# Las benditas redes sociales: Twitter y las elecciones presidenciales México 2018

González Franco, Nimrod
Escuela de Ingeniería y Ciencias
Instituto Tecnológico y de Estudios Superiores de Monterrey (ITESM)
Departamento de Ciencias Computacionales
Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET)
Cuernavaca, Morelos
nimrod.gonzalez@tec.mx
nimrod@cenidet.edu.mx

González Serna, Juan Gabriel
Departamento de Ciencias Computacionales
Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET)
Cuernavaca, Morelos
gabriel@cenidet.edu.mx

Castro Sánchez, Noé Alejandro
Departamento de Ciencias Computacionales
Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET)
Cuernavaca, Morelos
ncastro@cenidet.edu.mx

Astiazarán Yépiz, Guillermo José
Ingeniería en Negocios y Tecnologías de Información
Instituto Tecnológico y de Estudios Superiores de Monterrey (ITESM)
a00226365@itesm.mx

***Summary*— The analysis of comments shared via Twitter during election campaigns is a current field of research, however, the work carried out rarely focuses on elections held in Mexico, and in the case of doing so, they usually base their study on the manual analysis of tweets. This paper presents an analysis of tweets published in Mexico the days before the presidential elections of Sunday, July 1st, 2018. Using KNN, SVM, Decision Trees and Random Forest, the model identified whether or not the content of each tweet was related to the electoral process (binary classification), and if so, which of the candidates who participated in the elections was referred to (multiclass classification). In both binary and multiclass classification, the accuracy obtained with SVM was greater than the rest of the classification techniques - 0.860 and 0.8289, respectively -, while Decision Trees obtained the worst results. In addition, we found that Andrés Manuel López Obrador dominated the conversations on Twitter, even surpassing the rest of the candidates altogether.**

***Resumen*— El análisis de comentarios compartidos vía Twitter en tiempos de campañas electorales es un campo de investigación actual, sin embargo, los trabajos realizados pocas veces se centran en comicios realizados en México, y en el caso de hacerlo, normalmente basan su estudio en el análisis manual de tuits. En este trabajo se presenta un análisis de tuits compartidos en México los días previos a las elecciones presidenciales del Domingo 1 de julio de 2018. Empleando *KNN*, *SVM*, *Decision Trees* y *Random Forest*, se identificó si el contenido de cada tweet tenía o no que ver con el proceso electoral (clasificación binaria), y en caso de ser así, a cuál de los candidatos que participaron en las elecciones hacía referencia (clasificación multiclase). Tanto en la clasificación binaria como en la clasificación multiclase, la precisión obtenida con *SVM* fue mayor que con el resto de las técnicas de clasificación - 0.860 y 0.8289, respectivamente -, mientras que *Decision Trees* obtuvo los peores resultados. Además, encontramos que Andrés Manuel López Obrador dominaba ampliamente las conversaciones en Twitter, incluso superando en conjunto al resto de candidatos.**

***Palabras clave—Twitter, Técnicas de Clasificación, Elecciones Presidenciales***

## I. Introducción

Durante su primer discurso como virtual presidente electo de México, Andrés Manuel López Obrador usó la frase "… *mi gratitud a las benditas redes sociales*" para remarcar la trascendencia de estos medios durante los comicios presidenciales del 2018.

En una época en la que vivir la vida en un entorno en línea es la nueva normalidad [1], las redes sociales son un componente importante de la sociedad en el que las personas pueden compartir sus opiniones sobre candidatos y partidos políticos durante un periodo electoral. Debido a ello, en diversas investigaciones se ha trabajado en el análisis de comentarios compartidos en redes sociales para tratar de entender distintos fenómenos ocurridos durante comicios presidenciales, encontrado que Facebook y Twitter resultan ser las mejores opciones para tales análisis [2, 3, 4].

De forma particular, la literatura presenta a Twitter como una plataforma idónea para estudiar la popularidad e influencia que ejerce un líder político incluso a escala global, pues según algunos autores se usa comúnmente en la gestión de la comunicación política y gubernamental [16]. Además, dadas las características de Twitter, esta comunicación puede seguir múltiples patrones, tales como los esquemas arriba-abajo, abajo-arriba y de lado a lado [5], que, en el caso de un proceso electoral, se reflejan en los millones de tuits generados cuando candidatos y votantes hacen uso activo de dicha red social [2].

Aprovechando la gran cantidad de tuits generados en tiempos de campaña electoral, se han realizado estudios significativos enfocados a elecciones celebradas durante la última década en países como Estados Unidos, Reino Unido, Malasia, India, Italia o Pakistán, por ejemplo [6]. El caso de elecciones mexicanas, se han tenido trabajos que se centran en analizar la opinión y apoyo de los usuarios de Twitter respecto a un único candidato [7, 8], tratar de entender la otrificación asociada a la oposición política [9], estudiar la difusión de noticias sobre candidatos [10], estudiar el tipo y origen de las publicaciones sobre candidatos [11] e incluso buscar características especiales en los tuits publicados por las candidatas a gubernaturas [17].

En este artículo, presentamos un análisis de comentarios en Twitter relacionados con las elecciones presidenciales mexicanas de 2018, el cual se llevó a cabo buscando responder a la siguiente pregunta de investigación:

*¿Es posible utilizar diccionarios y técnicas de aprendizaje automático para identificar mensajes de Twitter que aluden a candidatos políticos?*

Con el fin de responder dicha cuestión, construimos un sistema que realiza la clasificación automática de tuits. La arquitectura de dicho sistema se presenta en la sección siguiente. En la sección 3, se describe el experimento realizado y finalmente, en la sección 4 se mencionan las conclusiones obtenidas.

## II. Clasificación automática de tuits

Tal como lo indica la Figura 1 y teniendo como entrada un **Conjunto de Tuits**, nuestra propuesta considera la ejecución de un proceso de cinco etapas para identificar tuits que hablan sobre determinados candidatos.

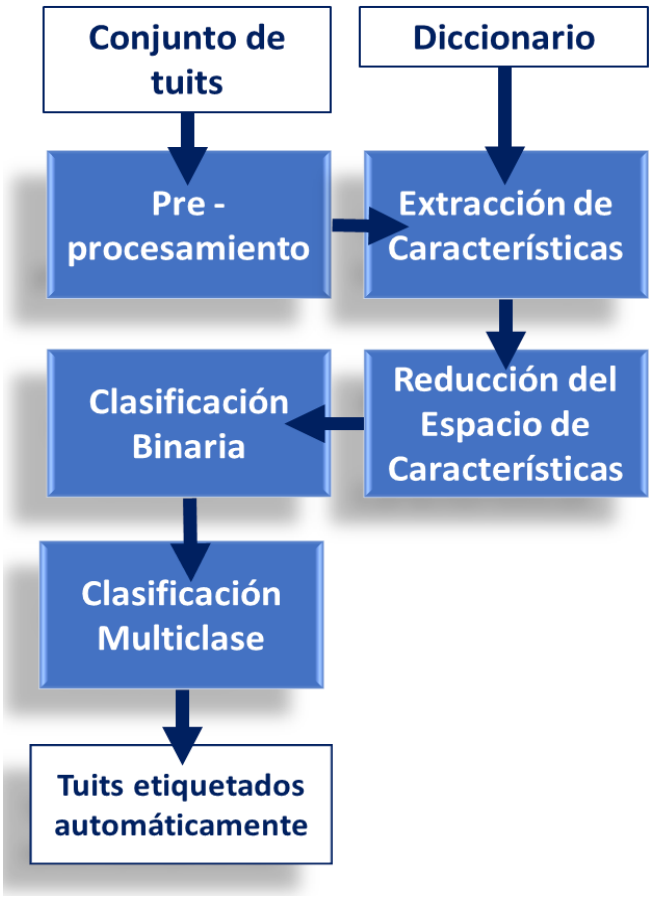


Fig. 1. Proceso para la clasificación automática de tuits

A continuación, procedemos a describir cada uno de los elementos del proceso.

**Conjunto de Tuits**: este componente corresponde a un conjunto de *n* tuits escritos en español y publicados en México que habrán de etiquetarse.

**Diccionario**: el componente Diccionario es un conjunto de *m* palabras o frases claves que nos ayudarán a decidir si un tuit tiene contenido relacionado con las elecciones y en caso de ser así, si habla de un determinado candidato.

En nuestra propuesta, consideramos 8 tipos de palabras o frases claves:

1. *Nombre del candidato*. Engloba las distintas combinaciones de nombres, apellidos e iniciales de cada candidato.
2. *Alias del candidato*. Incluye los apodos por los que se conoce a cada candidato.
3. *Cuentas oficiales*. Se refiere a las cuentas de Twitter tanto del candidato como del partido político de afiliación.
4. *Partido político*. Incluye el nombre e iniciales de un partido asociado a un candidato, ya sea por afiliación directa o por una alianza estratégica.
5. *Frases de candidato*. Engloba lemas de campaña y combinaciones de palabras comúnmente usados por el candidato.
6. *Etiquetas de candidato*. Se refiere a los hashtags asociados a cada candidato.
7. *Frases identificativas del proceso electoral*. Engloba lemas y frases comúnmente usados para referirse a las elecciones.
8. *Etiquetas del proceso electoral*. Se refiere a los hashtags asociados a las elecciones.

Las palabras y frases claves originalmente fueron definidas considerando cinco personas, pero tras la renuncia de la candidata independiente Margarita Ester Zavala Gómez del Campo, se procedió a reducir el Diccionario en función de los términos asociados a los siguientes candidatos:

- Ricardo Anaya Cortés (RAC)
- Andrés Manuel López Obrador (AMLO)
- José Antonio Meade Kuribreña (JAMK)
- Jaime Rodríguez Calderón (JRC)

**Módulo de Preprocesamiento**: en este módulo, se limpia el texto de cada elemento del **Conjunto de Tuits** para eliminar enlaces a páginas o recursos web, quitar caracteres especiales y evitar la repetición de letras (por ejemplo, convirtiendo la cadena "AMLOooooOOOooo" a la palabra "AMLO"). También se remueven los emoticones, aunque se considera analizarlos en trabajos futuros derivados de ésta investigación. Finalmente, el texto limpio queda en minúsculas y sin signos de puntación.

**Módulo de Extracción de Características***:* este módulo permite identificar características discriminatorias con base en detectar si cada uno de los *n* tuits del Conjunto de Tuits contiene alguna de las *m* palabras o frases claves del Diccionario.

**Módulo de Reducción del espacio de características:** este módulo permite establecer un subconjunto *m'* de palabras o frases claves que resultan determinantes al momento de decidir de qué trata el texto de un tuit.

**Módulo de Clasificación Binaria:** este módulo identifica a los tuits que tienen contenido relacionado con las elecciones. Para ello, se definieron dos clases ("Relacionado" y "No relacionado") que se asignan a los tuits usando individualmente las siguientes técnicas de clasificación:

- *K vecinos más cercanos (K-Nearest-Neighbor, KNN)*. Este algoritmo clasifica cada dato nuevo (en este caso, un nuevo tuit) en el grupo (clase) que corresponda, según se tengan k vecinos más cerca de un grupo o de otro. Para ello, se calcula la distancia del elemento nuevo a cada uno de los existentes, y se ordenan dichas distancias de menor a mayor para ir seleccionando el grupo al que pertenecer [12].
- *Máquinas de Soporte Vectorial (Support Vector Machines, SVMs)*. Una máquina de vectores de soporte construye un hiperplano óptimo en forma de superficie de decisión, de modo que el margen de separación entre las dos clases en los datos se amplía al máximo [13]
- *Árboles de decisión (Decision Trees, DTs)*. El algoritmo del árbol de decisión se incluye en la categoría de aprendizaje supervisado. Se pueden usar para resolver problemas de regresión y clasificación y utiliza la representación del árbol para resolver el problema en el que cada nodo hoja corresponde a una etiqueta de clase y los atributos (características) se representan en el nodo interno del árbol [14].
- *Bosques Aleatorios (Random Forest, RF)*. El bosque aleatorio, como su nombre lo indica, consiste en una gran cantidad de árboles de decisión individuales que operan como un conjunto. Cada árbol individual en el bosque aleatorio arroja una predicción de clase y la clase con más votos se convierte en la predicción de nuestro modelo [15].

Una vez que cada una de estas técnicas se ejecuta individualmente, se comparan los resultados de cada clasificador para encontrar aquél con mejor desempeño.

**Módulo de Clasificación Multiclase:** este módulo aplica de forma independiente los mismos algoritmos usados en la clasificación binaria para decidir a cuál(es) candidato(s) hace referencia un tuit relacionado con las elecciones. También compara los resultados de cada técnica de clasificación buscando cuál es la mejor de ellas.

**Tuits etiquetados automáticamente:** como salida del proceso, se tiene un conjunto de tuits a los que automáticamente se les ha asignado una y sólo una clase durante la clasificación binaria y una y solo una clase durante la clasificación multiclase.

## III. Experimentación

En la fase de experimentación, se implementó en R cada uno de los módulos presentados en la Figura 1, para luego ser probados con un **Conjunto de Tuits** conformado por 2000 publicaciones de Twitter escritas en español y publicadas en México durante junio de 2018.

Los tuits se obtuvieron originalmente como archivos json, por lo que se realizó un proceso adicional mediante el cual se extrajo el texto de los tuits para almacenarlo en hojas de cálculo.

Todos los tuits fueron etiquetados manualmente por voluntarios caracterizados por ser usuarios de Twitter, tener el español como lengua natal, mostrar interés en los comicios presidenciales y estar registrados en el padrón electoral.

A los voluntarios se les explicó que debían indicar si un tuit estaba o no relacionado con las elecciones, y, además, decidir si el tuit mencionaba alguno de los candidatos registrados considerando las combinaciones mostradas en la Tabla I. Para validar las etiquetas asignadas por los voluntarios, se aplicó una revisión de pares y en caso de controversia, se asignó una clase por consenso.

Tabla I. Combinaciones de menciones de candidatos

| *Combinación* | *Candidato(s)* |
|---|---|
| 1 | AMLO |
| 2 | RAC |
| 3 | JAMK |
| 4 | JRC |
| 5 | AMLO-RAC |
| 6 | AMLO-JAMK |
| 7 | AMLO-JRC |
| 8 | RAC-JAMK |
| 9 | RAC-JRC |
| 10 | JAMK-JRC |
| 11 | AMLO-RAC-JAMK |
| 12 | AMLO-RAC-JRC |
| 13 | AMLO-JAMK-JRC |
| 14 | AMLO-RAC-JAMK-JRC |
| 15 | Sin candidato |

De acuerdo con los voluntarios, 1523 tuits estaban “Relacionados” con las elecciones y 477 estaban “No relacionados con las elecciones”. Además, de los tuits “Relacionados” con las elecciones y ya fuese de manera única o en combinación, 311 hacían referencia a Ricardo Anaya Cortés, 951 hablaban de Andrés Manuel López Obrador, 242 mencionaban a José Antonio Meade Kuribreña y 27 estaban asociados a Jaime Rodríguez Calderón.

El **Diccionario** usado incluyó 146 palabras o frases, de las cuales 120 se asociaban directamente a uno de los cuatro candidatos registrados.

Todos los tuits sirvieron como entrada del **Módulo de Preprocesamiento**, cuyo funcionamiento se ejemplifica en la figura siguiente.

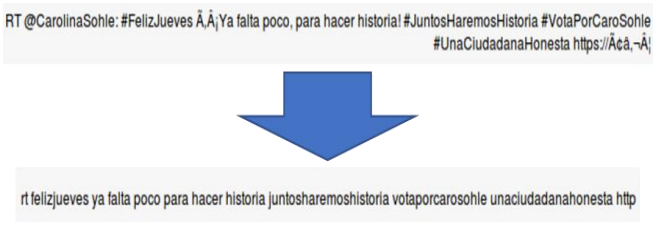


Fig. 2. Representación de la limpieza de tuits

Como se puede ver en el ejemplo anterior, a diferencia de otros trabajos en los que se elimina completamente cada cadena de enlace a un recurso web, nosotros la reducimos a la palabra reservada http. Esto fue hecho así para permitir un posterior análisis sobre la proporción de tuits relacionados con procesos electorales que incluyen en su contenido enlaces a sitios o recursos web.

En el **Módulo de Extracción de Características** se construyo una matriz de 2000 individuos y 146 características. Debido a tal dimensionalidad y para fines de demostración, sólo se incluye un fragmento de dicha matriz en la figura siguiente.

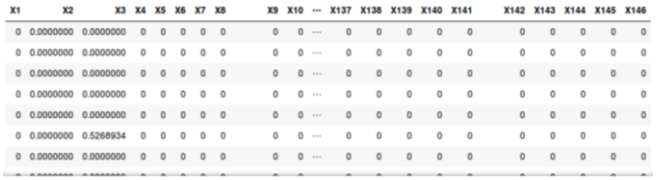

Fig. 3. Fragmento de la matriz de caracteristicas creada en la experimentación

En el **Módulo de Reducción del espacio de características** se logró identificar que 79 de los elementos del diccionario eran relevantes para el proceso de clasificación.

El **Módulo de Clasificación Binaria** fue entrenado obteniendo una muestra aleatoria equivalente al 90% del **Conjunto de Tuits** y el 10% restante sirvió para la realización de pruebas, en las cuales se pudo comparar los resultados obtenidos por los clasificadores automáticos con las etiquetas asignadas manualmente por los voluntarios. Como se muestra en la tabla siguiente, los mejores resultados se obtuvieron con SVM.

TABLA II. COMPARACIÓN ENTRE EVALUADORES PARA LA CLASIFICACIÓN BINARIA

| Clasificador | Precisión |
|---|---|
| KNN | 0.850 |
| SVM | 0.860 |
| DT | 0.750 |
| RF | 0.855 |

Por otra parte, el **Módulo de Clasificación Multiclase** se entrenó con el 90% de los tuits "Relacionados" y el otro 10% se reservó para pruebas como en el caso del **Módulo de Clasificación Binaria**. Nuevamente, SVM con kernel laplaceano obtuvo los mejores resultados, como se muestra a continuación.

TABLA III. COMPARACIÓN ENTRE EVALUADORES PARA LA CLASIFICACIÓN BINARIA

| Clasificador | Precisión |
|---|---|
| KNN | 0.7258 |
| SVM | 0.8289 |
| DT | 0.4013 |
| RF | 0.8026 |

Al analizar los tuits "Relacionados" con las elecciones encontramos que Andrés Manuel López Obrador fue referenciado en el 55.39% de ellos, dejando en segundo lugar a Ricardo Anaya Cortes con el 18.11%, en tercer lugar a José Antonio Meade Kuribreña con el 14.09% y en cuarto lugar a Jaime Heliodoro Rodríguez Calderon con el 1.57%, lo cual coincide con las posiciones alcanzadas por cada candidato en las elecciones

## IV. CONCLUSIONES

En este artículo se presenta un análisis de tuits relacionados con los comicios presidenciales mexicanos de 2018 basado en el uso de clasificadores automáticos. Como resultado, se observó que es posible utilizar diccionarios y técnicas de aprendizaje automático para identificar mensajes de Twitter que aluden a candidatos políticos en dos fases, la primera orientada en definir si un tuit está relacionado o no con unas determinadas elecciones, y la segunda, cuyo propósito es definir si el tuit habla sobre un candidato en específico. Como se ha mencionado previamente, se obtuvieron los mejores resultados con una Máquina de Vectores de Soporte implementada con un kernel Laplaciano.

## ACKNOWLEDGMENT *(Heading 5)*

Agradecemos al Instituto Tecnológico y de Estudios Superiores de Monterrey (ITESM) y al Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET) por facilitar los medios necesarios para la realización de ésta investigación.

## REFERENCIAS

[1] Allmer, T. (2019). Tim Highfield (2016) Social Media and Everyday Politics. Concept, 10(2), 2-2.

[2] Murthy, D. (2015). Twitter and elections: are tweets, predictive, reactive, or a form of buzz?. Information, Communication & Society, 18(7), 816-831.

[3] Caldarelli, G., & Chessa, A., Pammolli, F., Pompa, G., Puliga, M., Riccaboni, M., & Riotta, G.

[4] Williams, C. B., & Gulati, G. J. (2008, March). The political impact of Facebook: Evidence from the 2006 midterm elections and 2008 nomination contest. Politics & Technology Review, 1, 11–21.

[5] Yoo, J. J. S. (2019). Opinion leaders on Twitter immigration issue networks: combining agenda-setting effects and the two-step flow of information (Doctoral dissertation).

[6] Ahmed, S., Jaidka, K., & Skoric, M. M. (2016, March). Tweets and votes: A four-country comparison of volumetric and sentiment analysis approaches. In Tenth International AAAI Conference on Web and Social Media.

[7] Sandoval-Almazan, R. (2019). Using twitter in political campaigns: The case of the pri candidate in mexico. In Civic Engagement and Politics: Concepts, Methodologies, Tools, and Applications (pp. 710-726). IGI Global.

[8] Montes de Oca López, J. C., & Sandoval Almazan, R. (2019). Estudio del uso de las Redes Sociales en las Candidaturas Independientes a Presidente de México 2018.

[9] Corona, A. Mecanismos de otrificación entre la oposición política en Twitter durante las elecciones estatales de 2017 en México Othering mechanisms among political opposition on Twitter during the 2017 Mexico state elections.

[10] Glowacki, M., Narayanan, V., Maynard, S., Hirsch, G., Kollanyi, B., Neudert, L. M., & Barash, V. (2018). News and political information consumption in Mexico: Mapping the 2018 Mexican presidential election on Twitter and Facebook. The Computational Propaganda Project.

[11] Corona, A., & Muñoz, B. A. (2018). Twitter y la organización partidista local durante la elección estatal de Coahuila, 2017. Question.

[12] Sergio Ruiz. (2017). El algoritmo K-NN y su importancia en el modelado de datos. 2019, de Analítica Web Sitio web: https://www.analiticaweb.es/algoritmo-knn-modelado-datos/

[13] Staff. (2019). Algortimos de Machine Learning para clasificación (SVM). 2019, de MathWorks Sitio web: https://la.mathworks.com/discovery/support-vector-machine.html

[14] Staff. (2019). Decision Tree Introduction with example. 2019, de GeeksforGeeks Sitio web: https://www.geeksforgeeks.org/decision-tree-introduction-example/

[15] Toni Yiu. (2019). Understanding Random Forest. 2019, de Towards Data Science Sitio web: https://towardsdatascience.com/understanding-random-forest-58381e0602d2

[16] Moncayo, N. A. A., Osorio, A. V. E., & Paredes, M. L. (2018). La política en Twitter. Un estudio comparativo de las estrategias discursivas de los candidatos finalistas a la Presidencia de Ecuador en 2017. En: adComunica. Revista Científica de Estrategias, Tendencias e Innovación en Comunicación, nº16. Castellón: Asociación para el Desarrollo de la Comunicación adComunica y Universitat Jaume I, 25-44. DOI: http://dx.doi.org/10.6035/2174-0992.2018.16.3

[17] Marañón Lazcano, F., González, M., María, C., & Saldierna Salas, A. (2018). La mujer política en Twitter: análisis de los mensajes emitidos por las candidatas a gubernaturas en México. adComunica. Revista Científica de Estrategias, Tendencias e Innovación en Comunicación, nº16. Castellón: Asociación para el Desarrollo de la Comunicación adComunica y Universitat Jaume I, 71-92. DOI: http://dx.doi.org/10.6035/2174-0992.2018.16.5