiothoracic Imaging **3**(1), e200596 (2021). https://doi.org/10.1148/ryct.2021200596, https://doi.org/10.1148/ryct.2021200596
7. Álex Escolà-Gascón: New techniques to measure lie detection using covid-19 fake news and the multivariable multiaxial suggestibility inventory-2 (mmsi-2). Computers in Human Behavior Reports **3**, 100049 (2021). https://doi.org/https://doi.org/10.1016/j.chbr.2020.100049, https://www.sciencedirect.com/science/article/pii/S245195882030049X

8. Hu, Z., Ge, Q., Li, S., Jin, L., Xiong, M.: Artificial intelligence forecasting of covid-19 in china (2020)
9. Iwendi, C., Ponnan, S., Munirathinam, R., Srinivasan, K., Chang, C.Y.: An efficient and unique tf/idf algorithmic model-based data analysis for handling applications with big data streaming. Electronics **8**(11), 1331 (2019)
10. Kolluri, N.L., Murthy, D.: Coverifi: A covid-19 news verification system. Online Social Networks and Media **22**, 100123 (2021). https://doi.org/https://doi.org/10.1016/j.osnem.2021.100123, https://www.sciencedirect.com/science/article/pii/S2468696421000070
11. Salinas Zuñiga, J.I.: Estudio descriptivo de la aplicación del Semáforo Epidemiológico y su efecto en el comercio de la Ciudad de Babahoyo. B.S. thesis, BABAHOYO: UTB, 2020 (2020)
12. Salisu, A.A., Vo, X.V.: Predicting stock returns in the presence of covid-19 pandemic: The role of health news. International Review of Financial Analysis **71**, 101546 (2020). https://doi.org/https://doi.org/10.1016/j.irfa.2020.101546, https://www.sciencedirect.com/science/article/pii/S1057521920301903
13. Thabtah, F., Hammoud, S., Kamalov, F., Gonsalves, A.: Data imbalance in classification: Experimental evaluation. Information Sciences **513**, 429–441 (2020)
14. Umadevi, S., Marseline, K.J.: A survey on data mining classification algorithms. In: 2017 International Conference on Signal Processing and Communication (ICSPC). pp. 264–268. IEEE (2017)
15. Vaishya, R., Javaid, M., Khan, I.H., Haleem, A.: Artificial intelligence (ai) applications for covid-19 pandemic. Diabetes & Metabolic Syndrome: Clinical Research & Reviews **14**(4), 337–339 (2020). https://doi.org/https://doi.org/10.1016/j.dsx.2020.04.012, https://www.sciencedirect.com/science/article/pii/S1871402120300771
16. Wang, H., Czerminski, R., Jamieson, A.C.: Neural networks and deep learning. In: The Machine Age of Customer Insight. Emerald Publishing Limited (2021)

Este preprint fue presentado bajo las siguientes condiciones:

- Los autores declaran que son conscientes de que son los únicos responsables del contenido del preprint y que el depósito en SciELO Preprints no significa ningún compromiso por parte de SciELO, excepto su preservación y difusión.
- Los autores declaran que se obtuvieron los términos necesarios del consentimiento libre e informado de los participantes o pacientes en la investigación y se describen en el manuscrito, cuando corresponde.
- Los autores declaran que la preparación del manuscrito siguió las normas éticas de comunicación científica.
- Los autores declaran que los datos, las aplicaciones y otros contenidos subyacentes al manuscrito están referenciados.
- El manuscrito depositado está en formato PDF.
- Los autores declaran que la investigación que dio origen al manuscrito siguió buenas prácticas éticas y que las aprobaciones necesarias de los comités de ética de investigación, cuando corresponda, se describen en el manuscrito.
- Los autores declaran que una vez que un manuscrito es postado en el servidor SciELO Preprints, sólo puede ser retirado mediante solicitud a la Secretaría Editorial deSciELO Preprints, que publicará un aviso de retracción en su lugar.
- Los autores aceptan que el manuscrito aprobado esté disponible bajo licencia Creative Commons CC-BY.
- El autor que presenta el manuscrito declara que las contribuciones de todos los autores y la declaración de conflicto de intereses se incluyen explícitamente y en secciones específicas del manuscrito.
- Los autores declaran que el manuscrito no fue depositado y/o previamente puesto a disposición en otro servidor de preprints o publicado en una revista.
- Si el manuscrito está siendo evaluado o siendo preparando para su publicación pero aún no ha sido publicado por una revista, los autores declaran que han recibido autorización de la revista para hacer este depósito.
- El autor que envía el manuscrito declara que todos los autores del mismo están de acuerdo con el envío a SciELO Preprints.