# Estado de ánimo de los tuiteros en los Estados Unidos Mexicanos


Documento metodológico
Segunda edición

INSTITUTO NACIONAL
DE ESTADÍSTICA Y GEOGRAFÍA

**Obras complementarias publicadas por el INEGI sobre el tema:**

Estado de ánimo de los tuiteros en los Estados Unidos Mexicanos. Documento metodológico (2015).

**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.-- 2da. Edición.-- México : INEGI, c2018.

13 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**

Primera edición: 2015
Segunda edición: 2017

# Documento Metodológico de la Herramienta Estado de Ánimo de los tuiteros en los Estados Unidos Mexicanos. V 2.0

## Antecedentes

En 2015 el **Instituto Nacional de Estadística y Geografía (INEGI),** con la colaboración de investigadores de INFOTEC y del Centro Geo, así como con el apoyo del Positive Psychology Center of the University of Pennsylvania (PPC-UPenn), y de la Universidad Tec Milenio (UTM), publicó la primera versión del Estado de Ánimo de los Tuiteros. Esto representó un paso hacia el uso de fuentes alternativas de Big Data para generar nuevas estadísticas experimentales. Se logró concretar una aplicación, para el ámbito de Bienestar Subjetivo, en la que se presentaba el resultado de todo el ciclo necesario en el uso de técnicas de Big Data utilizando fuentes no tradicionales como lo son las redes sociales. Este ciclo abarca desde la recolección de la fuente de datos, el análisis de éstos, su preprocesamiento, la clasificación y el procesamiento de cada dato, su almacenamiento, la generación de indicadores, hasta la representación visual para difundir dichos resultados. Cabe mencionar que durante todo el ciclo se presentan retos muy diferentes a los enfrentados en la generación de estadísticas tradicionales, como son el almacenamiento masivo y su procesamiento en paralelo, sin dejar de lado los algoritmos de *machine learning* para la clasificación automática de nueva información que se está recolectando continuamente.

Para esta segunda versión que se libera en el 2017, se sigue utilizando la misma metodología descrita en el 2015, en cuanto a la recolección, análisis, preprocesamiento, clasificación del sentimiento y generación de indicadores. La diferencia y principal mejora radica tanto en la temporalidad, que ahora es diaria, para clasificar los tuits y generar los indicadores, como en el procesamiento para distinguir los tuits de los visitantes (turistas) y de los locales. Así mismo, se generó una nueva aplicación con más funcionalidades para visualizar los resultados en la temporalidad seleccionada, el sentimiento de todos los tuiteros por entidad federativa, de los tuiteros locales o de los visitantes, así como ver la nube de #hatshtags por día o las noticias de un día en particular, para que el usuario pueda tener más elementos para tratar de entender el sentimiento de ese día.

Por lo anterior, la diferencia entre la metodología publicada en el 2015 y esta metodología del 2017 la encontrará en las dos últimas secciones de este documento, "Procesamiento para distinguir locales de visitantes" y "Herramienta para la visualización de la estadística del ánimo de los tuiteros en México".

## Introducción

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 **Instituto Nacional de Estadística y Geografía (INEGI),** 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; de búsquedas web relacionándolas con estadísticas laborales, de sitios de comercio electrónico para estadísticas de precios, y de 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,* editada por el International Statistical Institute (ISI),[1] sin embargo el concepto

[1] http://www.datascienceassn.org/sites/default/files/Data%20Science%20An%20Action%20Plan%20for%20Expanding%20the%20Technical%20Areas%20of%20the%20Field%20of%20Statistics.pdf

ya había sido utilizado a finales de los sesenta 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 el 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, así como profundizar y dar sentido a los datos que pueden ser recolectados de distintas fuentes y por diversos medios, así como crear nuevos productos y servicios basados en versiones recicladas de los mismos datos.

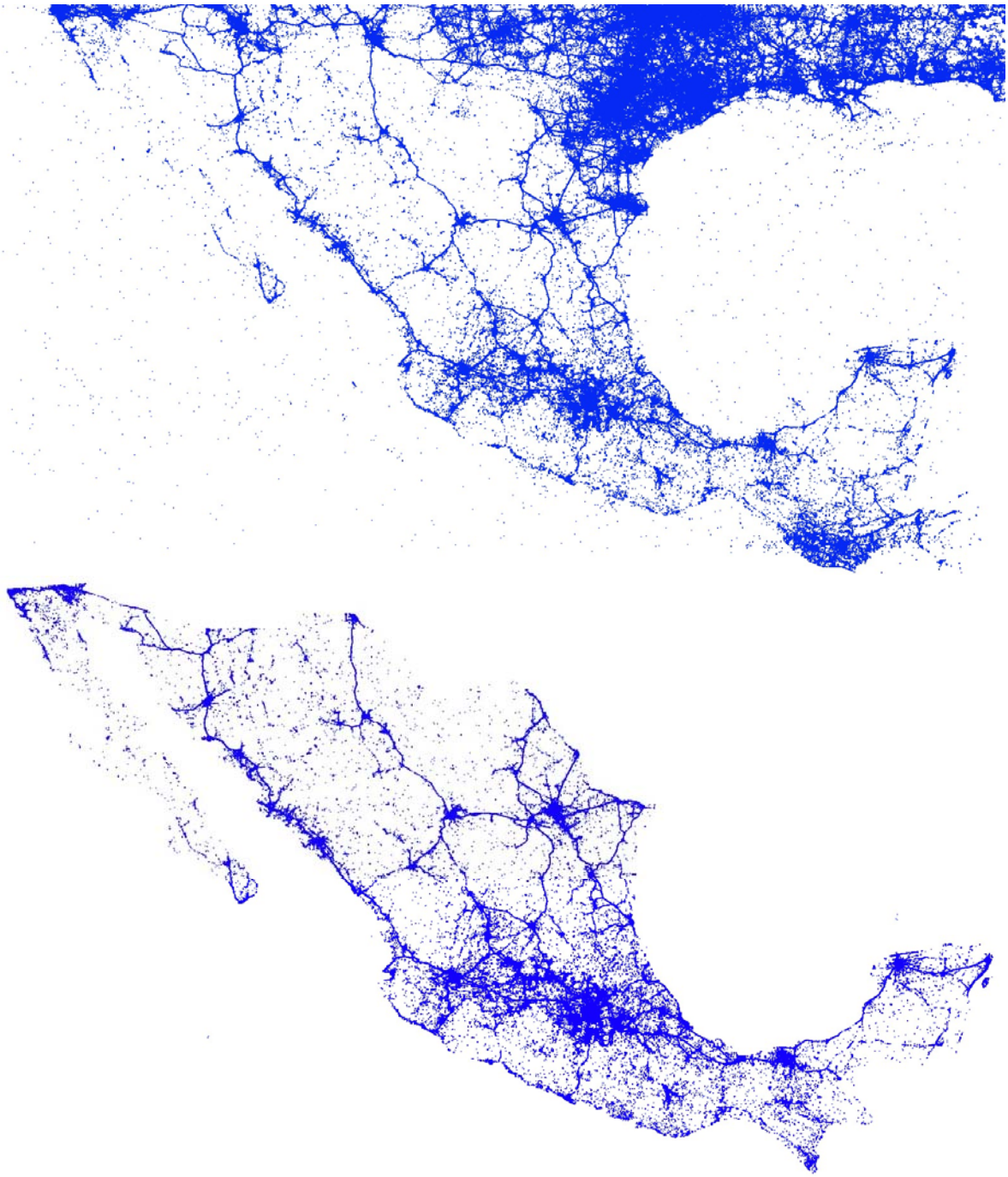
Desde el 2010 la Ciencia de Datos ha venido evolucionando de manera acelerada. Actualmente es una disciplina que incorpora diferentes áreas, entre ellas: matemáticas, estadística, ingeniería de datos, reconocimiento de patrones y aprendizaje, computación avanzada, visualización, modelado de la incertidumbre, 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.

**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.

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 los Estados Unidos de América 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 público y georreferenciado, publicado entre febrero de 2014 hasta el 15 de mayo de 2015 (125 millones de tuits) que, en conjunto delinean la República Mexicana y sus principales vías de comunicación.



63 millones de tuits en la República Mexicana desde febrero de 2014 hasta el 15 de mayo de 2015. Cabe mencionar que, después de esta fecha, se continúa de manera permanente recolectando tuits a diario.

**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 del 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, por ejemplo, no se analiza si proviene de un tuitero que habitualmente tuitea desde esa entidad o

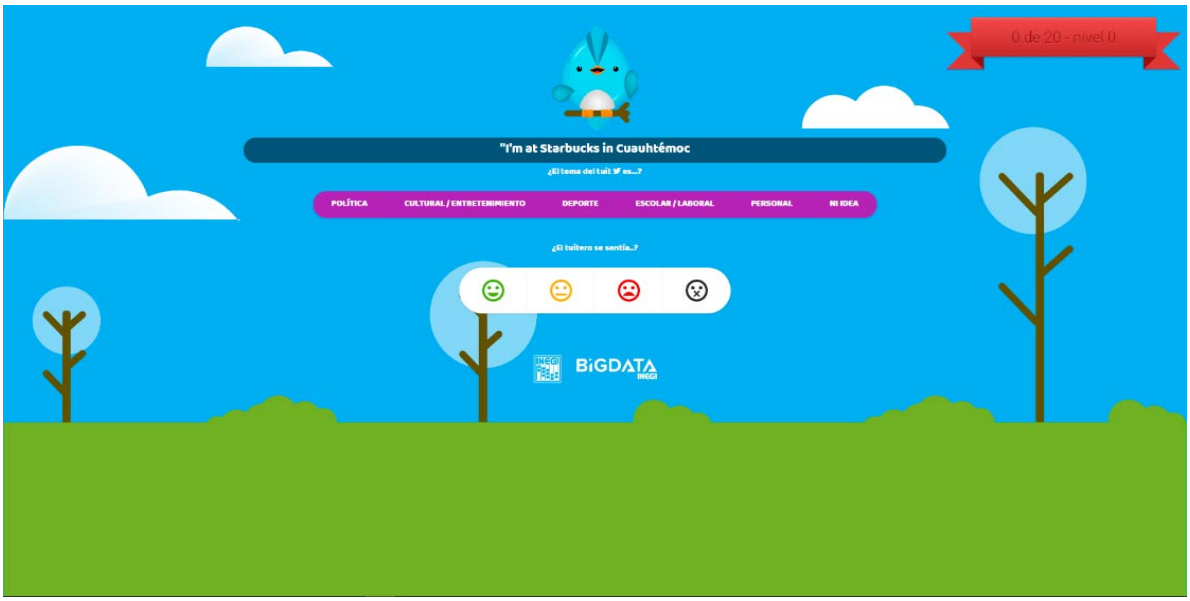
si es un turista que se encuentra ahí por un periodo corto de tiempo. El resultado del análisis geográfico permite clasificar los tuits en función de la entidad desde donde se publican.

**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 tamaño reducido 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 ocasiones a los estudiantes con la finalidad de que un solo tuit fuera etiquetado varias veces y de esta manera lograr 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.

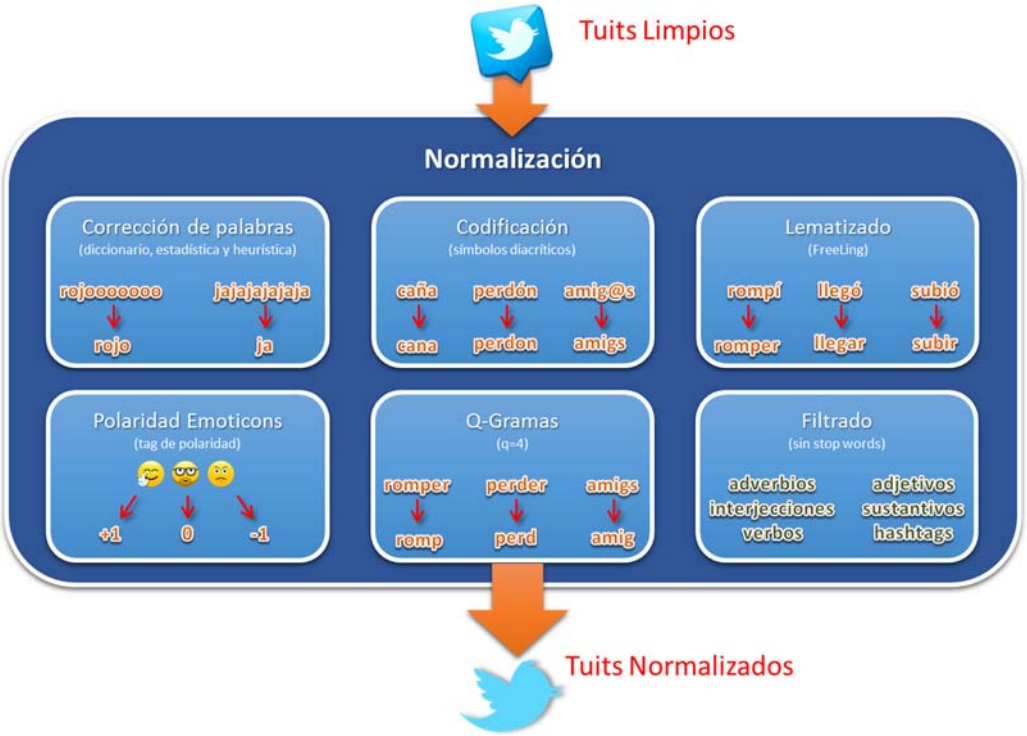
**Limpieza y normalización de tuits**

Posteriormente, a los tuits etiquetados se le realizó un proceso analítico de limpieza en el que se buscó disminuir 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 empleó un proceso de normalización que consiste en la ejecución de varios pasos, como corrección de errores, anonimización de usuarios y de URL, 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, por ejemplo: 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.

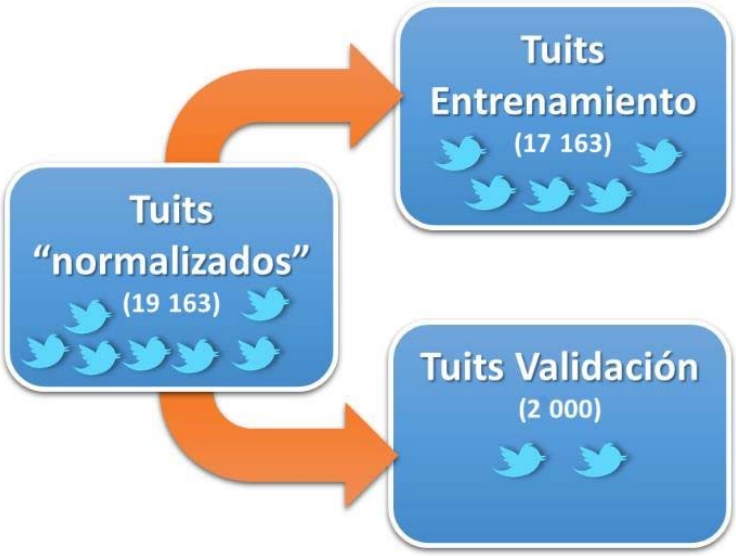
En el caso del uso de etiquetas especiales, se removieron los usuarios de twitter (@user) y las URL por medio de búsquedas basadas en patrones; además, se clasificaron 512 emoticones populares en cuatro clases (Positivo, Negativo, Neutro y Ninguna), las cuales fueron reemplazadas por una etiqueta de polaridad en el texto; por ejemplo: emoticones positivos como :) o :D se reemplazaron por la etiqueta _positivo, y emoticones negativos como :( o :S se reemplazaron por _negativo. En el paso de etiquetado de partes de oración, todas las palabras fueron lematizadas, es decir, se les dio la forma de una entrada de diccionario: comemos → comer; comimos → comer, etcétera; 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; por ejemplo: 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**

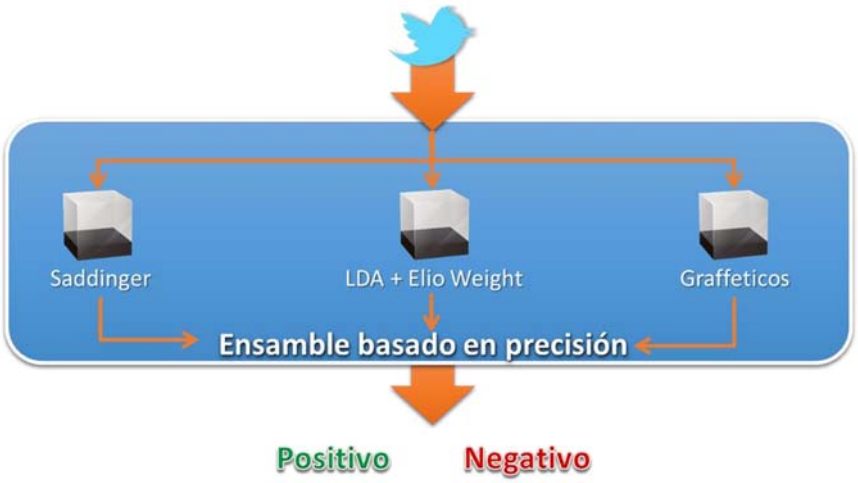
Una vez normalizados los tuits, el conjunto se partió en dos conjuntos independientes, uno con el 89% de los tuits para utilizarlo como conjunto de entrenamiento y el otro para emplearlo como conjunto de validación, el cual sirve para verificar la calidad de la clasificación realizada automáticamente.



**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 del Centro Geo. Estos algoritmos fueron integrados en un mecanismo de ensamble.

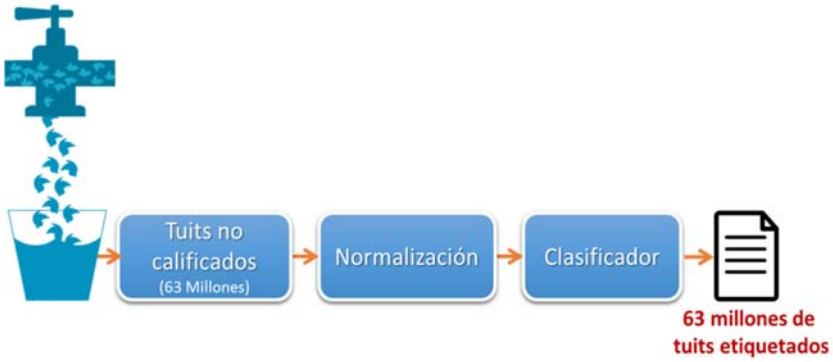
El ensamble basado en la precisión de la clasificación de los algoritmos individuales aprovecha lo mejor de cada algoritmo logrando un 80% de acierto en el etiquetado de nuevos tuits.




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.[2]

**Clasificación masiva de tuits**

Utilizando el ensamble 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.




[2] Este proyecto es resultado del trabajo de los siguientes investigadores: Dr. Elio Villaseñor (INFOTEC), Dr. Mario Graff (INFOTEC), Dr. Eric Téllez (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).

**Procesamiento para distinguir locales de visitantes**

Una de las características más destacadas en los resultados publicados en 2015 se refiere a las importantes diferencias entre las entidades federativas del país. En particular, destacan dos entidades (Nayarit y Quintana Roo) como las más optimistas. Ambas tienen en común tamaños poblacionales relativamente pequeños y destinos turísticos importantes (Nuevo Vallarta y Cancún, respectivamente). Lo anterior da lugar a la hipótesis de que el estado de ánimo de los residentes habituales es ocultado por el optimismo de los vacacionistas. Con el propósito de llevar a cabo comparaciones que no estén sesgadas por la presencia de visitantes, se hace entonces necesario distinguir sus tweets de los de los locales.

**Determinación de lugar de residencia**

**Algoritmo de local/visitante:**

Para cada usuario se toman los tuits que haya publicado en el último año. Se cuenta el lapso cubierto por una secuencia de tuits publicados todos desde la misma entidad federativa, sin interrupciones

Se hace la sumatoria del tiempo por cada entidad federativa y se asigna como lugar de residencia a la entidad federativa en donde más tiempo pasó en ese año, de acuerdo con el criterio anterior.

**Excepciones:**

Si el usuario no tiene historial de tuits (ej., turistas extranjeros) se asigna como lugar de residencia la entidad federativa en la que publicó el tuit.

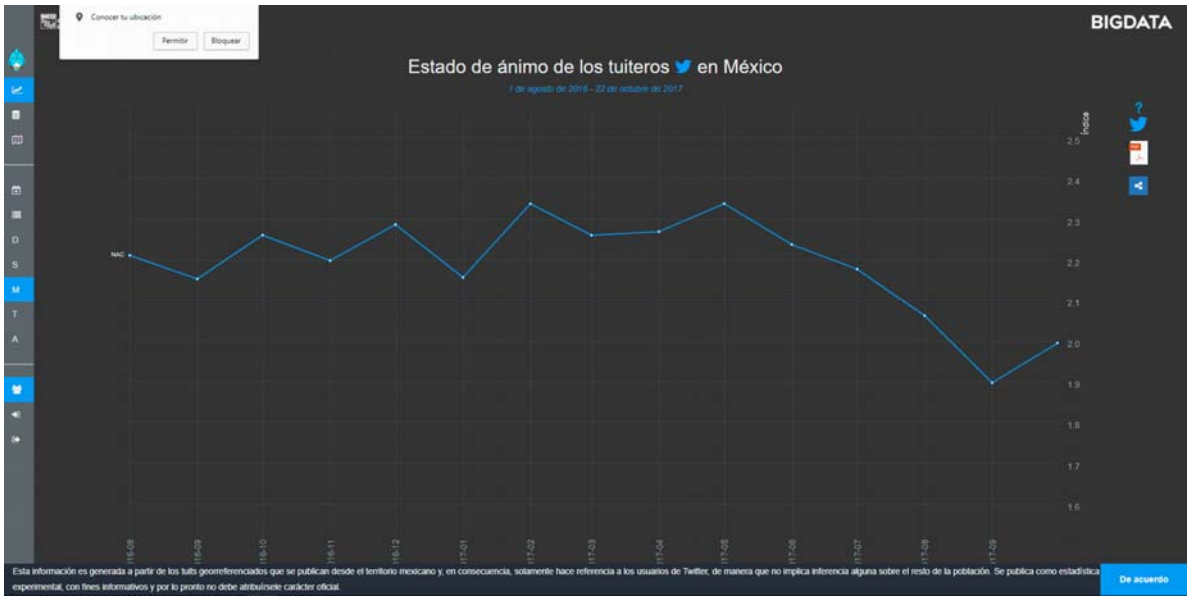
Si cada uno de los tuits contenidos en su historial tiene diferente estado, el lugar de residencia se asume que es la entidad federativa donde se creó su último tuit.

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

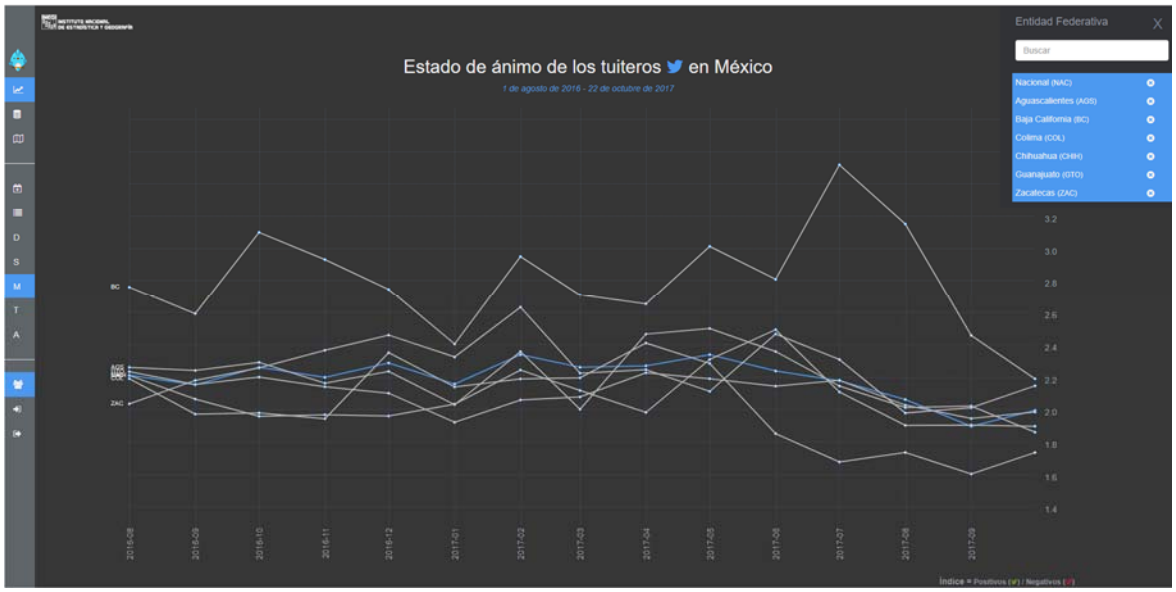
Finalmente, se desarrolló una herramienta de visualización que toma el resultado de la clasificación automatizada de todos los tuits recolectados diariamente, para representar el ánimo de los tuiteros en México, mostrando desagregaciones a nivel estatal por año, trimestre, mes, semana o día, además de por total, locales o visitantes.

Se calculó un índice que representa la relación del número de tuits positivos entre el número de tuits negativos, y sus valores se representan tanto geográfica como gráficamente.

Al inicio, la aplicación pedirá al usuario autorizar su ubicación para mostrar el índice de la entidad federativa en la cual se encuentra, y mostrará su entidad y el promedio nacional.

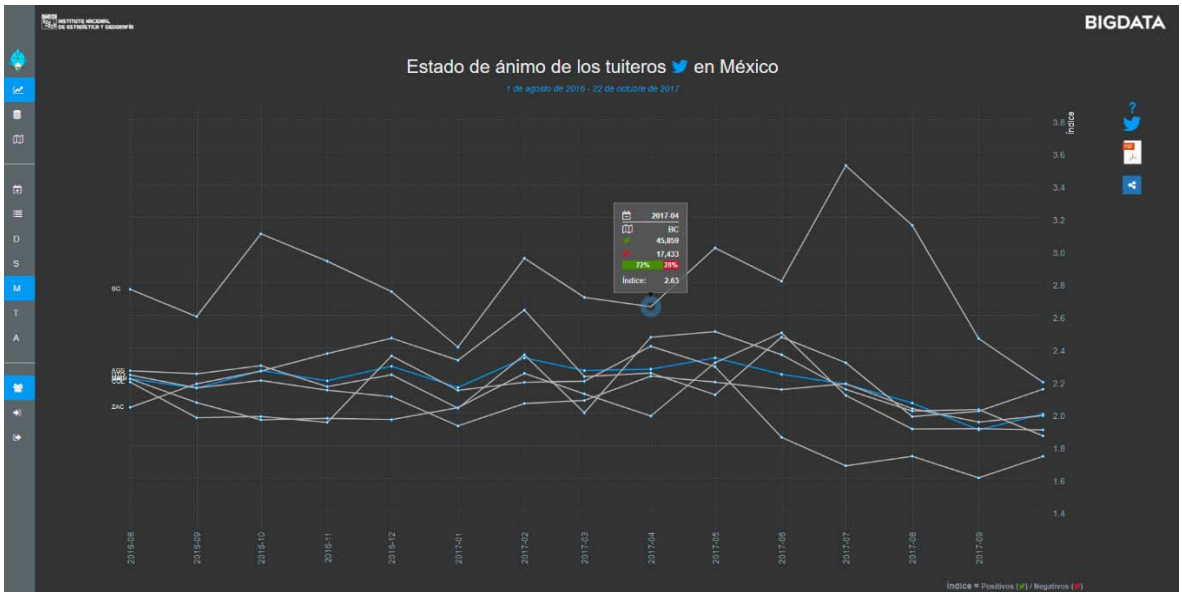


Se pueden visualizar todos los estados de la República Mexicana con el fin de que sean comparables entre sí.



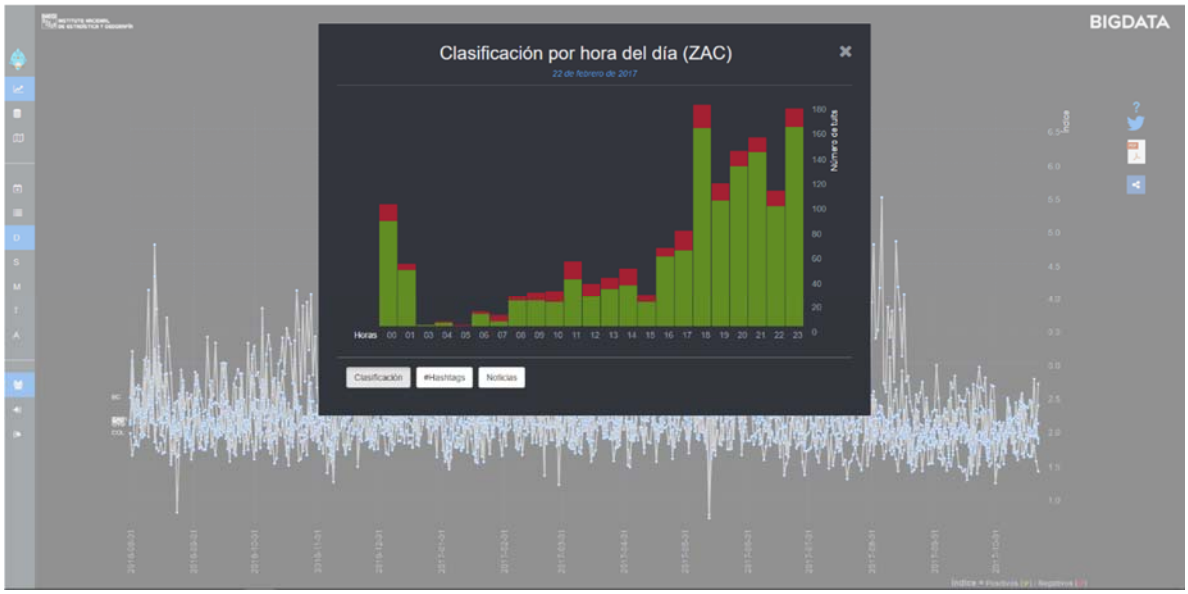

En el componente de filtros para fechas se puede visualizar tanto los tuits recolectados como el estado de ánimo en un rango de días específico, el mes actual, año actual, últimos 5, 15, 30, 60 y 90 días.


La gráfica del índice del Estado de ánimo de los tuiteros se muestra a nivel nacional y por entidad federativa. Al pasar el cursor sobre cualquier círculo de medición aparecen, la cantidad de tuits positivos y negativos, y el índice de ese punto (día, mes, etcétera).




También se visualizan los indicadores diarios, semanales, mensuales, trimestrales o anuales.

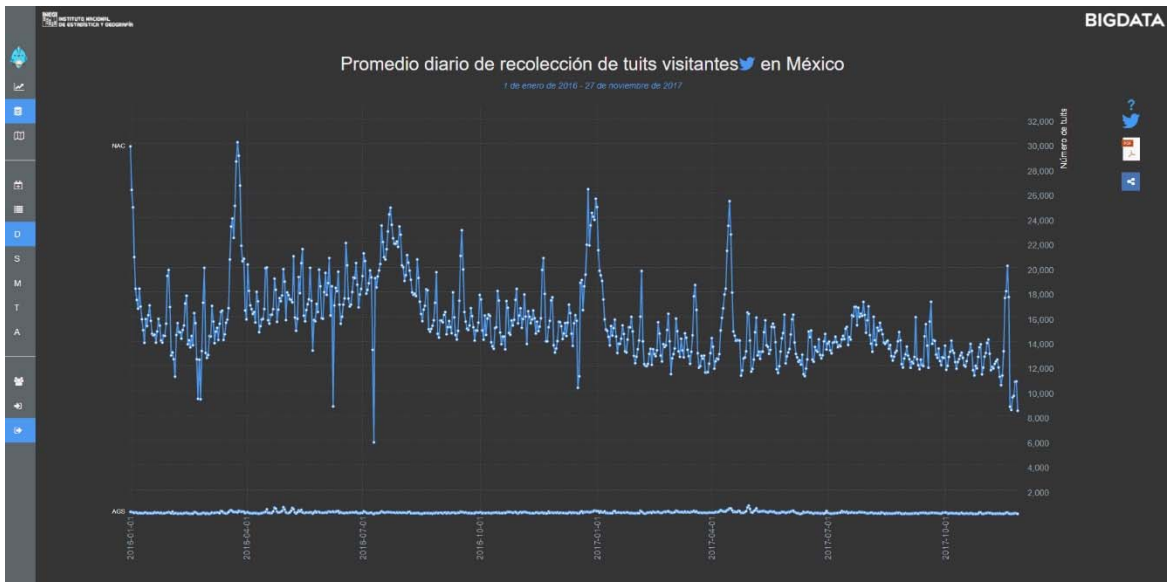
En el caso del indicador diario, es posible dar clic en el círculo de medición para visualizar la clasificación por horas del día seleccionado, así como una nube de *#hashtags* más utilizados en el día. Se puede también hacer clic en alguno de los #hashtags y la aplicación nos redirige directamente a Twitter para ver las publicaciones originales. Adicionalmente, se tiene una sección de los noticieros digitales que nos redirecciona a los sitios web de las noticias destacadas del día seleccionado.




La escala es relativa tomando como máximo el valor más grande de todos los índices y como mínimo el valor más pequeño de los mismos, del periodo seleccionado; se usa la misma escala para todo el 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, el color es más verde, y mientras más negativo, el color es más rojo. Cabe mencionar que para el cálculo del índice no se tomaron en cuenta los tuits clasificados como neutros, y por ende no se visualizan.

En el mapa de emotividad, al pasar el cursor por una entidad federativa, se muestran la cantidad de tuits positivos, la de tuits negativos y el valor del índice. Además, al hacer clic en la entidad, se presenta la gráfica del índice en el período seleccionado.


En la aplicación siempre se visualizan los datos de todos los tuits recolectados (locales y visitantes), pero se pueden visualizar los datos de locales y visitantes por separado, haciendo clic en los íconos del menú de lado izquierdo/abajo; tanto en la recolección como en el ánimo.




**Notas:**

En Internet Explorer solo funciona a partir de la versión 10.

En Internet Explorer los enlaces de la nube de *hashtags* no funcionan debido que Internet Explorer no soporta dentro de la etiqueta *text* código HTML por ejemplo: <a href="">