Estado de ánimo de los tuiteros en los Estados Unidos Mexicanos

Documento metodológico

INEGI
INSTITUTO NACIONAL
DE ESTADÍSTICA Y GEOGRAFÍA

Catalogación en la fuente INEGI:

302.97201 Instituto Nacional de Estadística y Geografía (México).
Estado de ánimo de los tuiteros en los Estados Unidos Mexicanos : documento metodológico / Instituto Nacional de Estadística y Geografía.-- México : INEGI, c2015.

7 p.

1. Redes sociales en línea - Estadísticas- Metodología 2. Twitter - Aspectos sociales - México.

Conociendo México

01 800 111 4634
www.inegi.org.mx
atencion.usuarios@inegi.org.mx

INEGI Informa **@INEGI_INFORMA**

Instituto Nacional de Estadística y Geografía

Estado de ánimo de los tuiteros en los Estados Unidos Mexicanos

Documento metodológico

INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

Documento Metodológico de la Herramienta: Estado de Ánimo de los tuiteros en México

La información proveniente de sistemas en Internet y de dispositivos electrónicos conectados a esta red, puede contribuir en la producción de información estadística y geográfica, razón por la cual Organismos Internacionales y Oficinas Nacionales de Estadística de varios países, entre ellas el INEGI de México, están incursionando en aplicaciones prácticas de Ciencia de Datos destinadas a resolver problemas de Big Data, en particular usando información proveniente de dispositivos móviles explorando la factibilidad de generar estadísticas de movilidad y turismo, búsquedas *web* relacionándolas con estadísticas laborales, sitios de comercio electrónico para estadísticas de precios, y redes sociales para confianza del consumidor, entre otras aplicaciones.

El término "Ciencia de Datos" fue definido como una nueva disciplina hace más de una década por William S. Cleveland quien escribió el artículo "*Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics*", publicado en el Volumen 69, No. 1, de la "*International Statistical Review / Revue Internationale de Statistique*" editado por el "*International Statistical Institute*" (ISI), sin embargo el concepto ya había sido utilizado a finales de los 60's por Peter Naur. En el trabajo de Cleveland se describe un plan para crear un campo de la ciencia que cubre diversas áreas técnicas entre las que se incluyen: análisis de datos, modelos estadísticos, métodos de construcción de modelos, métodos de estimación para realizar inferencia estadística, sistemas de *hardware* y *software*, algoritmos computacionales, herramientas estadísticas, etc. Todo ello con el objetivo de llevar al análisis de datos a un nivel en el que sea posible aprender de los propios datos.

El propósito de la Ciencia de Datos es hacer análisis cuantitativo, profundizar y dar sentido a los datos que pueden ser recolectados de distintas fuentes y por diversos medios y crear nuevos productos y servicios basados en versiones recicladas de los mismos datos.

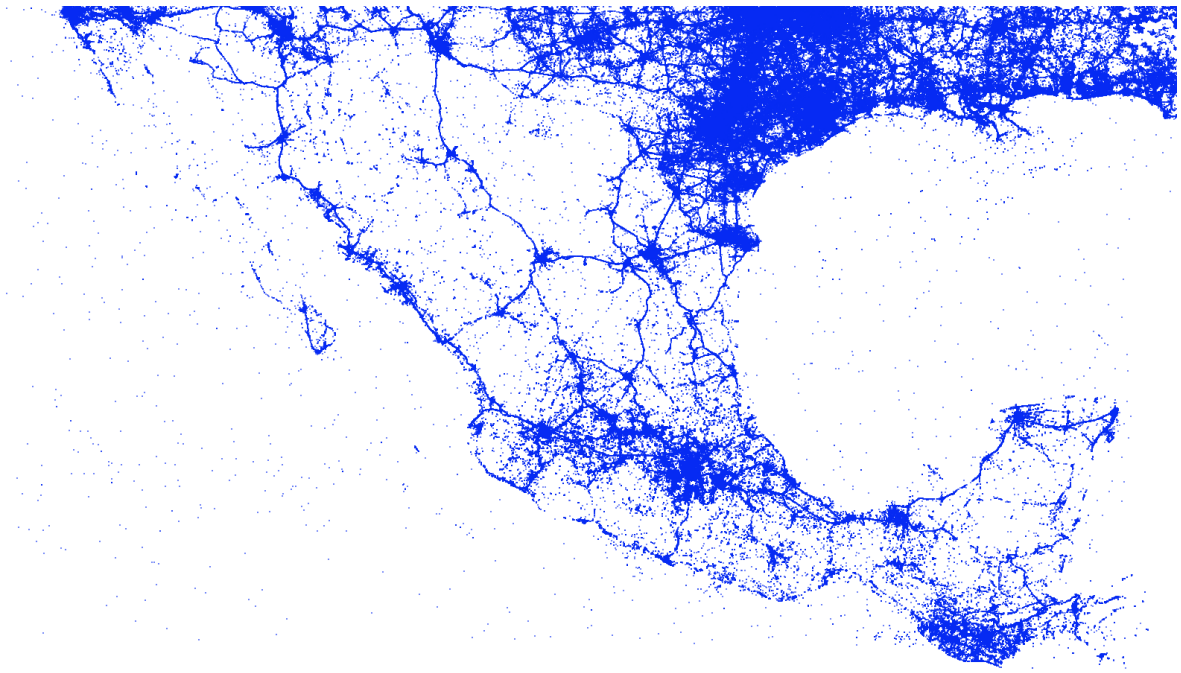
Desde el 2010 la Ciencia de Datos ha venido evolucionando aceleradamente. Actualmente es una disciplina que incorpora diferentes áreas, entre ellas: matemáticas, estadística, ingeniería de datos, reconocimiento de patrones y aprendizaje, computación avanzada, visualización, modelado de la incertidumbre, almacenamiento de datos y cómputo de alto desempeño. En el campo de la estadística, la Ciencia de Datos provee elementos para utilizar todos los datos disponibles (*Big Data*) y relevantes para emplearlos como insumos en otros procesos estadísticos.

Como parte de los estudios del INEGI en el ámbito del Bienestar Subjetivo se decidió usar *Twitter* como fuente de *Big Data* para determinar el estado de ánimo de los tuiteros en México. Se conformó un grupo de trabajo multidisciplinario con personal de INEGI, INFOTEC y Centro Geo, para incursionar en el tema de ciencia de datos, y se contó además con el apoyo de la Universidad de Pensilvania y de la Universidad de Tec Milenio.

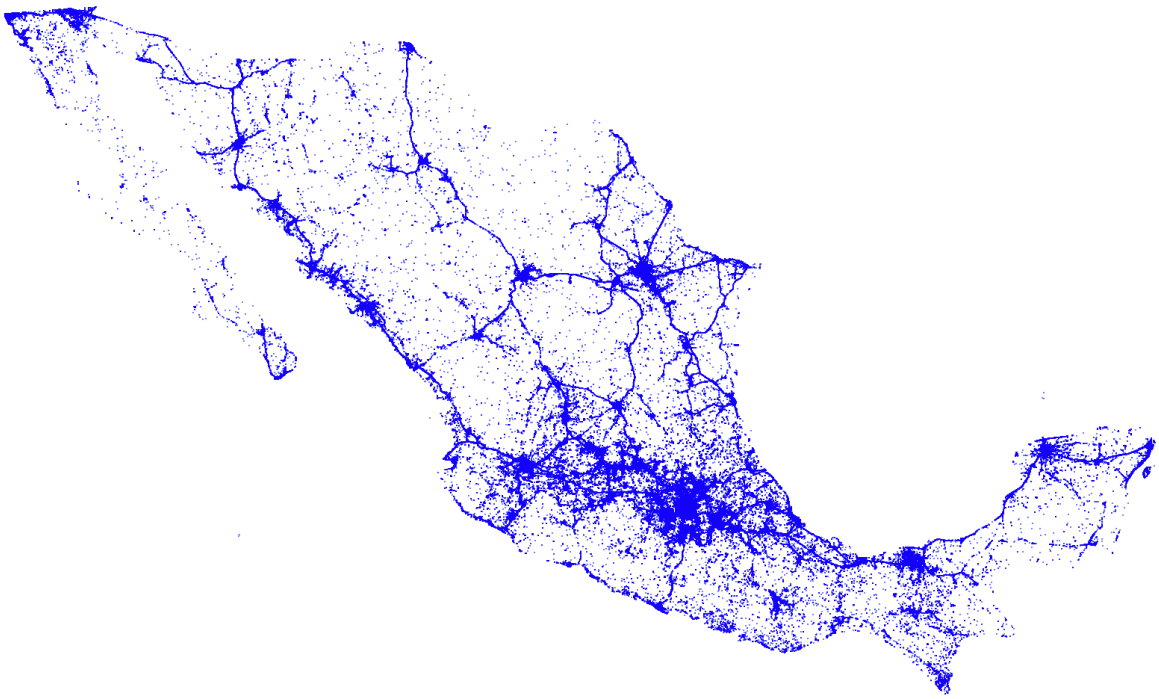
Recolección de datos provenientes de *Twitter* para fines estadísticos.

Twitter es una red social en la que los usuarios escriben textos cortos de hasta 140 caracteres que quedan visibles públicamente, es decir cualquier persona puede leer lo que se escribe en *Twitter*, no solamente aquellos que están vinculados al usuario que escribió el tuit. Adicionalmente el tuitero tiene la alternativa de georreferenciar sus tuits, etiquetando cada tuit con las coordenadas geográficas de su ubicación en el momento de publicarlo. El análisis del ánimo de los tuiteros se centró en estos tuits georreferenciados, debido a que es posible descargarlos mediante filtros geográficos sin importar el tema del que hable el tuitero, la desventaja de esto es que no todos los tuits se emiten con el atributo geográfico. El equipo de trabajo integrado por investigadores de INEGI, INFOTEC y Centro Geo, y contó además con el valioso apoyo del *Positive Psychology Center de la University of Pennsylvania* así como de la Universidad Tec Milenio y su Instituto de Ciencias de la Felicidad.

Mediante el uso de mecanismos que Twitter pone a disposición de cualquier usuario, el INEGI ha recolectado tuits públicos y georreferenciados dentro del territorio nacional, la parte sur de USA y norte de Centroamérica. Las siguientes dos gráficas muestran visualmente, gracias a su atributo de georreferenciación, todos los tuits recolectados por INEGI entre febrero de 2014 y mayo de 2015. Cada punto azul es un tuit y en conjunto dibujan la República Mexicana y sus principales vías de comunicación.



Cada punto azul es un tuit público y georreferenciado desde febrero de 2014 hasta el 15 de mayo de 2015 (125 millones de tuits).



63 millones de tuits al interior de la República Mexicana desde febrero de 2014 hasta el 15 de mayo de 2015.

Geocodificación de tuits

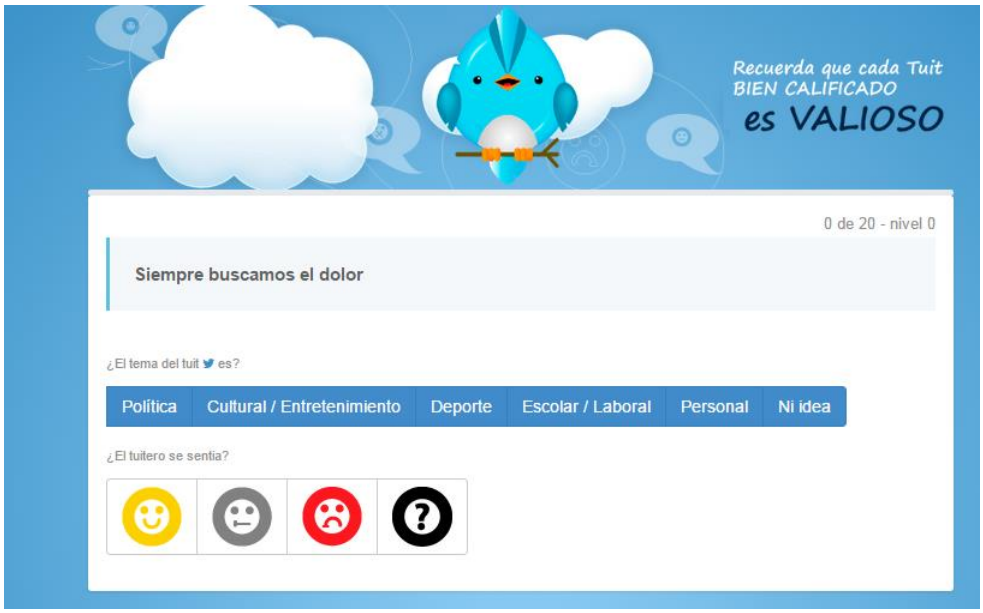
Para poder generar estadísticas a nivel estatal se llevó a cabo un análisis geográfico de cada tuit georreferenciado, y se le asignó el código geoestadístico del estado y el municipio de la República desde donde se emitió el tuit. Este primer análisis no considera la entidad habitual del tuitero, es decir, si el tuit es generado desde Nayarit no se analiza si proviene de un tuitero que habitualmente tuitea desde esa entidad o si es un turista que se encuentra ahí por un periodo corto de tiempo. El resultado del análisis geográfico permite clasificar los tuits en función de la entidad desde donde se publican.

Generación del conjunto etiquetado manualmente

Para generar la estadística del estado de ánimo de los tuiteros en México es necesario calificar cada tuit de acuerdo a la carga emotiva que identifique el estado de ánimo que tenía el tuitero cuando escribió el tuit. Si esto tuviera que hacerse manualmente sería una tarea monumental, por ello se utilizan técnicas de "*Machine Learning*".

Primero se requiere la clasificación manual de un subconjunto de tuits en la que se asigna una etiqueta de acuerdo a la carga emotiva de cada tuit. La etiqueta asignada a cada tuit se define como positiva, negativa o neutra.

Para generar este subconjunto de tuits etiquetados, se realizó una colaboración con la Universidad Tec Milenio, en la que más de 5 000 estudiantes etiquetaron manualmente miles de tuits. En este ejercicio cada tuit se presentó múltiples veces a los estudiantes con la finalidad de que un solo tuit pueda ser etiquetado varias veces y de esta manera buscar un consenso en la etiqueta.



Los estudiantes de la Universidad Tec Milenio, tuvieron acceso a una herramienta con la que etiquetaron múltiples veces 4 000 tuits previamente anonimizados

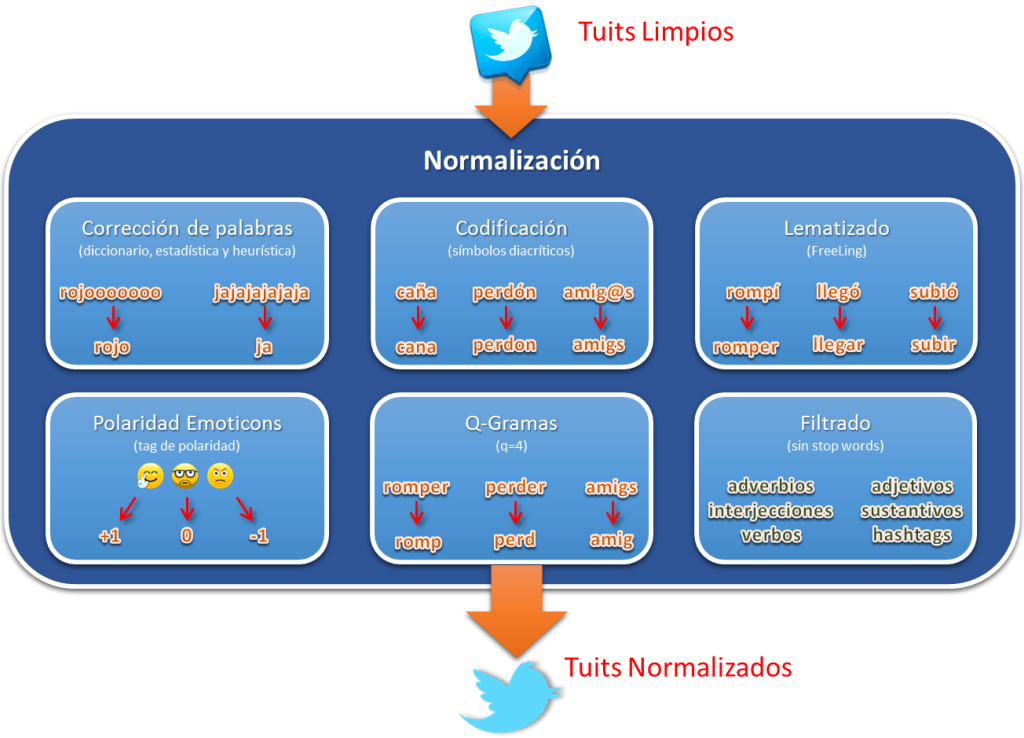
Limpieza y normalización de tuits

Posteriormente, a los tuits etiquetados se le realizó un proceso analítico de limpieza en el que se buscó disminuir la incertidumbre mediante técnicas basadas en entropía con la finalidad de disminuir el desorden en las calificaciones. Así, se identificaron y eliminaron los tuits de los etiquetadores inconsistentes, se desecharon contradicciones y repeticiones, y se identificaron aquellos tuits con mayor consenso en su etiqueta, así como también aquellos provenientes de estudiantes que mostraron mayor consistencia en su forma de asignar etiquetas.

El proceso de limpieza sirvió para eliminar redundancias e inconsistencias, dejando un conjunto menor de tuits pero con mayor calidad.

Además, en los tuits se usa argot y están escritos con incorrecciones, por lo que después de su limpieza, se usó un proceso de normalización que consiste en la ejecución de varios pasos como corrección de errores, anonimización de usuarios y de URLs, aprovechamiento de emoticones, identificación de la sintaxis de la oración y su negación. Todo ello se realizó con el fin de obtener una buena representación de la información del tuit y poder clasificarlo adecuadamente. La corrección de errores consiste en reducir las palabras/*tokens* con vocales y consonantes duplicadas inválidas a palabras del español estándar (representación de diccionario) o *tokens* válidos, p. ej., ruidoooo → ruido; jajajaaa → ja; jijijji → ja. Este proceso usa un enfoque basado en diccionarios, un modelo estadístico para letras dobles comunes y reglas heurísticas para las interjecciones comunes.

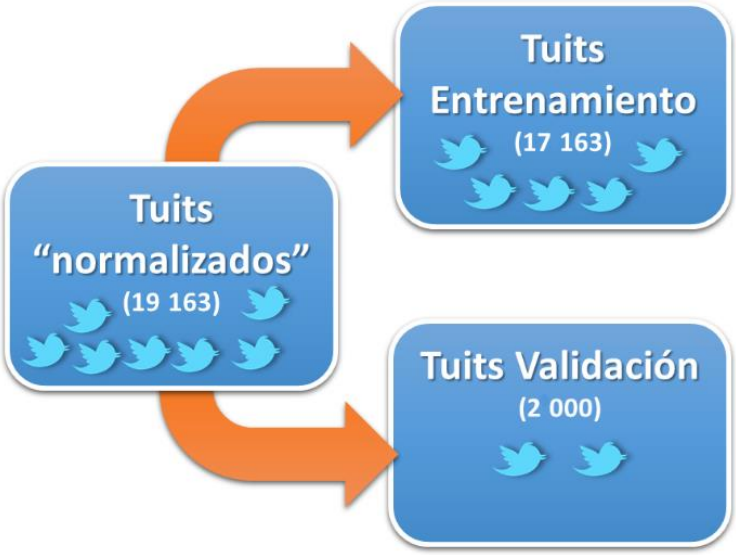
En el caso del uso de etiquetas especiales, se removieron los usuarios de *twitter (@user)* y las URLs por medio de búsquedas basadas en patrones; además, se clasificaron 512 emoticones populares en cuatro clases (Positivo, Negativo, Neutro y Ninguna), las cuales fueron reemplazadas por una etiqueta de polaridad en el texto, p. ej., emoticones positivos como :) :D se reemplazaron por la etiqueta _positivo, y emoticones negativos como :(:S se reemplazaron por _negativo. En el paso de etiquetado de partes de oración, todas las palabras fueron lematizadas, es decir, tienen la forma de una entrada de diccionario, comemos → comer; comimos → comer, etc.; se removieron las palabras que no aportan significación al contenido, dejando únicamente aquellas que sí la aportan como sustantivos, verbos, adjetivos, adverbios, las interjecciones, los *hashtags*, y las etiquetas de polaridad. En el proceso de negación, los marcadores de negación de español se unieron a la palabra de contenido más cercana, p. ej., "no seguir" → "no_seguir", "no es bueno" → "no_bueno", "sin comida" → "no_comida"; se usaron reglas heurísticas para las negaciones. Finalmente, se eliminaron todos los símbolos diacríticos y puntuación del contenido.



El proceso de normalización convierte cada tuit a una representación que facilite su clasificación automatizada.

Definición de conjuntos de entrenamiento y validación

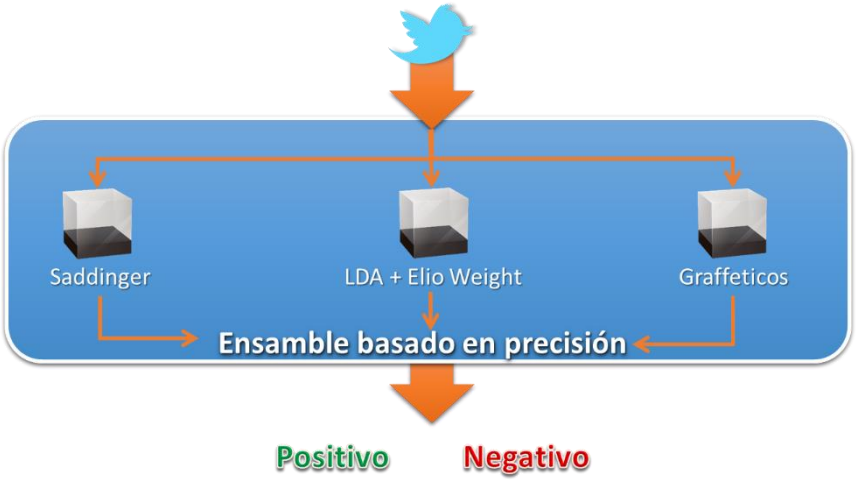
Una vez normalizados los tuits, el conjunto se partió en dos conjuntos independientes, uno con el 89% de los tuits para utilizarlo como conjunto de entrenamiento y el otro conjunto para utilizarlo como conjunto de validación, el cual sirve para verificar la calidad de la clasificación realizada automáticamente.



Desarrollo y entrenamiento de clasificadores automáticos

Se desarrollaron algoritmos innovadores de aprendizaje estadístico aplicando técnicas de inteligencia artificial por parte de Investigadores de INFOTEC y de Centro Geo. Estos algoritmos fueron integrados en un mecanismo de "Ensamble".

El ensamblado basado en la precisión de la clasificación de los algoritmos individuales, aprovecha lo mejor de cada algoritmo logrando un 70% de acierto en el etiquetado de nuevos tuits.

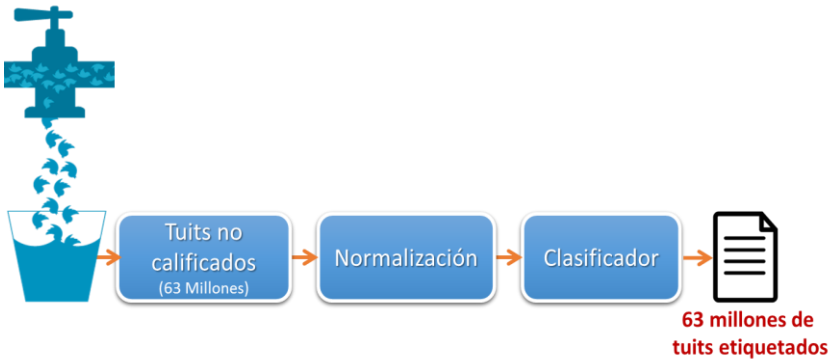


El resultado final de la fase de entrenamiento consistió en un ensamblado innovador desarrollado por el consorcio INFOTEC - Centro Geo, los nombres de cada uno de los algoritmos hacen referencia a los de sus autores: Saddinger (Dr. Eric Saditt), LDA+Elio Weight (Dr. Elio Villaseñor) y Graffeticos (Dr. Mario Graff), los tres investigadores de INFOTEC.[1]

Clasificación masiva de tuits

Utilizando el ensamblado de algoritmos ya entrenado, se prosiguió a procesar todos los tuits restantes, a los cuales se les aplicó previamente la función de normalización, dando como resultado una base de datos de tuits con un nuevo atributo que indica la carga emotiva de cada tuit.

[1] Este proyecto es resultado del trabajo de los siguientes investigadores: Dr. Elio Villaseñor (INFOTEC), Dr. Mario Graff (INFOTEC), Dr. Eric Tellez (INFOTEC), Dr. Sabino Miranda (INFOTEC), Dr. Oscar S. Siordia (Centro Geo), Dra. Daniela Moctezuma (Centro Geo), Dr. Gerardo Leyva (INEGI), Dr. Alfredo Bustos (INEGI), Dr. Juan Muñoz López (INEGI), Ing. Silvia Fraustro (INEGI), Mtro. Abel Coronado (INEGI), Ing. Ricardo Olvera (INEGI), Lic. Marco Ibarra (INEGI)



Herramienta para la visualización de la estadística del ánimo de los tuiteros en México

Finalmente, se desarrolló una herramienta de visualización que toma el resultado de la clasificación automatizada de los 63 millones de tuits para representar el ánimo de los tuiteros en México, mostrando desgloses a nivel estatal por mes.

Se calculó un índice que representa la relación de número de tuits positivos entre el número de tuits negativos y se representan tanto geográficamente como gráficamente.

La escala es relativa tomando como máximo el valor más grande de todos los índices mensuales y como mínimo el valor más pequeño de los mismos índices, utilizándose la misma escala para todos los meses del periodo con el fin de que sean comparables entre sí. La escala de colores utilizada en el mapa, indica la intensidad del sentimiento de cada entidad federativa, mientras más positivo es más verde y mientras más negativo es más rojo.

En el mapa estatal de emotividad, al hacer clic en un estado se muestra su serie de tiempo en la gráfica. También se cuenta con la opción demarcar todas las entidades federativas.

Al pasar el *mouse* en la gráfica sobre cualquier círculo de medición se muestra la cantidad de tuits positivos y negativos y el índice de ese mes.

En la parte inferior está una barra de tiempo. Al seleccionar el botón de *Play*, se pueden visualizar los cambios a través del tiempo en el mapa, a partir del mes seleccionado, o se puede seleccionar un mes en particular.