Modelamiento de tópicos aplicado al análisis de contenido de los tweets sobre el dengue en Colombia

Yanci Katherine Arenas Silva

Director:
Yuly Andrea Ramírez Sierra
M.sc en Ingeniería Industrial

Codirector:
Henry Lamos Diaz
PhD. en Física-Matemática

Universidad Industrial de Santander
Facultad de Ingeniería Fisicomecánicas
Bucaramanga
2022

Tabla de Contenido

Lista de Figuras

Lista de Tablas

Lista de Apéndices

Ver apéndices adjuntos y pueden ser consultados en la base de datos de la Biblioteca UIS

Resumen

Título: Modelamiento de tópicos aplicado al análisis de contenido de los tweets sobre el dengue en Colombia[*]

Autor: Yanci Katherine Arenas Silva[**]

Palabras Clave: tópicos; tweets; dengue; Colombia; modelamiento.

Descripción:

El uso de las redes sociales en la actualidad permite aprovechar la interacción y la conectividad entre usuarios, mediante el aprovechamiento de la cantidad de datos que se generan, que se consideran de gran utilidad al ser procesados mediante técnicas de modelamiento de tópicos para generar conocimiento mediante el cual se puede dar apoyo a la toma de decisiones entorno a modelos de control o creación de estrategias de acción.

En este proceso investigativo se aplicaron técnicas de modelamiento de tópicos que permiten analizar contenido de Twitter particularmente relacionado con el dengue en Colombia con los cuales se detectaron los tópicos más predominantes dentro de un periodo de estudio entre desde enero de 2019 hasta diciembre de 2020 y realizar su respectivo análisis y profundización en los mismos.

Luego de la extracción y procesamiento de los datos, se aplican los algoritmos para modelamiento de tópicos: Latent Dirichlet Allocation y Correlated Topic Models, para escoger el que mejor se aplique al caso estudiado mediante el uso de la medida de Loglikehood, dando como resultado el modelo LDA como el mejor, a través del cual posteriormente se determinaron los tópicos más significativos de los cuales se habla y que están relacionados con el dengue en específico.

[*] Trabajo de Grado

[**] Facultad de Ingenierías Físico-Mecánicas. Escuela de Estudios Industriales y Empresariales. Director: Yuly Andrea Ramírez Sierra. MS.c en Ingeniería Industrial

Abstract

Title: Topic modeling applied to content analysis of tweets about dengue fever in Colombia [*]

Author: Yanci Katherine Arenas Silva [**]

Key Words: Topics; Tweets; Dengue; Colombia; Modeling.

Description:

The use of social networks nowadays allows taking advantage of the interaction and connectivity between users, through the use of the amount of data generated, which are considered very useful when processed by topic modeling techniques to generate knowledge that can support decision making around control models or the creation of action strategies.

In this research process, topic modeling techniques were applied to analyze Twitter content, particularly related to dengue in Colombia, with which the most predominant topics were detected within a study period from January 2019 to December 2020, and to perform their respective analysis and deepening in them.

After the extraction and processing of the data, the algorithms for topic modeling are applied: Latent Dirichlet Allocation and Correlated Topic Models, to choose the one that best applies to the case studied through the use of the Loglikehood measure, resulting in the LDA model as the best, through which the most significant topics that are related to dengue specifically were subsequently determined.

[*] Trabajo de Grado

[**] Facultad de Ingenierías Físico-Mecánicas. Escuela de Estudios Industriales y Empresariales. Director: Yuly Andrea Ramírez Sierra. MS.c en Ingeniería Industrial

Introducción

En la última década, el uso de redes sociales y el aprovechamiento de la conectividad han aumentado su desarrollo significativamente y han mejorado la interacción entre los usuarios, revolucionando campos como la administración, educación, el marketing, la salud y el comercio e inclusive la vida cotidiana (Ortiz, 2019); la redes sociales han generado una gran cantidad de datos diariamente que pueden ser analizados por medio de distintas métodos como el modelamiento de tópicos, con el propósito de obtener conocimiento significativo para el apoyo de toma de decisiones de las entidades interesadas (Kalyanam et al. , 2017).

Es posible encontrar información epidemiológica en estas plataformas casi que, en tiempo real, pues los usuarios generan nuevos contenidos, nuevas publicaciones y fomentan las conexiones con otros usuarios continuamente (Missier et al. , 2016). Esto significa para los investigadores una oportunidad para aprovechar el uso de las redes sociales como fuente de información. Particularmente Twitter es una red de información abierta que conecta a una amplia variedad de usuarios para compartir información en torno a muchos temas e intercambiar ideas en línea, su característica clave es la naturaleza pública de sus tweets, permitiendo la recuperación de la información generada por medio de una interfaz de programación de aplicaciones (API) que permite recopilar tweets, retweets o hashtags; con esta información es posible realizar análisis descriptivo, análisis de contenido y análisis de redes, estudiar métricas, sentimientos, composición de la red de usuarios y la relación entre elementos de Twitter con la oportunidad de descubrir conocimiento de valor (Stieglitz y Bruns, 2013; Twitter, 2018).

Con el desarrollo de esta investigación se busca aplicar modelamiento de tópicos para analizar el contenido de los tweets sobre información relacionada con el dengue en Colombia, para detectar tópicos en auge, estudiar métricas, entre otros; todo esto con el propósito de contribuir al desarrollo investigativo del campo de la salud desde los métodos de analítica de datos. Cabe aclarar que se entiende al estudio de métricas como una herramienta de investigación apoyada tanto en modelamiento matemático como en el apoyo de indicadores cuantitativos para la generación de

conocimiento con el cual se puede tanto interpretar como profundizar en la interrelación que existe entre resultados científicos y resultados que pertenecen a la vida cotidiana (Spinak, 1996).

Este proyecto investigativo está integrado por la revisión literaria entorno a la temática propuesta, compuesta por un análisis bibliométrico y un análisis preliminar de literatura; el planteamiento del problema; la construcción del marco de referencia (marco de antecedentes y marco teórico); la metodología propuesta y, por último, el desarrollo del caso de estudio de la investigación en donde se desarrolla toda la metodología para la obtención de conocimiento.

Tabla de Cumplimiento de Objetivos

Tabla 1

Cumplimiento de objetivos

Objetivo	Cumplimiento (Numeral)
Realizar una revisión de literatura para identificar los usos del modelamiento de tópicos aplicado al análisis de datos de redes sociales en información sobre eventos de salud pública.	2
Comparar métodos de modelamiento de tópicos aplicados a los datos obtenidos de la red social Twitter con el fin de identificar el modelo más representativo de acuerdo a métricas de evaluación y validación.	6.6
Identificar tendencias en los datos objeto de estudio relacionado con el dengue, con el fin de categorizar la información.	6.4, 6.5 y 6.7
. Elaborar un artículo de carácter publicable, en donde se den a conocer los resultados de la investigación.	Apéndice C

1. Objetivos

1.1 Objetivo general

Aplicar modelamiento de tópicos al análisis de contenidos de los tweets sobre información relacionada con el dengue en Colombia.

1.2 Objetivos específicos

Realizar una revisión de literatura para identificar los usos del modelamiento de tópicos aplicado al análisis de datos de redes sociales en información sobre eventos de salud pública.

Comparar métodos de modelamiento de tópicos aplicados a los datos obtenidos de la red social Twitter con el fin de identificar el modelo más representativo de acuerdo a métricas de evaluación y validación.

Identificar tendencias en los datos objeto de estudio relacionado con el dengue, con el fin de categorizar la información.

Elaborar un artículo de carácter publicable, en donde se den a conocer los resultados de la investigación.

2. Revisión de la literatura

2.1 Análisis bibliométrico

Para valorar la actividad científica y el impacto que ha tenido este tema en la investigación, a continuación, se presenta el análisis bibliométrico de las publicaciones científicas desarrolladas en los últimos años, detallando la estrategia de búsqueda para la identificación de artículos.

Inicialmente, se establece el tópico a investigar junto con los criterios de inclusión, calidad y exclusión, con el objetivo de seleccionar los artículos que se toman como referencia para el desarrollo de esta investigación. Dentro de la estrategia de búsqueda se elaboró la matriz de palabras de interés, incluyendo sus sinónimos o variantes ortográficas.

Tabla 2

Palabras clave

infectious disease	arboviral diseases	Arboviruses
machine learning	social media	Twitter

Con los anteriores términos se construyen las ecuaciones de búsqueda, las cuales se utilizan en bases de datos tales como: PubMed, Web of Science y Google académico.

Tabla 3

Ecuaciones de búsqueda

Buscador	Ecuación de búsqueda
PubMed	TS= ("infectious disease" OR "arboviral diseases" OR Arboviruses) and "machine learning" and (“social media” or twitter)

Web of Science	TS= ("infectious disease*" OR "arboviral diseases" OR Arboviruses) and "machine learning" and (“social media” or twitter)
Google académico	TS= ("infectious disease*" OR "arboviral diseases" OR Arboviruses) and "machine learning" and (“social media” or twitter)

Partiendo de los resultados, derivados de las ecuaciones resultantes, se emplean los criterios de exclusión, de inclusión y de calidad, con el fin de seleccionar los documentos potenciales.

Tabla 4

Criterios de selección

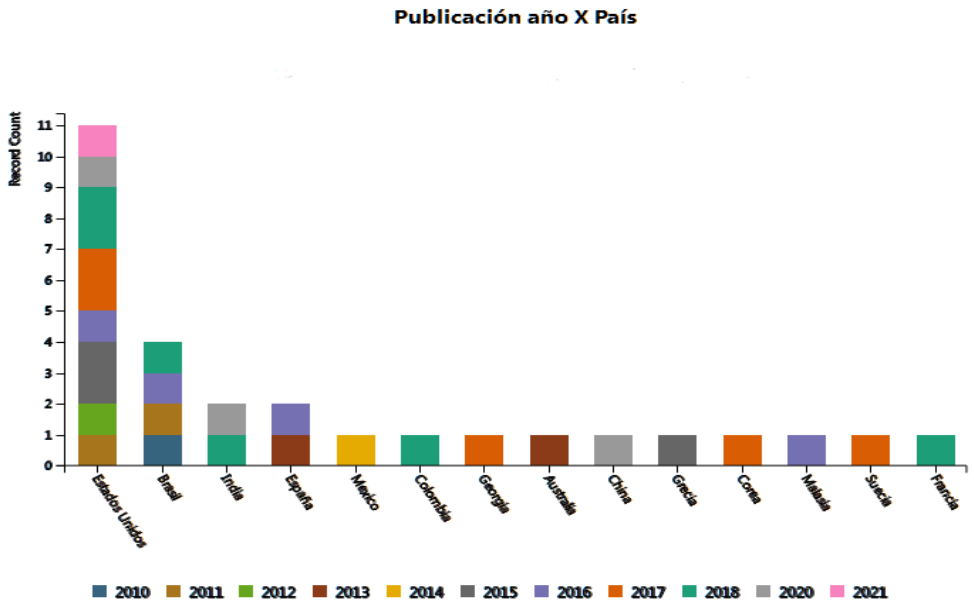
Criterios	Descripción
De exclusión	Boletines, comunicados de prensa o conferencias. Documentos de versión resumida. Artículos que no están orientados a observar una pandemia o enfermedades arbovirales. Artículos que no encajan dentro de ningún criterio de inclusión.
De inclusión	Área relacionada con la temática. Documentos en inglés, español u otros idiomas. Documentos tipo artículos o literatura gris. Últimos diez años, es decir, periodo comprendido entre 2010-2021
De calidad	Artículos que, aunque cumplen con los términos de búsqueda no están relacionado de forma directa con el tema de interés.

Como resultado de la búsqueda en bases de datos, se obtienen 146 resultados en Web of Science, 19 en PubMed y 77 en google académico, obteniendo en total 242; seguidamente, se eliminan los artículos duplicados, disminuyendo a 212. A continuación, se analiza el título, el resumen con el fin de identificar la pertinencia de cada documento, lo que lleva a que se elijan 40 artículos como parte de la revisión de literatura.

Posteriormente, se procede a realizar el análisis bibliométrico de los documentos haciendo uso de la aplicación de minería de texto denominada Vantage Point, software para el análisis de texto y la creación de informes estadísticos coherentes, relevantes y de precisión, y permite crear múltiples herramientas que coadyuvan al investigador a comprender y analizar de manera diversa y didáctica la información científica recolectada.

Figura 1

Número de Publicaciones por año y por país



Nota. Adaptado de Vantage Point (2021).

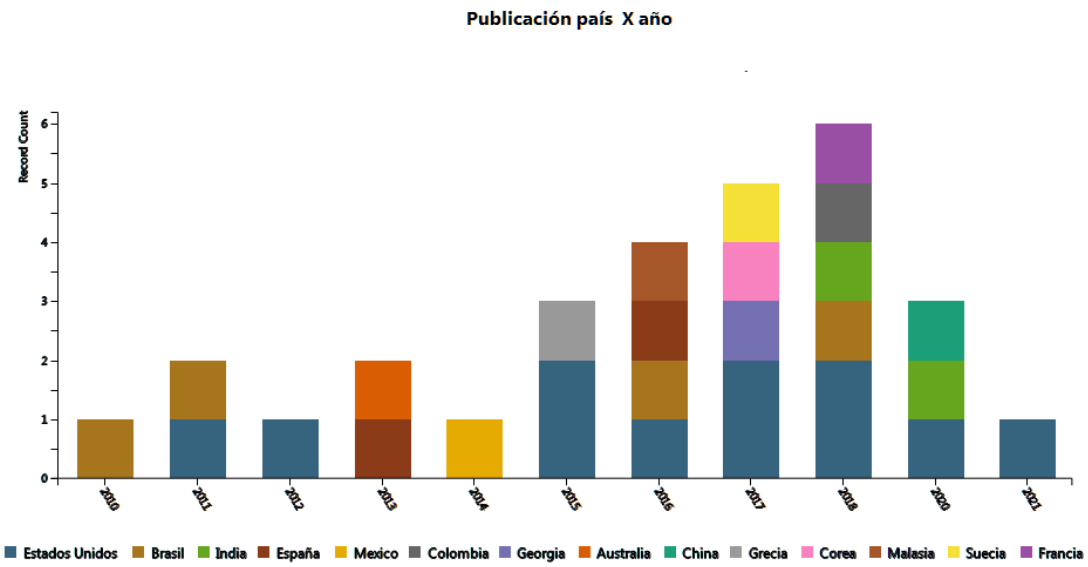
En el primer Informe visualizado en la figura 1, se encontró:

- Para un periodo comprendido entre 2010-2021, el país que mayor número de publicaciones tiene es Estados Unidos con once, correspondiente al 25 % de la totalidad, seguido Brasil con cuatro (9%), e Italia y España con dos; los demás países han publicado únicamente un artículo.

- Brasil es intermitente en sus publicaciones, se ausento cinco años entre 2011 y 2016, por el contrario, Estados Unidos publica con mayor periodicidad.

Figura 2

Número de Publicaciones por país y por año



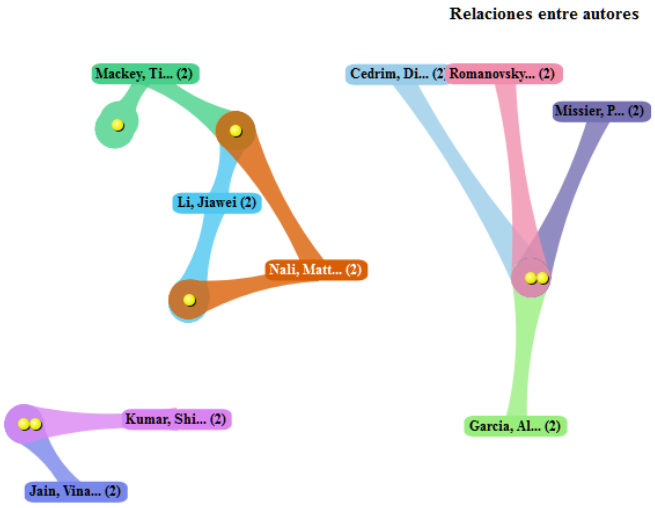
Nota. Adaptado de Vantage Point (2021).

A partir de la figura 2 se puede concluir que, en el periodo analizado, en promedio se publican 2 artículos por año, 2018 es el año que más número de publicaciones tiene en torno al tema; el comportamiento entre años es variable, por ejemplo, entre 2015 y 2018 fue ascendente y desciende en los últimos tres años.

- Además, el interés de Estados Unidos en investigar se observa desde el año 2011, en Brasil desde 2010, en Francia en el 2018 al igual que en la India y Colombia.

Figura 3

Análisis de co-ocurrencia de autores

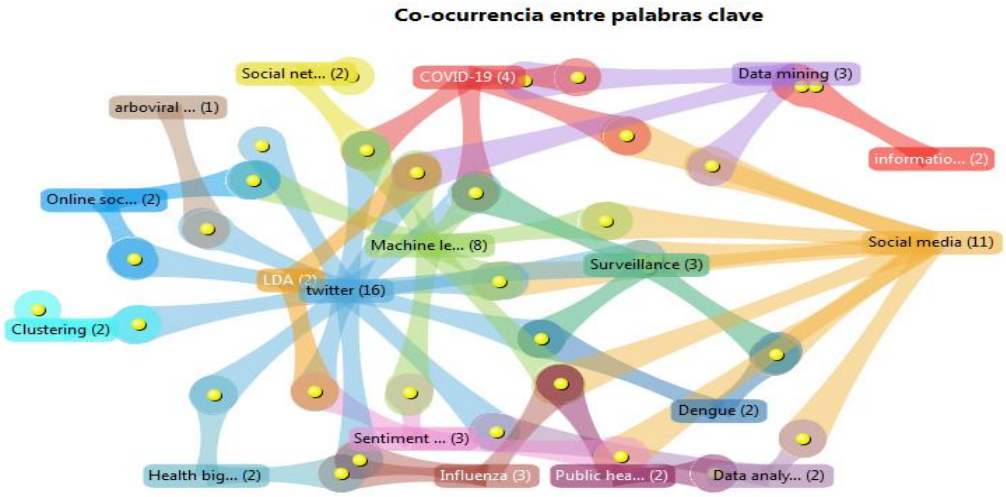


Nota. Adaptado de Vantage Point (2021).

En la anterior figura, se observan las relaciones de co-ocurrencias entre autores, el número de artículos en los que se presenta este comportamiento por cada nodo relacional son dos, para todas las visualizaciones presentes; resaltando a Kumar, Shishir; Li, Jiawei; Nali, Matthew; Mackey, Tim K; Cedrim, Diego; Jain, Vinay Kumar; García, Alessandri y Missier, Paolo como los autores que publican con mayor periodicidad.

Figura 4

Análisis de co-ocurrencia de palabras claves



Nota. Adaptado de Vantage Point (2021).

De acuerdo con la figura 4, las temáticas más frecuentes son: Twitter, Social Media, Machine Learning, Data Mining, Dengue, LDA, influenza, sentiment analysis, Surveillance, Text Mining, arboviral disease, Covid 19 y Clustering.

2.2 Análisis preliminar de la literatura

La rápida evolución del internet ha dado lugar a la tecnología Web 2.0, un nuevo modelo para crear y compartir información, sobrepasando la distancia geográfica entre los usuarios, a través de la creación de plataformas de interacción, entre ellas Instagram, Facebook, Twitter, Youtube, Linkedin, WhatsApp, conocidas actualmente como redes sociales; si bien, una red social no es solamente esto, sino que también ayuda al fortalecimiento de la interacción, de los procesos colaborativos y de generación de ideas , del apoyo en situaciones y experiencias, entre otras cosas más.

Análisis de redes sociales

Cada instante los individuos interactúan entre si estableciendo relaciones, Sorin, (2011) afirma que el análisis de redes sociales es un conjunto particular de interrelaciones (en inglés, linkeages) entre un conjunto limitado de individuos, con la propiedad adicional de que las características de estas interrelaciones, consideradas como una totalidad, pueden ser utilizadas para interpretar el comportamiento social de las personas implicadas (p.408).

Entorno a diferentes ciencias, entre ellas la psicología, la antropología, las matemáticas, surgió el enigma de estudiar, analizar y sobretodo entender todos aquellos fenómenos relacionales que surgen a partir de la interacción en redes sociales; dando lugar al término de análisis de redes sociales definido por Santos Javier en 2017 como "una aproximación metodológica y teórica que enfatiza el estudio de las relaciones entre actores, tanto relaciones entre personas, organizaciones, países o cosas"(párr. 1.); es allí en donde a partir de la creación de teorías y métodos se investiga cómo se genera, cómo cambia y cuál es la influencia que se produce, mediante la implementación de distintas herramientas tanto matemáticas como informáticas las cuales se encarguen del procesamiento y el análisis de dicha información.

En el transcurso de la última década, la utilización de este estilo de investigación está en crecimiento continuo, cada vez son más los artículos o libros que incluyen dentro de sus líneas el "análisis de redes sociales (AR) o (ARS)", dado que se complementa y es transversal a distintas técnicas y perspectivas de investigación cuantitativa y cualitativa. Este avance está enfocados en investigaciones relacionadas con distintas temáticas entre ellas: el análisis y la descripción de relaciones interpersonales, el estudio de estructuras organizacionales o corporativas, el seguimiento a la innovación agrícola, el mejoramiento de la seguridad de entidades sanitarias, el nivel de cooperación de los integrantes de una organización, el estudio y seguimiento del comportamiento la salud pública, entre otras (Rodríguez, 2013).

Análisis de redes sociales en la salud pública

El análisis de redes sociales es un componente crucial dentro de la investigación en salud pública; gracias a la capacidad que tienen estos modelos de identificar y analizar estructuras horizontalmente, analizar su composición y resaltar los distintos puntos de interacción entre cada uno de los aspectos y áreas involucradas (Sánchez, 2014). Las investigaciones que comprenden esta área de la salud en los últimos diez años han aumentado, significativamente, a causa de la imposibilidad de acceder a dichas investigaciones a través de técnicas convencionales (Missier et al. , 2018).

Existe una vinculación casi directa entre los factores de riesgo y protección involucrados en la salud pública con las características observables de las redes sociales que se considera debe ser investigada dado que permite un mayor acercamiento a los distintos contextos sociales y ambientales, a los que están expuestos los individuos de la sociedad; a su vez es posible la comprensión, incluso a los procesos más básicos, permitiendo a quien investigue la posibilidad de mapear cualquier vínculo entre el grupo objetivo, identificar propagaciones, monitorear brotes de epidemias e incluso conocer los riesgos de enfermedades, todo esto a partir de la teorización, el modelamiento y la predicción que permite el ARS (PLoS Medicine, 2010).

Su gran importancia reside en que es posible inclusive llegar a identificar cómo actuar ante un evento en particular, dado que la obtención casi en tiempo real puede apoyar a la toma y acción de estrategias efectivas encaminadas a la intervención a tiempo de la situación, en particular en los casos de propagación de enfermedades. (Al-garadi et al. , 2016). De ahí que, en los últimos años, las investigaciones sobre el uso real del análisis de redes sociales en la salud pública sea un tema que ha generado bastante interés, sobretodo entorno a la investigación digital de detección de enfermedades e infodemiología, en vista que propone una manera más rápida de dar respuesta a grandes enigmas adaptando múltiples modelos al problema objeto de estudio (Marqués et al. , 2013). Algunos tópicos de investigación en esta área son: emergencias sanitarias, análisis epidemiológicos, utilidades del análisis de redes sociales en el ámbito sanitario, gestión sanitaria, ARS y los hábitos saludables, entre otros.

Para el campo de la epidemiologia se estudia, particularmente, los factores determinantes, los problemas de salud relacionados con enfermedades endémicas, epidemias, pandemias y brotes; destacándose los siguientes ejemplos: el proceso de identificación de un agente de cambio candidato con riesgo por infección de VIH para la prevención (Schneider et al. , 2012); el análisis tanto de la variación espacial como temporal del virus del dengue en el periodo comprendido entre 2011-2014 en Pakistán tomando como punto de partida el patrón de movilidad urbana. (Kraemer et al. , 2018); la detección integrada de temporadas de influenza y su actividad en entornos ya vigilados en Suecia con fines de predicción (Spreco et al. , 2017); el monitoreo de epidemias de gripe con la ayuda de datos recolectados en hospitales de Francia (Bouzillé et al. , 2018).

En sus estudios, Lampos y Cristianini (2010) y Gomide et al. (2011) señalaron que redes sociales como Twitter poseen un gran potencial para obtener información sobre epidemias, mediante las cuales las entidades sanitarias e investigadores pueden llevar a cabo un seguimiento en tiempo real y su vez descubrir conocimiento. Dado el volumen de datos que esta red maneja es viable implementar la metodología KDD (Knowledge Discovery in Databases).

Las fases que componen esta metodología KDD son: Recolección y Selección, preprocesamiento/ limpieza, transformación, minería de datos y conocimiento. La obtención o selección de conocimiento se ha realizado mediante distintas herramientas, dependiendo de la necesidad y los objetivos de quien investiga; una técnica es la preselección de tweets haciendo uso de palabras claves. Lampos y Cristianini (2010) la emplearon con el propósito de encontrar relaciones entre los síntomas de la gripe o influenza estacional; Hirose y Wang (2012) para identificar relaciones entre este contenido y los datos relacionados con enfermedades similares a la influenza y finalmente predecirse enfermedades infeccionas; Mackey et al. (2020) la usaron como método para posteriormente encontrar información relacionada con sintomatologías del Covid-19. También se ha usado el trabajo conjunto entre el uso de esta técnica y la implementación de algoritmos específicos de selección y extracción o mediante la combinación de análisis específicos de lenguaje y reglas de inferencia en la biovigilancia de enfermedades (Hartley et al. , 2013).

En otra investigación, esta recolección la hicieron creando diccionarios de inclusión y exclusión de palabras que estuvieran relacionadas con el dengue y luego mediante la API de Twitter extrajeron tweets que contuvieran dichas palabras, estableciendo a su vez un periodo de tiempo específico, este fue empleado por (Lauren et al. , 2017); este método, también fue empleado por Martínez et al. (2016) para el desarrollo de investigación en donde se estudiaba el flujo de en redes sociales relacionada con el consumo de fármacos y sus efectos adversos en la salud, de usuarios de Europa.

En la segunda etapa, el preprocesamiento, se remueven todos los datos sin relevancia para la investigación, también conocidos como datos ruidosos, y se establecen las estrategias y técnicas estadísticas para trabajar con conocimiento duplicado, nulo y desconocido (Timarán et al. , 2016). Una de las herramientas empleadas recientemente para la eliminación de texto es a través del kit "re" de Python (Zhu et al. , 2020); también es posible segregar el ruido describiendo detalladamente la temática, haciendo uso de palabras clave para posteriormente filtrar y clasificar la información en grupos dependiendo de la vinculación con el objeto de estudio. (Missier et al. , 2016; Zhou et al. , 2018).

En la siguiente etapa se mejora la calidad de los datos, conocida también como la etapa de reducción. Es acá donde se escogen las variables que se consideran útiles para la meta del proceso y se eliminan aquellas que se consideran redundantes (Timarán et al. , 2016). En su reciente investigación sobre una epidemia Odlum y Summoo (2015) transformaron el texto convirtiéndolo en un N-grama en donde le mantuvieron el uso del vector del formato y redujeron su dimensionalidad.

Seguido de esto, están las etapas de minería de datos e interpretación; en la etapa de minería lo que se busca es descubrir patrones, o tópicos relevantes mediante diversas técnicas a través de las cuales es posible crear tanto modelos predictivos como descriptivos. Algunas técnicas del modelo descriptivo pueden ser las reglas de asociación, las técnicas de aprendizaje automático, entre otros (Timarán et al. , 2016).

En 2017, un grupo de investigadores de la Universidad de California, señalaron que el aprendizaje automático sin supervisión aplicados al análisis de tweets, relacionados con la salud pública, es una estrategia eficiente para hallar patrones dentro de los datos casi que de forma inmediata y resumir el corpus concisamente, dada la eficiencia computacional de sus algoritmos y la ventaja de que la información no se ve afectada por ningún juicio personal. (Kalyanam et al. , 2017). A su vez dentro de su investigación sobre el abuso en el consumo de drogas y la automedicación, emplearon reiterativamente el aprendizaje BTM (Biterm Topic Model) para detectar patrones iniciales del conjunto de tweets, correlacionarlos y asegurarse de que la información es realmente relevante. Este modelo, también fue empleado para analizar la interacción entre usuarios de Twitter vinculada al movimiento “Libérate” durante la etapa inicial del contagio por Covid-19 en Estados Unidos (Haupt et al. , 2021).

En 2017, otro grupo de investigadores, estudiaban otra técnica conocida como la asignación de Dirichlet Latente (LDA) como herramienta para conocer cuáles eran los medicamentos para la influenza que se consumían en mayor proporción por ciudadanos chinos mediante un método de vigilancia a través de redes sociales (Kagashe et al. ,2017). En Corea, durante el mismo año se desarrolló una investigación que aplicaba esta misma técnica de descubrimiento de información para conocer las reacciones de toda la población frente a los brotes recurrentes de infecciones virales, sobre todo el relacionado con el síndrome respiratorio de Oriente Medio (MERS) (Choi et al. , 2017).

Tres años después, en una revisión sistemática entre 2010 y 2019 que estudiaba diversos sistemas de vigilancia de salud pública mediante el uso de redes sociales, los investigadores afirmaron que el modelo BTM (una técnica de aprendizaje no supervisado) ofrece un mejor enfoque de clasificación de contenido de Twitter que el modelo LDA, no supervisada, debido a que el nivel de control sobre el contenido puesto en análisis es mejor que el anteriormente mencionado, opinión sustentada en la investigación desarrollada por (Missier et al. , 2018) relacionada con la prevención del dengue haciendo uso de redes sociales (Gupta y Katarya, 2020).

3. Planteamiento del problema

Una sociedad exitosa está fundamentada en diferentes pilares, entre ellos, la promoción y prevención de la salud; gracias a esto es posible tener una población saludable en donde cada individuo se encarga de autocontrolar su bienestar físico, social y mental, coadyuvando a garantizar altos niveles de calidad de vida (Universidad Internacional de Valencia, 2018).

En este sentido, en busca de comprometerse con la contribución del logro de los objetivos de desarrollo sostenible a nivel mundial, y como estrategia de transformación de la vida de millones de personas en pro de garantizar la cobertura universal de información sanitaria, en la 72° asamblea mundial de la salud celebrada el 24 de mayo de 2019 cada estado miembro de la OMS acordó desarrollar acciones a corto y largo plazo encaminadas al fortalecimiento del monitoreo de la salud; así mismo acordaron trabajar en la recopilación, análisis y presentación de resultados obtenidos sobre el comportamiento de enfermedades arbovirales con el propósito de fortalecer la comprensión de cada uno de los desafíos, oportunidades e implicaciones que esto conlleva a la salud pública (Organizacion Mundial de la Salud , 2019).

Desde el año 2005 en la 58° asamblea mundial de salud se propuso a cada estado incluir los sistemas electrónicos de uso nacional dentro del ámbito de la salud pública (OMS, 2005). En Colombia, las primeras campañas de promoción de salud y prevención de enfermedades se desarrollaron mediante la implementación de folletos, vallas, divulgación en medios de comunicación radial, pautas de publicidad televisiva, anuncios en periódicos, entre otros; fue hasta el año 2012, en donde tanto entidades como población en coordinación con el Ministerio de Salud y siguiendo el plan decenal de salud pública 2012- 2021, se gestionaron y desarrollaron distintas acciones, campañas y planes de interacción en redes sociales con líneas de acción comunitaria, sectorial e intersectorial, todas estas enfocadas hacia la promoción, prevención y mantenimiento de la salud (MinSalud , 2021).

La importancia de las redes sociales en la salud pública “no se encuentra sólo en potenciar la comunicación horizontal, sino que el análisis automatizado del tráfico de información compartida

a través de ellas permite detectar patrones y comportamientos asociados a la búsqueda y suministro de información sobre salud" (párr.2) citado por Serri (2018). Esta incorporación de las redes sociales con fines de vigilancia de salud pública es de avance lento, sin embargo, logra conectar personas de distintos sectores y culturas del país con situaciones similares y con intereses en común; mediante un sin número de publicaciones, ya sea positivas o negativas, es posible, la generación de un contagio emocional, fenómeno que contribuye a determinar las acciones que se deben desarrollar para garantizar el control de la salud pública.

A nivel mundial y en la última década ha existido una creciente preocupación por los índices de crecimiento del número de contagiados por arbovirus, un grupo de aproximadamente 500 virus encontrados en la naturaleza que se transmiten por vectores y se puede propagar entre vertebrados. Dentro de los vectores más conocidos se encuentran las garrapatas, los chinches, los mosquitos, las pulgas y las moscas; aproximadamente 150 de estas especies virales pueden causar alguna enfermedad en los seres humanos y unas cuantas de estas se pueden diseminar de un ser humano a otro a través de los artrópodos, donación de órganos o transfusión de órganos, entre las más conocidas el dengue, el zika, la fiebre amarilla, el chikungunya, la fiebre de lassa, el ébola, la fiebre hemorrágica, entre otros (Manual MSD, 2020); actualmente el reto del control de estas enfermedades para el gobierno de las zonas afectadas se encuentra en la detección efectiva, en los métodos de prevención y en la identificación geográfica de la presencia del vector transmisor, esto último dado que están en constante evolución y procesos de adaptabilidad provenientes de cambios relacionados con factores tanto ambientales como sociales, entre los que se encuentran el cambio climático, los procesos de deforestación, de urbanización, de migración, entre otros, que permiten en gran medida la circulación variante de las zonas con presencia de brotes a una zona no endémica (Missier et al. , 2016).

De todas las enfermedades causadas por estos virus la más común y de mayor resurgimiento es el dengue, según registros de la OMS para el año 1970 solo nueve países a nivel mundial presentaban casos de contagio de esta epidemia , no obstante, en el año 2020 ascendió a más de 100 países (OMS, 2020); el 2018 ha sido registrado como el año con mayor número de muertos por esta enfermedad, sin embargo, las cifras muestran que en 2019 se reportó la mayor cantidad

de casos en todo el mundo siendo en Latinoamérica en donde se notificó la mayor propagación con cerca de 3,2 millones de contagios (BBC News, 2020).

A nivel nacional, en la semana epidemiológica 07, el Ministerio de Salud reporto en lo que iba corrido del año 2020 en total 24.489 casos de dengue, distribuidos entre los departamentos de Atlántico, Antioquia, Sucre, Santander, Cesar, Meta, Cundinamarca, Huila, Tolima y Valle del Cauca (MinSalud, 2020); de manera análoga, de acuerdo al último informe presentado por el observatorio de salud pública en Santander se registraron en total 3624 casos de dengue y 66 de dengue grave por cada 100.000 habitantes durante el año 2016 (OSPS, 2017).

Por tanto, la detección temprana y oportuna de estos brotes está dentro de los propósitos de monitoreo de todos los países que integran la OMS y de sus organizaciones de salud, enfocándose en la eliminación de los focos geográficos vectoriales de arbovirus; una fuente comúnmente utilizada para esto es la interacción de usuarios en redes sociales, debido al desarrollo de conversaciones espontáneas y la transmisión continua de información, aspecto de gran importancia que permite identificarlos como referentes a través de los cuales es posible obtener volúmenes de contenido valioso (Vasileios y Nello, 2010); especialmente twitter, que permite la producción de conocimiento en tiempo real sobre lo que sucede en el mundo, y se considerada como una red social con potencial único para monitorear epidemias, en particular las relacionadas con los brotes de influenza y el dengue. (Missier et al. , 2016; McCreadie et al. , 2013).

El análisis de datos de redes sociales hace parte de los nuevos enfoques desafiantes de la analítica de datos, y con el cual también surge la necesidad implícita de explorar las técnicas y herramientas de mayor uso y eficiencia, entre las que se destacan: el análisis de clúster, análisis de grafos, las reglas de asociación y las técnicas de aprendizaje automático tales como modelos de regresión, clasificadores de Bayes, arboles de decisión, algoritmos de agrupamiento con el objeto de identificar la más pertinente y la que permita obtener conocimiento de alta validez para ser utilizado para el desarrollo de estrategias eficientes y efectivas por parte de las entidades gubernamentales de salud.

Por lo tanto, se propone una investigación en la que se busca analizar información sobre el dengue que se encuentra en el contenido de la red social Twitter, abordando la campaña "Hay que cortarle las alas al dengue" adoptada por el Ministerio de Salud y Protección Social, como iniciativa para la prevención del contagio por dengue en Colombia durante el año 2019, además de la información recolectada de cualquier publicación asociada con el dengue en Twitter , con el fin de aplicar métodos de analítica de datos como el modelamiento de tópicos que permite categorizar la información e identificar aquellos aspectos destacables en la información objeto de estudio, para apoyar el análisis de tendencias en la red social y contribuir con conocimiento sanitario.

Además, los resultados de esta pasantía de investigación aportarán al proyecto de investigación titulado "Modelos De Analítica De Datos Para La Vigilancia De Enfermedades Arbovirales En El Departamento De Santander", aprobado por la VIE y liderado por el grupo OPALO con el apoyo del grupo de investigación de la UDES(Universidad De Santander), en lo que concierne a la captura de información de la red social Twitter a partir de palabras claves relacionadas con el dengue en el departamento de Santander, con el fin de analizar los tópicos que se presenten y evaluar la posible presencia de esta enfermedad arboviral que pueda representar una nueva amenaza a la salud pública.

4. Marco de referencias

4.1 Marco de antecedentes

La importancia de las redes sociales y la analítica de datos para la comprensión de los eventos han venido siendo estudiada durante años, es por esto que en los últimos años se han realizado investigaciones que permiten estudiar su validez y mejorar su comprensión.

En el trabajo de grado titulado "Análisis de tendencias en marcas deportivas a través de Twitter" elaborado por Mesas en el año 2015, en este se desarrolla una herramienta encaminada al análisis de campañas publicitarias particularmente en Twitter, haciendo uso de la metodología KDD. Para

dar inicio ellos realizan el proceso de extracción y análisis de tweets particulares, luego en la etapa de procesamiento se aplican técnicas específicas de Clustering, con el propósito de encontrar los términos de mayor relevancia y asociarlos mediante clúster, que les permitió posteriormente ver como estos se relacionaban e influían en la propagación y en el fomento de la campaña publicitaria, asimismo poder determinar en donde se está generando dicha relevancia para tales campañas. Acto seguido, se implementan métodos de geolocalización que ayudan a evaluar el alcance, finalmente se emplea la teoría de grafos para identificar quienes intervienen en mayor o menor proporción, es decir la red que se genera entorno a la misma. Este trabajo se puede aprovechar como medio de apoyo para el análisis de campañas y la determinación del impacto.

En el 2019, se desarrolló un proyecto de grado titulado "Reglas de asociación aplicadas al análisis de contenido de los tweets sobre enfermedades trasmitidas por vectores en Santander, Colombia". La metodología que propusieron los investigadores fue enfocada en el uso de algoritmos de clasificación supervisada junto con modelos de reglas de asociación para el análisis de tweets publicados, y que incluyera información relacionada con enfermedades de transmisión vectorial; enfocada a identificar que tópicos son más frecuentes al momento de expresar opiniones sobre alguna de estas enfermedades en particular, y que a su vez sirvieran como muestra destacada en el análisis de la salud pública respaldando la supervisión de estas epidemias. A partir de esta investigación fue posible concluir que Twitter si es una fuente útil además de que los modelos escogidos si pueden ser implementados para el apoyo a las estrategias de control y prevención de enfermedades de transmisión vectorial (Rodríguez y Rojas, 2019).

Un año después, Olarte y Ariza (2020) en su trabajo de investigación titulado "Evaluación de métodos de agrupamiento DBSCAN y LDA para el análisis de contenido de la red social Twitter" buscan estudiar la efectividad de estos dos algoritmos dado el uso reiterativo que han tenido en diferentes estudios por la garantía de la información generada. Estos han sido empleados generalmente con el propósito de desarrollar análisis de percepción de marcas y sentimientos que conlleven a proponer estrategias de marketing viables enfocándose en la información que circula en la red social Twitter. En este caso en particular se analizó contenido real de Twitter relacionada con información vinculada con la Universidad Industrial de Santander, UIS, en un periodo de

recolección de cinco meses. Se destaca el desarrollo de esta investigación porque los métodos implementaos para identificar tendencias y temas de relevancia en la información recopilada son relevante para el desarrollo de la investigación en desarrollo.

4.2 Marco teórico

Para ayudar a un mejor y más amplio entendimiento de la temática en investigación se presentará a continuación las principales referencias conceptuales que ayudan a entender con claridad tanto los resultados como el desarrollo del proceso investigativo.

Salud Pública

A nivel nacional y de acuerdo a la ley 1122 de 2007 se entiende la salud pública como:

El conjunto de políticas que buscan garantizar de una manera integrada, la salud de la población por medio de acciones de salubridad dirigidas tanto de manera individual como colectiva, ya que sus resultados se constituyen en indicadores de las condiciones de vida, bienestar y desarrollo del país. Dichas acciones se realizarán bajo la rectoría del Estado y deberán promover la participación responsable de todos los sectores de la comunidad (MinSalud, 2014, párr.10).

Enfermedades de transmisión vectorial

Estas contemplan todas aquellas enfermedades en humanos que provienen agentes patógenos como bacterias, parásitos o virus y que son causas por vectores, entendiendo estos como un grupo de organismos con vida que tienen la capacidad de trasmitir algún patógeno de infección en un campo de acción de personas a personas o de animales a personas, en gran parte los insectos ingieren el microorganismo patógeno a través de un portador ya infectado para luego trasmitirlo activa o pasivamente , llegando incluso a trasmitirlo toda el ciclo de vida (OMS, 2020).

Esta transmisión se presenta particularmente en zonas geográficas delimitadas por sus características físicas y ecológicas, que cuentan con elementos que permiten y favorecen su reproducción y vivencias (Padilla et al. , 2017). Esta epidemia ocasiona un poco más de un millón

de muertes anuales en todo el mundo, en América una de cada dos personas se padece alguna ETV. Entre las más frecuentes están el chikunguya, el dengue, la malaria y el zika (OMS, 2020).

Dengue. Este virus es transmitido por la picadura de un mosquito de la familia Aedes con presencia de infección, alrededor de 400 millones de personas cada año son infectadas por este animal de las cuales 100 millones superan la enfermedad y 22 000 mueren de cualquiera de los cuatro tipos de dengue (Centro para el control y la prevención de la enfermedad, 2019).

De acuerdo a un informe presentado por la OMS esta es la infección vírica, una de las mayores causas tanto de enfermedad como de muerte en el mundo, sobretodo en América y en Asia, actualmente no presenta un tratamiento estándar por lo que la prevención es una estrategia clave para ayudar en la disminución de los índices de contagio (OMS, 2020).

Chikungunya. Es una enfermedad vírica de la misma familia que el dengue, con mayor presencia en el continente africano y asiático, no obstante, América vivió un brote en el año 2013 y otro en el 2015; en Europa el primer brote se videncia en el 2013. Es una enfermedad de difícil diagnóstico, motivo por el cual se desconocen las cifras precisas, a nivel mundial, de personas afectadas por su contagio (OMS, 2018).

Zika. Es una enfermedad transmitida por mosquitos. África, América, Asia y regiones cercanas al pacifico se han visto afectadas por la presencia de este virus. No solo se transmite a animal a humano, sino también de la madre a feto, cuando hay casos de embarazo, e incluso puede llegar a propagarse mediante el contacto sexual, cuando existe transfusión de productos sanguíneos y en casos de trasplante de órganos (OMS, 2013).

Leishmaniasis. Es causada por un tipo de protozoo, organismo compuesto por una célula y que se transmite por la picadura de un flebótomo hembra. Existen tres tipos de esta enfermedad: Visceral, cutánea y mucocutanea. Se considera que las poblaciones con más baja cantidad de recursos son las que más se ven afectadas debido a la baja calidad

alimenticia, las precarias condiciones de vida y por ende la fragilidad del sistema inmunológico. A nivel mundial, el número de casos puede estar cerca del millón al año, pero no toda aquella persona que la porta padece los síntomas. En Casos en donde es visceral, casi el 100% derivan en la muerte, la más frecuente es la cutánea con padecimientos de ulceras y la mucocutanea puede llegar a ocasionar la pérdida total de la garganta, nariz o boca (OMS, 2018).

Chagas. Es una enfermedad con un índice de mortalidad bastante elevado, a nivel mundial anualmente cerca de siete millones de personas se ven afectadas, se han identificado presencia de ventores en 21 países latinoamericanos, sobre todo en zonas rurales. Su transmisión es vía heces u orina de un animal conocido como chinche, madre-hijo, trasplante de órganos, transducciones de donaciones sanguíneas y por consumo de alimentos contaminados (OMS, 2018).

Campaña "Hay que cortarle las alas al dengue"

Surge a raíz del pronóstico en 2019 que los casos de contagio por este enfermedad arboviral aumentarían considerablemente, y a su vez como acción para el fortalecimiento de la prevención, vigilancia y control de la misma; liderada por el Ministerio de Salud y Protección en Colombia, con el propósito de aumentar el nivel de vinculación por parte de la población a las actividades de control y al mantenimiento de las acciones de prevención catalogándose así como un trabajo de colaboración colectiva (MinSalud , 2019).

Juan Pablo Uribe (2019), ministro de salud en su momento, explicó:

"El dengue afecta habitualmente al país; tenemos 752 municipios que, al estar debajo de los 2.200 metros sobre el nivel del mar, presentan el vector y tienen un patrón de transmisión endémico en los que circulan los cuatro tipos del virus. A esto se suma que el 2019 es un año pico del ciclo interepidémico –cada tres años– y que el fenómeno de El Niño, con las altas temperaturas y lluvias ocasionales, favorece la reproducción del mosquito Aedes Aegypti, transmisor del virus" (párr. 6).

Redes Sociales

El concepto de red social es múltiple, sin embargo, es preciso mencionar que la definición más completa es la propuesta por Kuz et al., (2016): "Nuevos modos de socialización, a partir de ellas se puede tener una fuente de interacción entre las personas posibilitando la contextualización de fenómenos sociales entre los individuos y las relaciones inherentes que han surgido" (p. 1).

Twitter

Es una red social con un alto nivel de popularidad a nivel mundial, cuenta con más de 500 millones de usuarios, permite la difusión de información e interacción entre cuentas de forma transversal, genera en promedio 65 millones de tweets por día y fundada por Jack Darsey en 2006. Este canal es apto para desarrollar diversas actividades tales como crowdsourcing, networking, investigación en gestión del conocimiento, monitoreo, entre otras (Ignacio Santiago, 2019).

Twitter aprueba compartir información además del acceso a sus datos de forma programática con cualquiera de sus usuarios mediante su API, cabe señalar que este contenido ha sido establecido como de naturaleza pública al momento de ser publicada.

Elementos de Twitter. Dentro de esta red se implementan habitual y generalizadamente algunos términos presentados a continuación (Twitter, 2018):

- Los Tweets: así se les llama a los mensajes que publica cada usuario y tiene una extensión máxima de 280 caracteres, con excepción de Corea, Japón y China.

- #hashtags: es una forma de destacar temáticas vinculando el tema de conversación dentro de etiquetas; además, es una herramienta que usa Twitter para organizar la información. Se conocen cuatro tipos de hashtags: de marca, de tendencia, de eventos y de contenidos (Guerra, 2015).
    - Hashtags de marca: También conocidos como corporativos, estos son empleados únicamente con denominaciones que identifican una marca de forma concisa y corta.

  - Hashtags de tendencias: estos están directamente relacionados con la marca y están encaminados a dar apoyo a una nueva tendencia, ya sea promocionando un producto o un servicio.
  - Hashtags de eventos: estos están enfocados a en promocionar un evento en particular o en evaluar el desarrollo del mismo.
  - Hashtags de contenidos: están orientados a conectar a los usuarios con contenido en particular para hacerlo viral.

- Retweets: fueron establecidos para que contenidos publicados sean compartidos con rapidez y básicamente consiste en publicar de nuevo un tweet ya publicado.

- Followers: son aquellos usuarios que siguen a cada otro usuario y deciden ser parte de su red y recibiendo información relacionada con sus tweets y retweets.

API de Twitter

Definida como la interfaz de programación de aplicaciones que, mediante los puntos de conexión brindados a los usuarios permite interactuar y administrar la información compartida. Existen cinco tipos de puntos de conexión: Cuentas y usuarios, tweets y respuestas, herramientas y SDK del editor, y mensajes directos y anuncios (Twitter, 2018).

En busca de ofrecerles a los desarrolladores y analizadores facilidad en sus proyectos se han mejorado los niveles de acceso a los datos, dada la expansión de sus necesidades, ofreciendo una clasificación en torno al seguimiento de productos que están subdivididos en: estándar, investigación académica y negocios (Cairns y Shetty, 2020).

- Estándar: Diseñado para todo tipo de desarrolladores, para los que apenas inician como los que no.

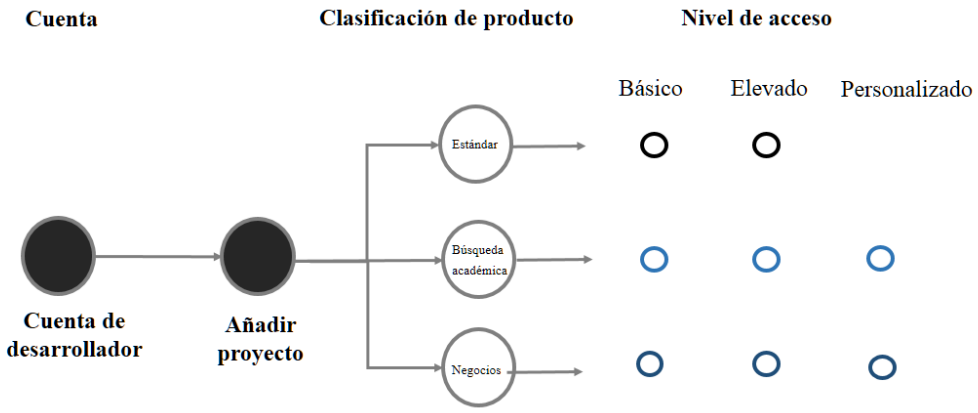
- Investigación académica: los accesos, herramientas y guías contenidos están determinados solo para investigadores calificados y para investigaciones de tipo académico.

- Negocios: solo para uso de diseñadores que plasmas negocios relacionados con esta API, en donde se encuentran los socios oficiales de Twitter y clientes específicos de datos.

A su vez estos se subdividen en niveles de acceso individuales enmarcados en básico, elevado y personalizado. Para mayor contextualización se presenta la figura 5.

Figura 5

Organización API de Twitter



Nota. Adaptado de Blog. Twitter

La API de Twitter requiere de unas credenciales de acceso particulares para cada tipo de desarrollador que se obtienen al ingresar a la cuenta personal una vez aprobado como desarrollador, estas son: API Key, API Secret, Access Token, Access Token Secret (Twitter, 2018).

Minería de datos

Es un proceso enmarcado dentro de un conjunto de técnicas y herramientas empleadas para la exploración, extracción y exposición de información implícita, útil y comprensible extraída de una gran cantidad de datos generalmente almacenados y transformados (Ribas, 2018); descubriendo dentro de la misma conocimiento, relaciones, desviaciones comportamentales, patrones y tendencias mediante la combinación de múltiples técnicas semiautomáticas de inteligencia

artificial, visualización de gráficos, informes y análisis estadísticos, además de bases de datos. Siendo enmarcada por tal motivo dentro de los procesos tecnológicos de análisis de datos en el nivel más alto con el cual es posible mejorar la toma de decisiones y extraer información estratégica (Beltrán, 2014).

Minería de datos a partir de datos de Twitter

A continuación, se presentará un análisis de los métodos y herramientas más de investigación y analítica para recopilación de conocimiento de los datos ofrecidos por Twitter.

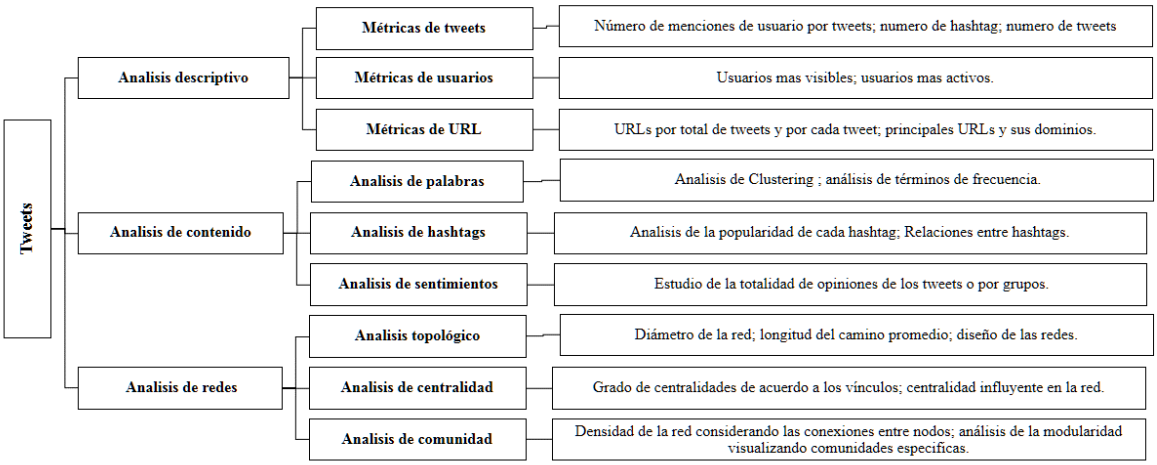
Como primera opción esta la analítica descriptiva, centrada en información estadística que comprende el número de hashtags, de tweets, de usuarios, de retweets, información relevante relacionada con la URL y todo ello que puede llegar a generar un usuario. Con dicha información se extraen numerosas métricas de usuarios o de grupos de usuarios para ser transformadas en conocimiento (Stieglitz y Bruns, 2013).

Seguido de esta tenemos la analítica de contenido, en donde mediante la minería de texto y el procesamiento de lenguaje natural (NLP) se extraen datos de categoría textual. Mediante esta información y el uso particular de algoritmos de aprendizaje automático es posible llegar a identificar temas importantes de discusión en torno a grupos o temas de interés, detección de tópicos en auge y su comportamiento o análisis de sentimientos, análisis de hashtags y análisis de palabras.

Finalmente, está el análisis de red que contempla todo tipo de relaciones entre los usuarios teniendo en cuenta desde el usuario hasta la temática que los relaciona; con esto se llega incluso a identificar núcleos comunitarios inmersos en una red, la capacidad de liderazgo de los usuarios sobre la misma y la extensión de la influencia de la red a nivel territorial, todo esto haciendo uso de medidas de centralidad, análisis topológicos o análisis de la comunidad.

Figura 6

Twitter Analytics



Nota. Adaptado de “Insights form hashtag #supply chain and Twitter Analytics: Considering Twitter and Twitter data for Supply chain practice and research”, de Chae, 2015, p 250.

Minería de **texto**

Contreras (2014) definió esta área del procesamiento automático de datos como un área que, con la ayuda de métodos, y algoritmos procesa de forma casi automática datos estructurados, no estructurados o semiestructurados mediante la cual es posible descubrir tanto nuevo conocimiento como patrones de interés dentro de una recopilación de texto (o documentos), teniendo en cuenta los propósitos particulares del investigador y de la investigación.

Esta Minería se subdivide en dos fases: la de procesamiento y la de descubrimiento; en la inicial, la recopilación de documentos o bien llamados documentos de entrada se transforman y representan en datos semiestructurados o estructurados, según corresponda, y en la siguiente fase, las representaciones son estudiadas para el hallazgo de información valiosa; ambas fases son tan directamente relacionadas entre sí, dependiendo del tipo de representación de datos que se forjan se escogen, posteriormente, las herramientas para el descubrimiento (Contreras , 2016).

Limpieza y Representación de la información

En esta etapa se realiza una limpieza a los documentos que contienen las palabras.

El corpus de texto. Se entiende como cualquier agrupación de información que está compuesta por uno o más documentos, independientemente si estos están escritos por el mismo autor o de la clasificación de los documentos; Atkins, en 1992 considero la existencia de cuatro tipos genéricos de colecciones de texto: Corpus, Archivos, bibliotecas de texto en formato magnético (ETL) y Subcorpus.

Si bien, un archivo es todo el conjunto de texto que es almacenado en formato magnético en un lugar común, si varios de estos están agrupados sin ningún criterio de selección especifico se conforma las ETL, si por el contrario se agrupan de acuerdo con criterios específicos se obtiene el corpus, y si este se subdivide mediante cualquier la ejecución de una consulta se denomina subcorpus.

Técnicas para el preprocesamiento de datos.

Es en esta fase en donde es posible hacer menos tedioso el proceso de representación de texto, un método que se emplea, particularmente es SWR más conocido como palabras de parada y consiste en suprimir todas aquellas palabras que no estén relacionadas directamente con el tema de investigación. A su vez, dos técnicas más usadas en la actualidad para el modelamiento de documentos son el Modelamiento de espacio Vectorial, VSM, y bolsa de palabras, BOM, cualquiera ayuda a convertir los datos en vectores de tipo alfanumérico para después ser tratados mediante operaciones algebraicas lineales (Huan & Xia, 2013).

Con el propósito de desarrollar un análisis preciso al caso de investigación en curso es correcto iniciar eliminando las palabras superfluas o también conocidas como “stop-words” en donde se encuentran los artículos, las preposiciones, las conjunciones y todas aquellas que no proporcionan valor a la investigación.

Transformación de un archivo tipo texto a un archivo tipo numérico

Consiste en transformar un documento integrado por palabras en un vector numérico que pueda ser empleado en la construcción del algoritmo. Para el caso particular de esta investigación se entiende un documento como un tweet.

Es relevante considerar que, si alguna palabra aparece en todos o en casi todos los documentos, deja de ser tan significativa para el análisis, para mejorar esto podemos apoyarnos en la medida numérica Tf-Idf.

Matriz termino documento (DTM)

Es una matriz conformada por filas y columnas, en las filas se ubican los tweets y en las columnas las palabras que integran cada tweet, para calcularla se tienen en cuenta tres tipos de medidas, la frecuencia inversa de termino, la frecuencia de aparición de termino y la frecuencia inversa de aparición de un documento.

La ponderación If-Idf se conoce como "frecuencia de término-frecuencia inversa de documento", con esta se puede determinar la relevancia de la palabra dentro de un documento que este contenido en una colección, este aumenta o disminuye dependiendo de la frecuencia con la que aparece, se calcula de la siguiente manera:

$$\text{Tf-Idf} = \text{Tf}(t, d) * \text{Idf}(t, D)$$

donde:

d: documento.

t: término.

D: número de documentos dentro de la colección.

El factor Tf es frecuencia en la que aparece cualquier término dentro de un documento que se divide en la frecuencia del término más frecuente visto dentro del documento, se obtiene de la suma de todas las veces que aparece un término t dentro de un documento d.

$$\text{Tf}(t, d) = f(t, d) / \max \{f(w, d): w \in d\}$$

El factor Idf o también conocido como la frecuencia inversa del documento, es el coeficiente que se encarga de hallar la capacidad discriminatoria de dicho término del documento en relación con toda la colección, de tal manera que cuanto menor sea el factor Tf, respecto a una alta presencia en el número de documentos, mayor será su factor Idf, este factor es único para cada término dentro de la colección y se calcula así:

$$\text{Idf}(t, D) = \log |D| / |\{d \in D: t \in d\}$$

donde:

$|\{d \in D: t \in d\}|$: número de documentos donde aparece el termino t.

Steamming. Una vez identificada la relevancia de las palabras, se procede a desarrollar el proceso de reducir las palabras a su mínima raíz mediante este método, con este paso es posible recuperar más documentos al repetir nuevamente el cálculo de Tf-Idf.

Lematización. Con este proceso se puede eliminar las pluralidades, las conjugaciones y todas las otras formas flexionadas de una palabra.

Stop Words. También llamadas palabras vacías y son aquellas no tienen ningún contexto semántico dentro de la oración y en la mayoría de los casos ayudan únicamente a conectar las oraciones o las palabras.

Modelamiento de tópicos

Es una técnica integrada dentro de la inteligencia artificial de gran utilidad para la clasificación de textos y documentos que identifica los tópicos (temas) o patrones de mayor relevancia. En donde cada uno de los tópicos se puede definir como $p_i(w)$ en función de la probabilidad que exista de la palabra *w* una vez establecido el tópico *i* (Hammoe L. , 2018). A su vez Hommoe, define al modelamiento de tópicos como “el modelo que permite la clasificación de documentos aún no procesados una vez que el modelo fue entrenado sin pasar por el corpus completo. Esto último es

un atributo muy importante para un modelo que se ejecuta con un stream de documentos (que es lo mismo que un corpus infinito)" (p.5).

El proceso de detección puede variar, no obstante, lo hace por el algoritmo que se usa, teniendo presente que se recolectan los datos, luego se analizan y finalmente se identifican los tópicos; siendo la cantidad de datos ilimitada con altos niveles de ruido (García, 2014).

Técnicas para el modelamiento de tópicos

Dados los avances tanto tecnológicos como informáticos de las últimas décadas han surgido algunas técnicas que permiten identificar tópicos dentro de una secuencia de datos y observar cómo se comportan en el tiempo; los más habituales se presentan a continuación.

Aprendizaje automático. Mediante este tipo de aprendizaje es posible llevar a cabo la construcción de modelos de generalización de comportamientos para posteriormente identificar tópicos, extrayendo automáticamente conocimiento implícito, considerándose en muchos casos como un modelo de la inteligencia artificial de gran importancia (Barrios, 2020).

Aprendizaje supervisado y Aprendizaje no supervisado. Mediante estos dos procesos se agrupan todos los elementos dentro de un espacio n- dimensional. En el aprendizaje supervisado ya se tiene conocimiento previo sobre los tipos de etiquetas de observación, para lo cual, posteriormente, se crean reglas o procedimientos cuyo fin es lograr clasificar objetos que se encuentren implícitas o explicitas dentro de las etiquetas, y que relacionen la entrada con la salida (Collins, 2014). En el aprendizaje no supervisado, por el contrario, se toma como referencia lo observado para parametrizar las condiciones de entrada y que por ende se obtenga una salida análoga, teniendo el enfoque sobre conocimiento conexo con la información disponible (Sancho, 2020).

Este último se divide en dos técnicas: Clustering, en donde se agrupan datos dependiendo de las similitudes y aprendizaje topológico, se emplea la estructura interna con límites del espacio original.

Modelamiento de tópicos LDA

Latent Dirichlet Allocation es un modelo probabilístico generativo de tópicos en donde los documentos del corpus son representados en una combinación aleatoria de los tópicos. En LDA, un documento puede verse como una composición de tópicos, y a cada tópico como una mezcla de palabras permitiendo la "superposición" entre documentos entre sí. (D & single, 2018) , dicho de otra manera, permite que un documento pueda ser parte de varios tópicos, cada uno con un peso distinto. Así mismo, como una mezcla de varias categorías con una distribución a priori de Dirichlet con probabilidades que se intercambian y que abordan distintos métodos matemáticos, para ello este tipo de modelamiento parte del supuesto de que no importa el orden de las palabras ni el lugar en el que se encuentre dentro del documento, siempre están comunicando la misma información (Valvuena & Benitez, 2011).

En este método se puede considerar que al existir dentro de los documentos palabras de menor frecuencia que otras pero que a la vez es común en documentos distintos puede existir la posibilidad de existir un tema en común entre los documentos, permitiendo en última instancia seleccionar la información más relevante con posibilidades mínimas de repetición. De igual forma se asume que en un corpus puede existir una "estructura oculta de tópicos" y una estructura de "variables observadas"; lo que hace este método de modelamiento de tópicos es correlacionar dicha suposición en un "modelo de variables ocultas" en donde se practica y se postula una estructura oculta en los datos observados en el corpus correspondiente, de la cual se aprende usando "probabilidades a posteriori", a lo cual se nombra como modelo probabilístico generativo (Chandia B. , 2016).

Distribución de Dirichlet

Es una distribución de probabilidades multivariada de tipo continuo con un vector alfa (α) de números reales positivos, a su vez es una generalización de la distribución beta (Chandia B. , 2016)..

$$f(x_1, \ldots, x_K; \alpha_1, \ldots, \alpha_K) = \frac{1}{\mathrm{B}(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1},$$

donde $\{x_i\}_{i=1}^{K}$ al k-1 simplex abierto.

Proceso generativo de LDA

Dentro de este proceso se intenta modelar un proceso generativo real por otro que se aproxime, dentro del cual exista la posibilidad de hallar parámetros que se acoplen de la mejor forma posible a los datos (Cenditel, 2016).

El proceso inicia suponiendo que un documento surge como producto de la mezcla de palabras que conforman tópicos con determinada probabilidad, luego de esto se debe:

1. Determinar el vocabulario que se usara.
2. Establecer el número de tópicos o temas (*k*) con su respectiva distribución multimodal de palabras.
3. Establecer el número de documentos (*d*) que tendrá el corpus.
4. Para cada uno de estos documentos (*d*):
    a. Establecer el número de palabras (*v*) que el documento dentro de acuerdo con la distribución.
    b. Elegir la distribución de tópicos θ para el documento, de acuerdo con la distribución de Dirichlet (α) sobre el conjunto de tópicos.
    este vector debe ser de tamaño *k* y define la probabilidad de que un tema ocurra o no en un documento, de manera que los valores pequeños de alfa indican que los documentos están conformados por un pequeño número de temas y por el contrario si el valor es alto su número de temas también lo es.
    c. Para cada una de las palabras (*v*)

i. Se selecciona el tópico Z*v* siguiendo una distribución multinomial (θ)-
ii. Se selecciona una palabra del tópico de acuerdo con la distribución de palabras en el tópico establecida en el paso 2.

Modelamiento de tópicos CTM

Es un modelo jerárquico de colección de documentos presentado por primera vez en el trabajo de Brea & Lafferty (2007) , que permite modelar la correlación dentro de una colección, dicho de otra manera, modelar palabras de cada documento mediante un modelo de mezcla en donde sus componentes son compartidos por todos los documentos de la colección; permitiendo que cada documento muestre tópicos con distintas proporciones en donde se captura la heterogeneidad en datos agrupase que contienen múltiples patrones latentes. La principal característica de este modelo para poder correlacionar los tópicos es que se basa en la distribución normal logística en vez de la de Dirichlet.

La terminología más común empleada es:

1. Palabras y documentos: son las únicas variables aleatorias que se pueden observar y considerar, las palabras están organizadas dentro de los documentos.
2. Temas: un tema β es una distribución que se aplica sobre el vocabulario, un punto en la v-1 simplex. de tal forma que el modelo contiene k temas $\beta_{1:k.}$
3. Asignación de tema: cada palabra es extraída de uno de los k temas y es asociada con la enésima palabra d del documento, $Z_{d,n.}$
4. Proporciones temáticas: cada documento este asociado con un conjunto de proporciones de temas $\theta_{d,}$ que es en un punto en el simplex k-1 por lo tanto, θ_d es una distribución sobre los índices de los temas y refleja las probabilidades con las que las palabras se extraen de cada tema de la colección, normalmente se considera parametrización natural multinomial n= log(θi/θk).

Proceso generativo CTM

1. Extraer el tópico asignado.
2. Extraer la palabra.

3. Mapear la parametrización natural de las proporciones del tema mediante un modelo grafico probabilístico.

El modelo CTM toma como base el modelo de Asignación Latente de Dirichlet (LDA) que al presentar una limitación relacionada con la incapacidad de modelar algunas correlaciones de temas debido a la variabilidad de estos, este modelo al cambiar la distribución por una normal logística ofrece un mejor ajuste de modelamiento de la correlación que LDA para el análisis de una colección de documentos, además de explorar conjuntos de datos que no están estructurados (Lafferti & Blei, 2007).
Desde ambos modelos se considera que los tópicos son grupos de palabras correlacionadas y que dicha correlación se puede observar mediante patrones de co-ocurrencia de las palabras que se encuentran dentro de la colección de documentos (Hammoe L. , 2018).

5. Estimación de parámetros

Tanto la estimación de parámetros con LDA como CTM es llevada a cabo con ayuda de la maximización de la probabilidad de todos los documentos, en particular si el corpus del documento y los parámetros de las distribuciones de Dirichlet y normal ayudan a maximizar la probabilidad de registro de los datos.

Para que se ajuste el modelo LDA o CTM a las matrices de probabilidad, el número de temas (k) se debe fijar a priori, una forma de estimarlo es mediante el muestreo de Gibbs o el muestreo de VEM, sin embargo, para la estimación utilizando el muestreo Gibbs se requiere la especificación de los valores para los parámetros de las distribuciones anteriores (α, μ); Griffiths y Steyvers, en el año 2004 propusieron un valor de 50 /k para los parámetros de estas dos distribuciones en LDA y CTM. Debido a que el número de tópicos es óptimo se determina con ayuda de medidas como por ejemplo la verosimilitud.

6. caso de estudio

A continuación, se presenta una explicación del desarrollo del caso de estudio compuesto por la etapa de extracción de datos de la red social Twitter, la limpieza y representación de la información, el preprocesamiento, procesamiento de los datos y el desarrollo del proceso de minería de texto, para finalmente obtener el conocimiento requerido.

6.1 Extracción de datos de Twitter

Inicialmente se realiza una revisión e identificación de las palabras claves, los hashtags que están vinculadas con la temática en estudio, además de las autoridades de Salud en Colombia que se vinculan a las campañas de promoción y prevención del dengue en Colombia a través de Twitter, entre estas se escoge particularmente: "dengue", arbovirus", "aedes aegpyti", "# hayquecortarlelasalasaldengue", el Ministerio de Salud y protección Social (@MinSaludCol), Instituto Nacional de Salud (@INSColombia) y la Superintendencia Nacional de Salud (@Supersalud).

Para la extracción de los datos se hace uso de la API de Twitter, al generarse una conexión con Python mediante la librería Request; haciendo una selección de datos relacionadas con la búsqueda escogida en un periodo comprendido entre 01/01/2019 desde las 00:00:00 y el 01/01/2021 hasta las 00:00:00 horas, este periodo se ha escogido teniendo en cuenta el año en que se dio anuncio a la campaña "hay que cortarle las alas al dengue" y el tiempo de desarrollo del presente proyecto. Así mismo a través del parámetro place_country se limita la búsqueda a los tweets publicados dentro de Colombia.

La base de datos está compuesta por información que representa cada tweet vinculado con la búsqueda, seleccionando las variables más importantes (fecha de creación del tweet, id del usuario, tweet, cantidad de retweets, hashtags, lugar de publicación del tweet y país de publicación del tweet), esta información se presenta en formato de codificación UTF-8. Se obtuvo un total de 1989 tweets, almacenados en formato csv, obteniéndose una mayor cantidad de información de la búsqueda con la palabra "dengue".

6.2 Limpieza y representación de la información

Esta etapa es fundamental para optima aplicación de los métodos de analítica de datos; se inicia con la limpieza de datos.

Limpieza de datos

Para el desarrollo de este proceso se copilan los datos dentro de una única base de datos en formato csv., para posteriormente iniciar el proceso de limpieza apoyado en distintos paquetes entre ellos tm (text Mining), wordcloud, qdap, stringr, entre otros del software R.

En esta etapa se selecciona y se realiza lo siguiente:

- Se trasforma todas las letras mayúsculas en minúscula con el fin de hacer más uniforme el proceso de visualización de contenido.
- Se eliminan tanto las palabras que se catalogan como artículos y conjunciones debido a que son palabras que carecen de contenido semántico y su aporte se puede considerar nulo.
- Se eliminan las abreviaturas.
- Se elimina todo tipo de direcciones de correos institucionales, debido a que no es contenido que sirva para el proceso de generación de conocimiento del objeto de estudio.
- Se elimina las Urls de direccionamiento; en su mayoría se encontró que la mayoría de tweets contiene este tipo de información con la cual el usuario pretende ampliar el contenido de su mensaje.
- Se elimina del mismo modo las menciones de usuarios, pues no se considera información relevante dentro del contenido de los tweets.
- Se elimina todo tipo de Imágenes, caracteres y videos, pues esto lo usan comúnmente los usuarios para hacer más uniforme y entendible su mensaje, de tal manera que con el mismo se puede mejorar la visualización de quien lo observa.
- Se elimina símbolos y signos de puntuación pues únicamente representan ruido dentro del contenido del tweet.

- Se elimina las numeraciones, dado que en particular señalan las fechas o las horas, información que no es relevante para el desarrollo de este proyecto.
- Se elimina de la misma manera los caracteres no alfanuméricos, porque representan ruido para el contenido a evaluar.
- Se elimina las palabras vacías, es decir palabras que carecen de contenido semántico o que no aportan de manera significativa al contenido del documento.
- Se elimina las filas vacías, estas surgen una vez que se realiza todas las eliminaciones mencionadas anteriormente, y se dan porque no contienen información dado que su único contenido se encontraba enmarcado por Urls, conectores, abreviaturas, etc.
- Se reducen las palabras a su raíz,

La base de datos, luego de finalizado todo el proceso de limpieza de esta base de datos, termina un grupo base para continuar con el proceso de representación compuesta por 1862 tweets, correspondiente al 93 % del total recolectado.

6.3 Representación

Para esta etapa se requiere como entrada únicamente la variable tweet que está dentro del archivo csv obtenido de la etapa de limpieza, la cual es procesada con ayuda de la función "vectorsource" del software R, con el propósito de identificar cada tweet como un documento que luego será representado dentro de un espacio vectorial en donde cada palabra estará separada de forma independiente y se conoce como variable, que seguidamente integrará una colección de documentos o más conocido como corpus.

Seguido se procesa cada corpus y se convierte a un formato integrador de una matriz termino documento, dado que esta es la entrada que permitirá realizar el análisis pertinente de minería de datos. En esta matriz se asigna un peso de termino frecuencia a cada palabra del corpus, con esta misma se calcula posteriormente el Tf- Idf en donde es posible conocer la importancia que tiene un término dentro del documento dependiendo del número de veces que se presenta en el mismo.

Haciendo uso de la función Term- Document Matrix de la librería tm (software R) se obtuvieron los resultados expuestos en la tabla 5.

En esta se detalla que la matriz Término- Documento está compuesta en total por 4248 términos encontrados en 1862 documentos, ubicados todos ellos en las filas de la matriz y 1862 documentos (tweets), en las columnas; para dar un total de 79.099.776 entradas no dispersa, se ha encontrado dentro de la totalidad de documentos que la longitud máxima que puede tener cada termino es de 16 letras.

Tabla 5

Características principales de la TDM

Document- Term Matrix	**Numero de documentos: 1862/términos:4248**
Entradas no dispersas	79099776
Dispersión	100%
Longitud máxima de cada término	16

Dentro de los tweets analizados para este caso de investigación, para el término “dengue” el valor asignado dentro de la matriz Tf-Idf , entendido como la frecuencia de ocurrencia de este término dentro de colección de tweets, tiene un valor de 37,5079, de tal manera que este se considera más relevante que el término “aedesaegypti”, con un puntaje de 32,5413 y menor respecto a términos como “Colombia” ,“mosquito”, “zancudo”, “enfermedad”, “fumigación”, “epidemia” y “control” todas ellas con ponderaciones entre 70 y 40 , es decir, al compararse estos resultados la palabra dengue es más común dentro de la colección total de tweets que las palabras “pandemia”, “miedo”, “combatir”, “vacuna” y “contagio”.

Se continua con la creación de una nube de palabras, medio que permite representar el Tf-Idf de aquellas palabras que tienen mayor ocurrencia dentro del grupo de tweets. Esta se puede visualizar en la figura 7; en la misma se representa la importancia de cada termino de acuerdo al tamaño de la fuente y el color, y si su valor dentro de los términos disminuye su prominencia también lo hace. De tal forma que, en este caso para un total de 4248 términos, los de mayor

relevancia son: muerte, caso, prevención, Colombia, seguido por zancudo, mosquito, salud, enfermedad, control, dengue, entre otras; dentro del rango de menos importantes se encuentran algunas palabras como arbovirus, respirar, diariamente, aislamiento, inmunosuprimido.

Figura 7

Nube de palabras

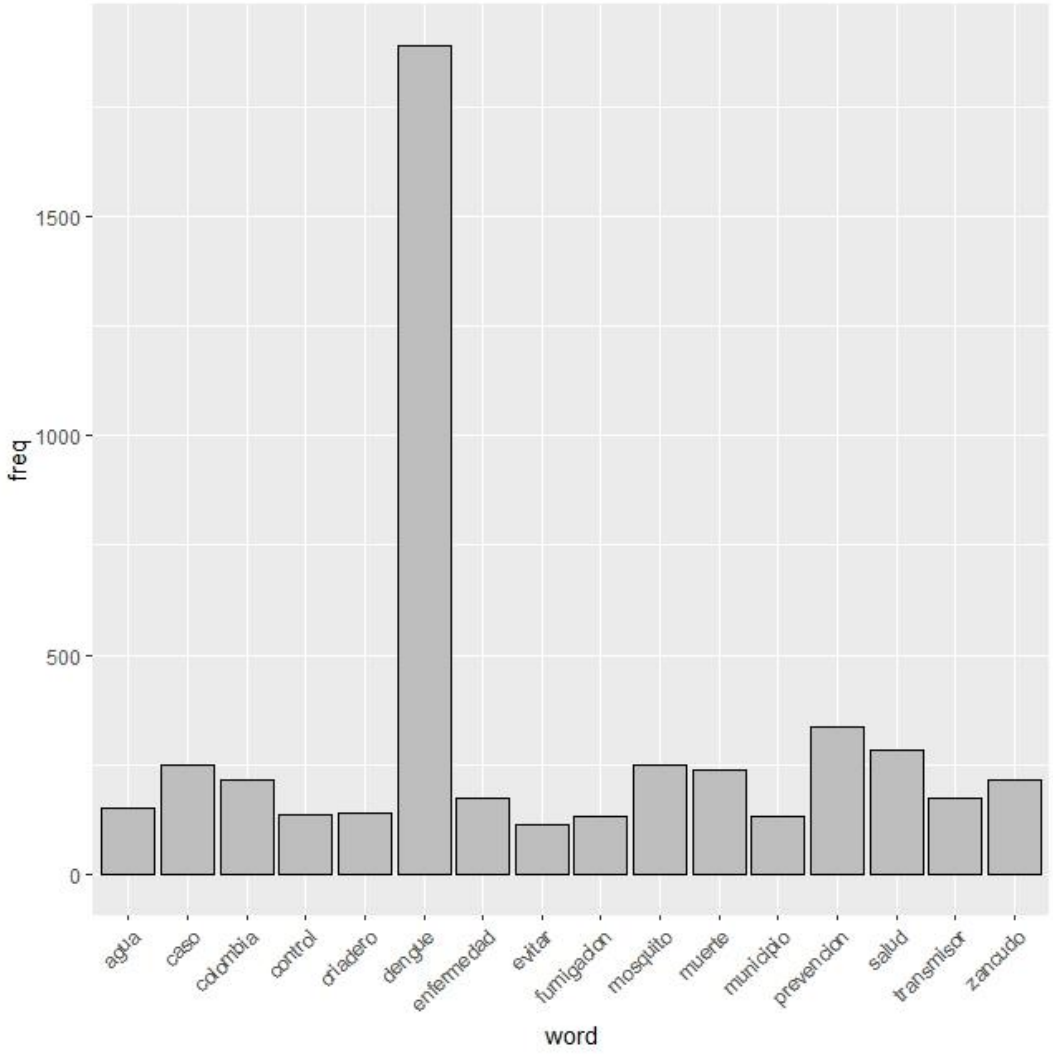
Nota: creada en Rstudio.

6.4 Análisis de las palabras más frecuentes

A continuación, en la figura 8, se presenta un análisis grafico mediante un diagrama de barras en donde se agrupan las palabras más frecuentes y el respectivo número de veces que se repite dentro de la colección para el periodo de tiempo estudiado, además, se determina que las palabras que más predominan son “dengue” con una frecuencia de aparición dentro de la base de tweets de 1891 veces, a esta le siguen “prevención” con 335, “salud” con 281, “mosquito” con 250 y “caso” con 248 veces.

Figura 8

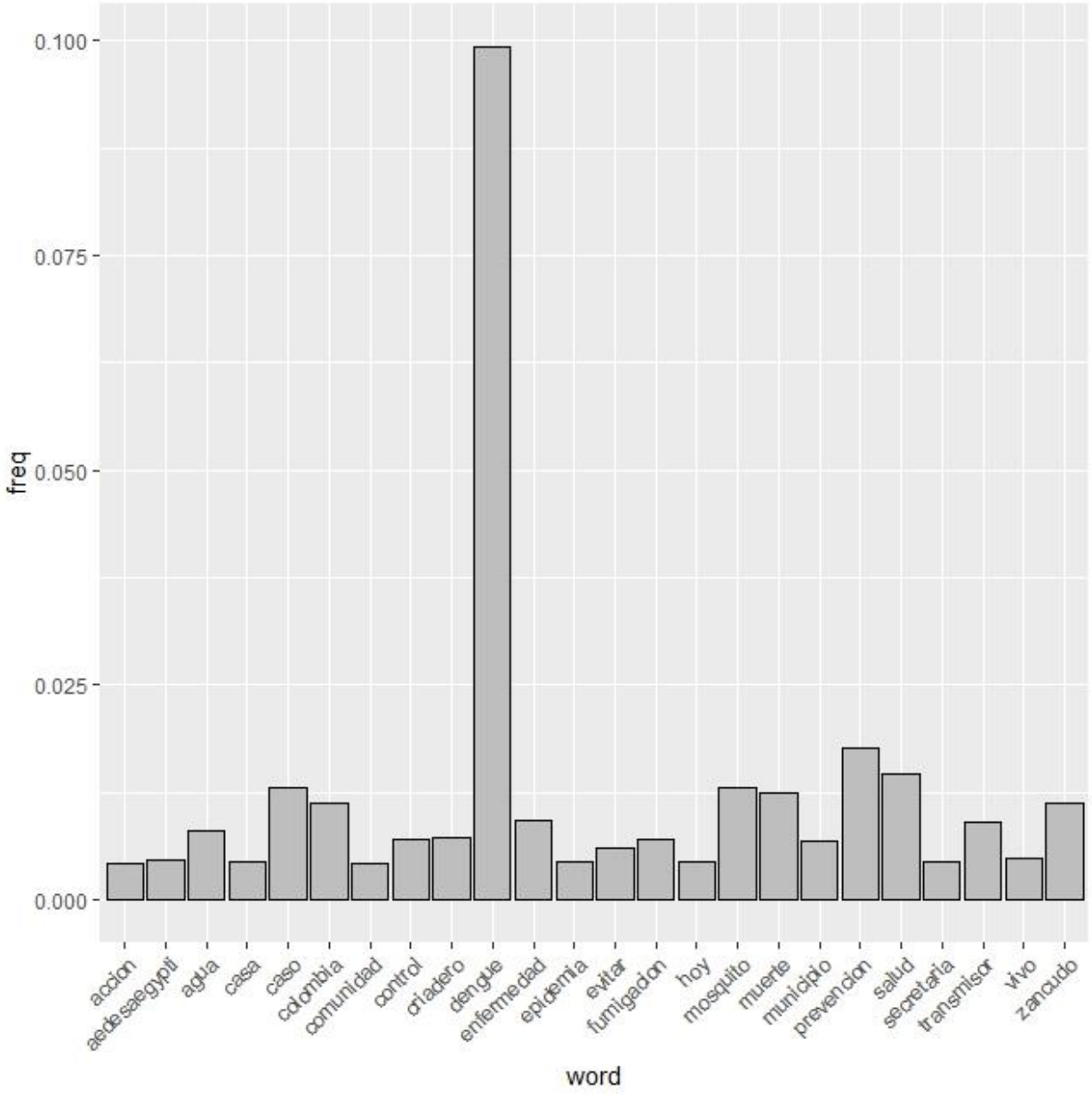
Frecuencia de palabras en el tiempo estudiado



En la figura 9, se muestra un diagrama de barras con las frecuencias relativas de algunas palabras. En donde de la totalidad de repeticiones de los términos el 9,9 % corresponde a la palabra “dengue”, el 1,7 % a la palabra “prevencion”, el 1,4 % a “salud”.

Figura 9

Frecuencia relativa de los términos más comunes



Asimismo, como se observa en las figuras 10 y 11, durante el año 2019 las palabras más frecuentes fueron: “dengue” con una frecuencia de 379 repeticiones, que corresponde al 20% de la totalidad del periodo estudiado en el proyecto, seguido de “caso” con 82 (33%) y “mosquito” con un total de 76 (30%); para el año 2020 “dengue” continúo siendo la palabra de mayor ocurrencia con 1482 (80%), sin embargo, esta va seguida por la palabra “prevención” con 258

(77%) y "muerte" con 215 (90%); cabe resaltar que la palabra "caso" aumento a 161 y "mosquito" a 159.

Figura 10

Frecuencia de palabras año 2019

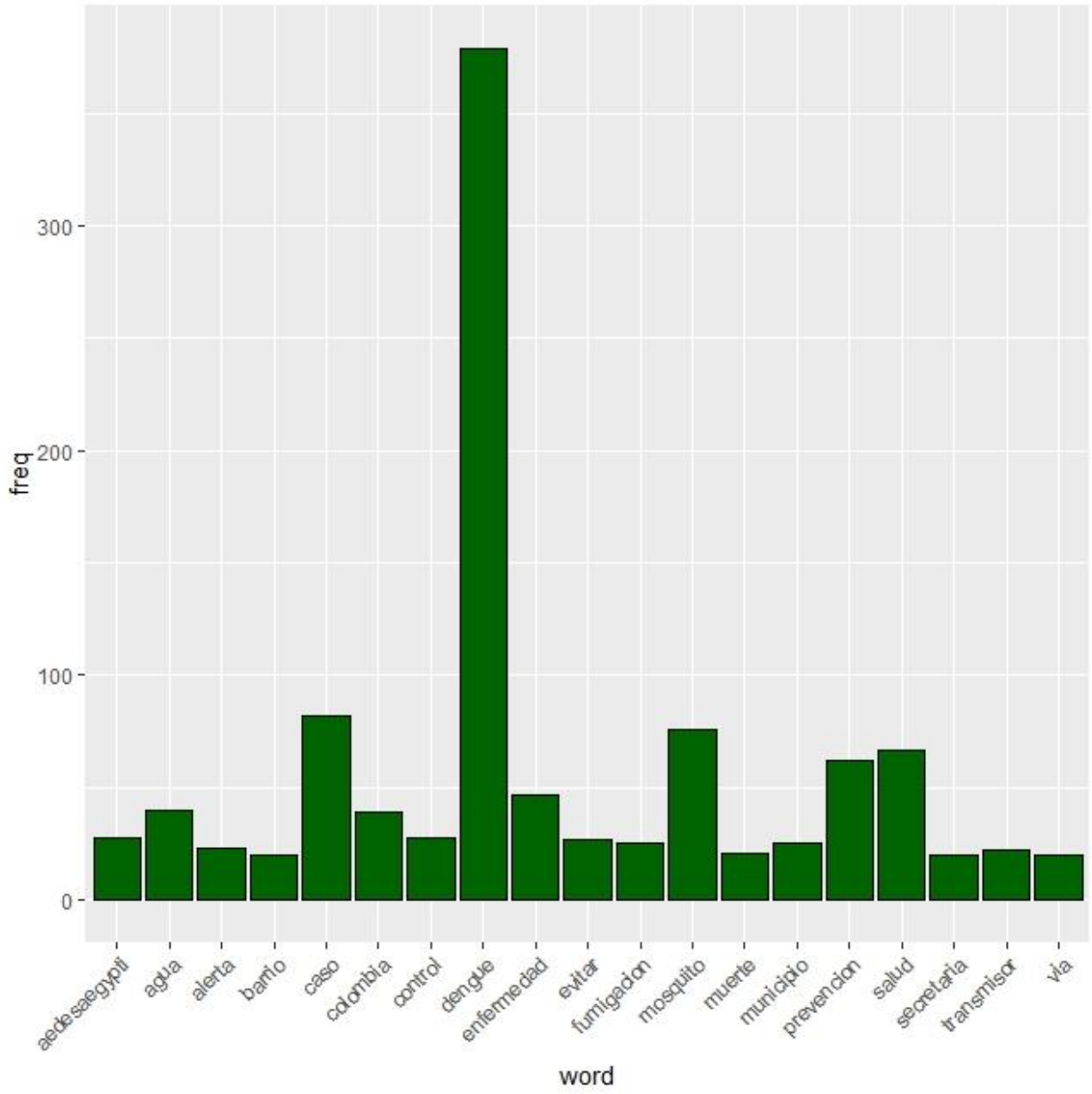
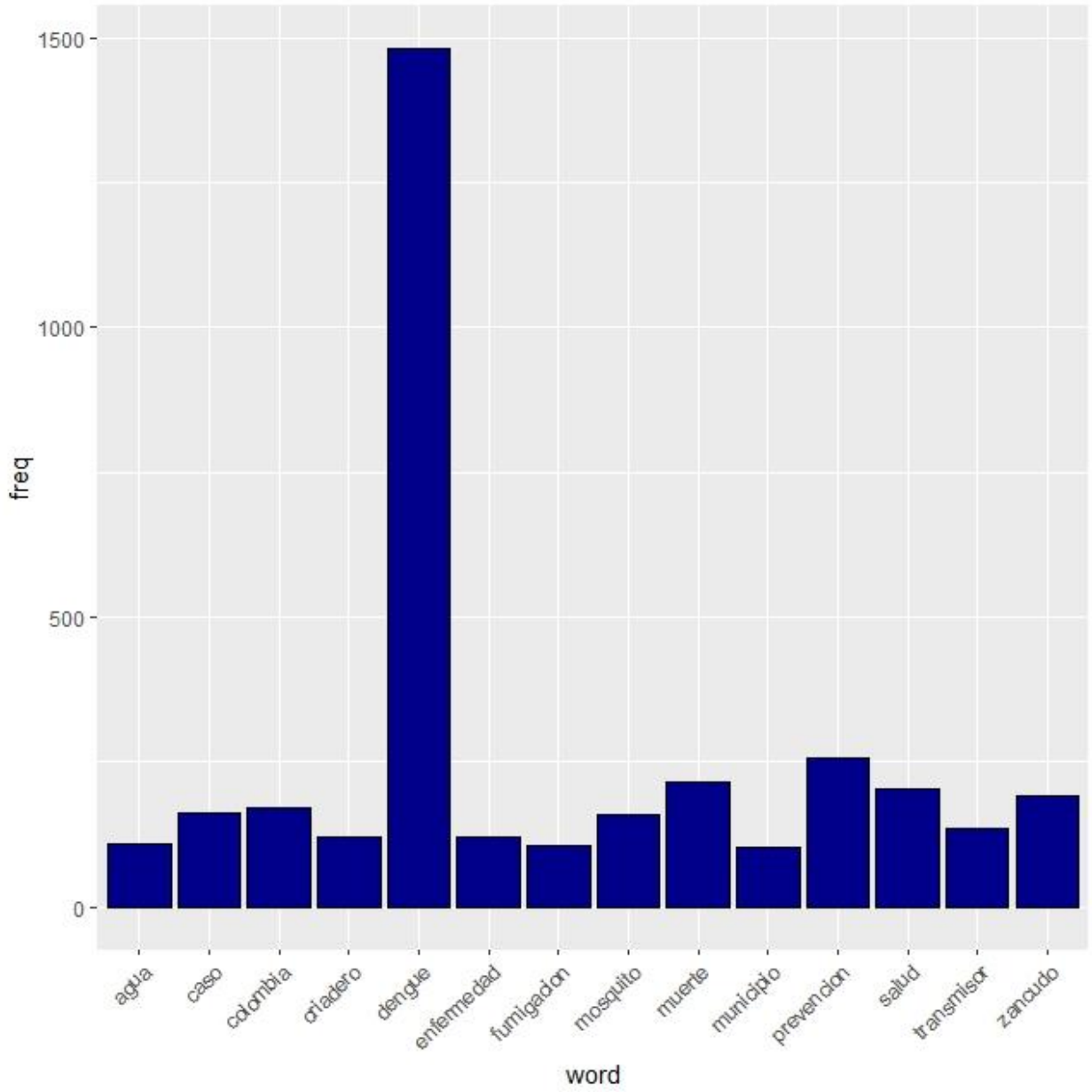


Figura 11

Frecuencia de palabras año 2020

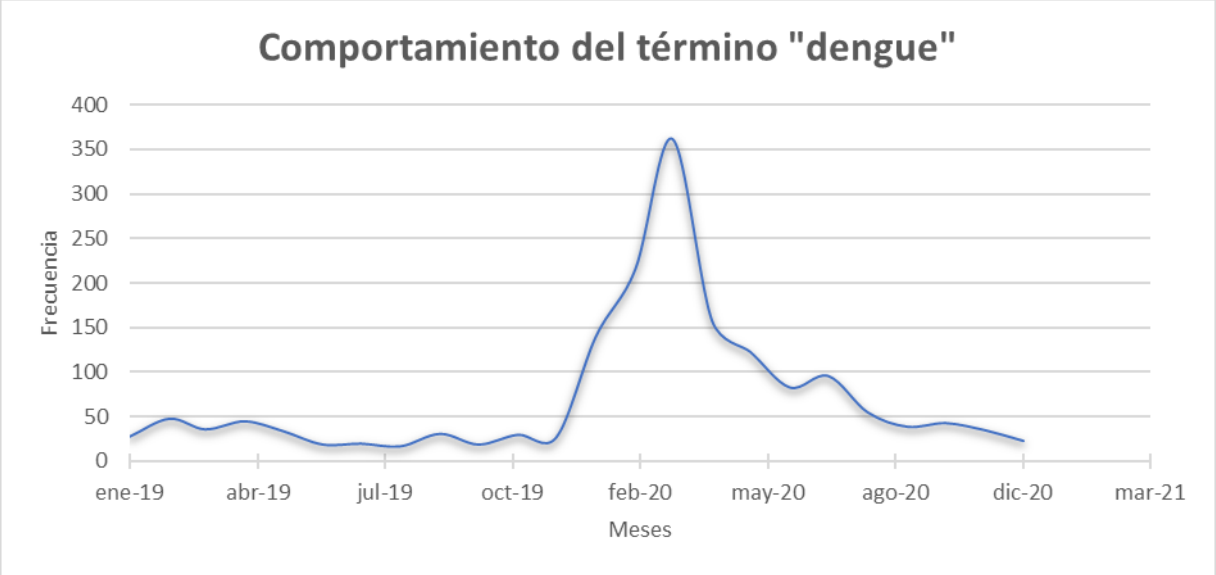


Ahora bien, en la figura 12, se presenta el comportamiento de la palabra “dengue” por mes. A partir de esta se puede concluir que el pico más alto de ocurrencia está en el mes de marzo de 2020,

seguido de los meses de febrero y abril con frecuencias predominantes, respecto a los demás meses; el periodo en el que menos predomina el uso de esta palabra esta entre inicio del mes de junio hasta finales de diciembre del 2019.

Figura 12

Comportamiento de la palabra dengue, mes a mes



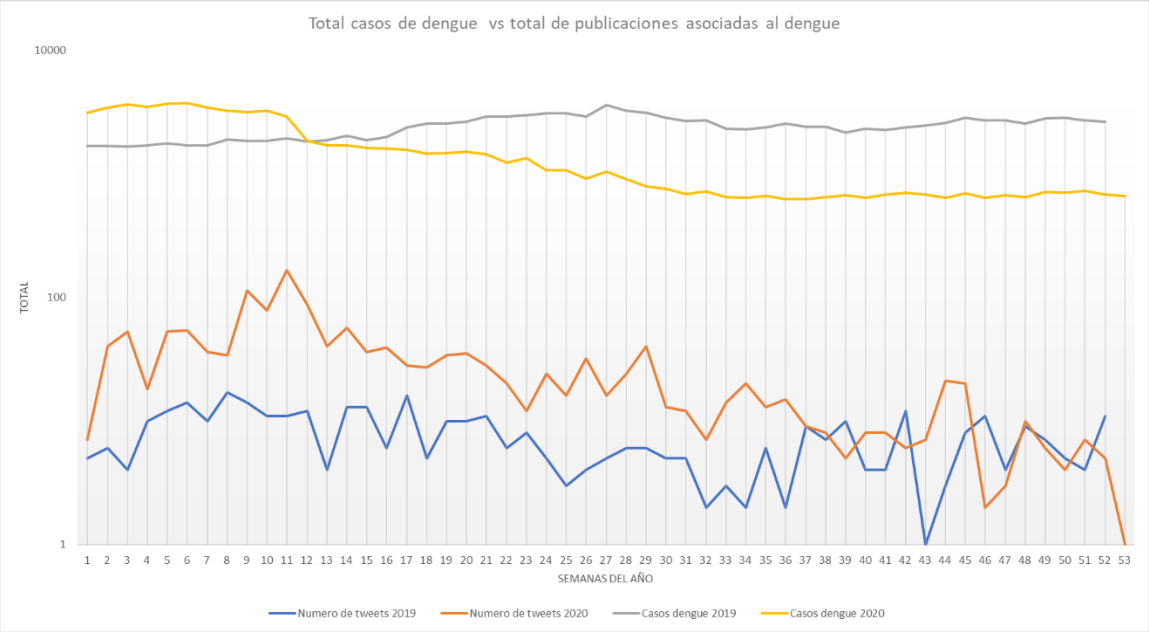
6.5 Analisis de las publicaciones versus los casos de dengue en Colombia del año 2019 al año 2020

En la figura 13, se presenta una comparación entre el total de publicaciones por semana versus el total de casos registrados en cada una de estas; este último se construyó con información presentada por el Ministerio de Salud de Colombia entorno a información epidemiológica (MinSalud, 2021). Se muestra un aumento de los casos de dengue del año 2020 respecto al año 2019 hasta la semana 12, de ahí en adelante hasta la última semana es superior este último; si se estudia la tendencia de las publicaciones hasta la semana 11, para el 2020 y 10 para el 2019 es más creciente que decreciente; se conserva una superioridad de publicaciones de este último año sobre el anterior hasta la semana 36, que son iguales y posteriormente a esta, en la mayoría de las

semanas sucede lo contrario. En general, se concluye que la tendencia de publicaciones presenta una alta fluctuación entre una semana u otra, caso contrario para los casos registrados de dengue, en donde luego de la semana 12, para el 2019 empieza a crecer y para el 2020 a decrecer.

Figura 13

Publicaciones vs casos de dengue



Para complementar, se ha calculado la correlación entre la cantidad de publicaciones tweets de cada semana para los dos años y el número de casos de dengue del 2019 y 2020 agrupados en uno solo por cada semana, para ello se usó la función “coef.de.correl” de Excel, este dio como resultado un coeficiente de correlación con un valor de 0,549516497; de tal manera que es posible concluir que entre ambas variables estudiadas existe una relación de tipo directo, de tal manera que al aumentar la magnitud de una, a su vez aumenta la de la otra y viceversa.

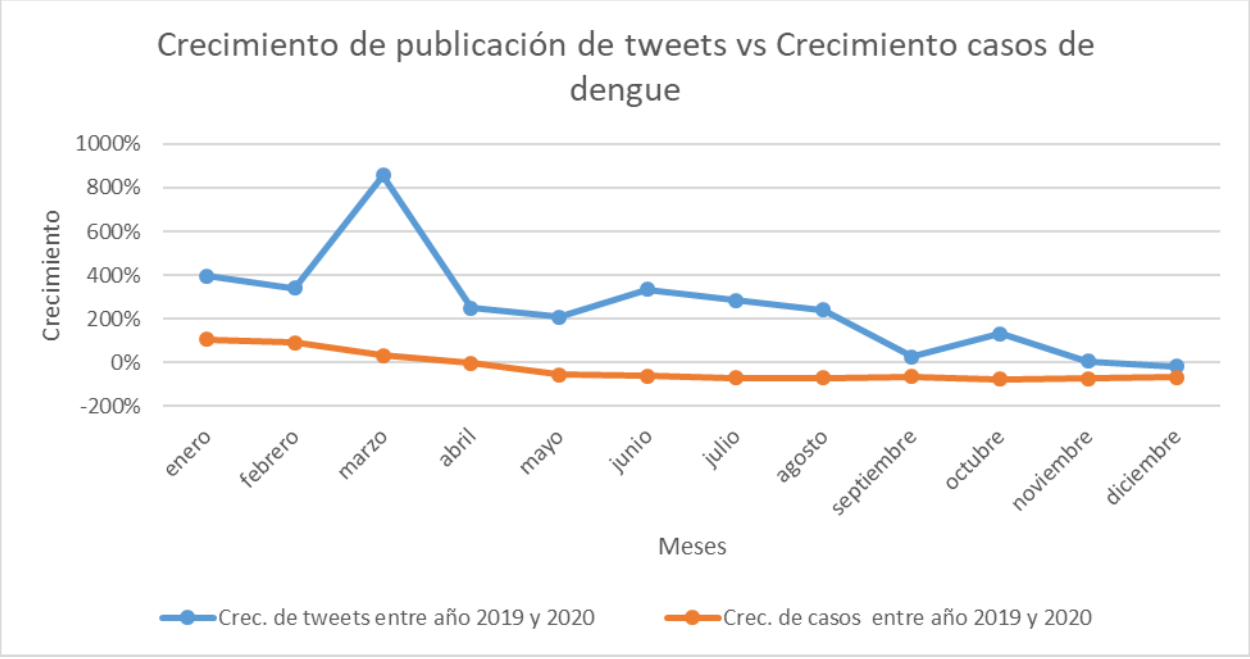
Adicionalmente, en la figura 14, se muestra el porcentaje de crecimiento que tiene la cantidad de publicaciones del año 2020 respecto al año anterior, al mismo tiempo que se presenta el crecimiento de los casos de dengue.

A partir de esta, se puede analizar que en total las publicaciones crecieron en un 275%, entendiendo esto como 392 tweets publicados en el transcurso del año 2019 y para el siguiente año 1470 dentro del territorio nacional. Caso contrario, ocurre con los casos de dengue registrados, en donde se evidencia una disminución; mientras que en el primer año se registran 125.111 casos, para el año dos, 77.348 casos, representando una disminución del 38 %.

Es importante señalar que el comportamiento de ambas situaciones es similar, mientras para la situación de casos registrados la tendencia durante los meses analizados es a decrecer para el caso de las publicaciones es fluctuante entre enero a junio teniendo un crecimiento predominante en el mes de marzo, para los meses siguientes su tendencia es de decrecimiento.

Figura 14

Comparación de crecimientos



En la misma se evidencia que si bien a pesar de ocurrir un decrecimiento de publicaciones, únicamente en el mes de diciembre del año 2020 son menores a su antecesor; por el contrario,

si se estudia la situación de los casos, el mes de enero fue el que presento un crecimiento predominante con un 104% de crecimiento (de 8479 casos registrados en 2019 aumento a 17.322 casos), seguido de los meses de febrero con un 90 % y marzo con 30%; para los meses siguientes esta situación es contraria, el mes en donde menos crecimiento se evidencio fue en octubre con -78% (pasando de 11.959 casos registrados a 2690) .

Analisis de crecimiento de las publicaciones

Al analizar todos los meses, marzo fue en donde más aumentó la cantidad de publicaciones entre un año y el otro, pasando de 40 a 343 tweets, del mismo modo fue el mes en el cual se publicaron más tweets; caso contrario diciembre, con un crecimiento negativo, la población paso de realizar 29 a 22 publicaciones al mes. En la tabla 6, se muestran los meses con más alto crecimiento dentro del periodo de tiempo analizado, siendo el más bajo junio.

Tabla 6

Meses con más alto crecimiento

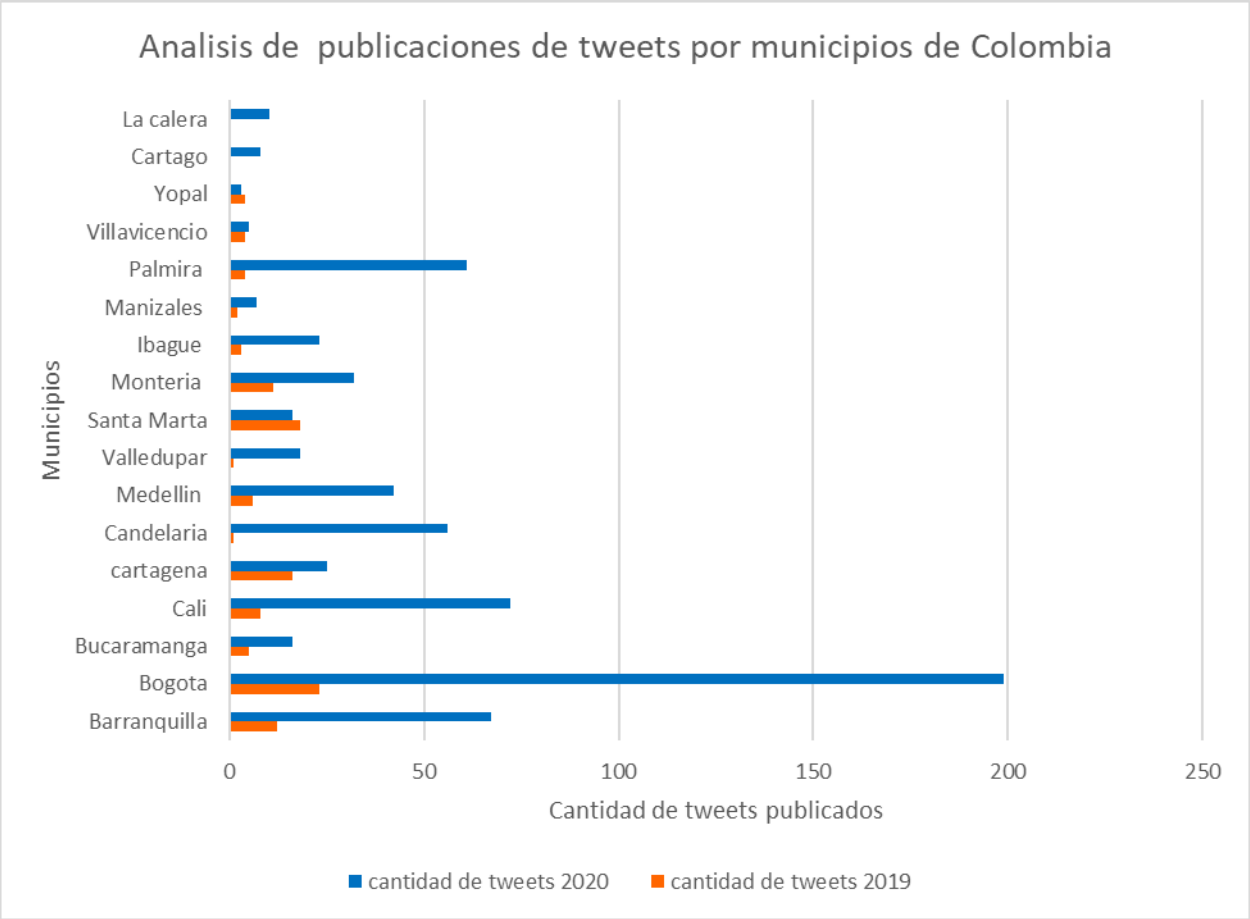
Mes	**Porcentaje de crecimiento de publicaciones 2020 respecto 2019**
Enero	397%
Febrero	340%
Marzo	858%
Junio	333%

Cantidad de tweets por cada ciudad

A continuación, se observa la cantidad de tweets que se publican por cada ciudad durante estos meses.

Figura 15

Analisis de publicaciones por municipio



Las ciudades en las que para estos meses estudiados se presentó la mayor cantidad de publicaciones fueron Bogotá con 222, seguido de Cali con 80 y Barranquilla con 79. Situación que para inicios del año 2020, preocupaba al país, pues el Ministerio de Salud consideró un aumento superior al 90 % de los casos de dengue respecto al año anterior, situación que causo inquietud y tras la cual se dio la medida de alerta roja para todo los departamentos a los que pertenecen las ciudades mencionadas anteriormente, además de otros como Tolima, Huila, Cesar y Santander (Aguirre, 2020).

6.6 Minería de texto

Teniendo de base la revisión literaria realizada y los objetivos propuestos que permiten el desarrollo de este proyecto los algoritmos que se usan son el modelo LDA (Latent Dirichlet Allocation) y CTM (correlated Topic Models) para identificar a través de una comparación de las medidas de máxima verosimilitud entre tópicos cual es el mejor modelo. Para el respectivo análisis de los dos modelos propuestos se siguen los siguientes pasos:

Ingreso de Bases de datos

Como base de datos para la etapa inicial del proceso de minería de texto se inició con un documento de entrada que es una Matriz termino-Documento compuesta por 4248 términos, es importante aclarar que este es un documento de salida obtenido de la etapa de preprocesamiento de datos. Esta matriz se procesa y se convierte en una matriz documento-termino, para poder usar el paquete Topic Models de R.

Selección del mejor algoritmo de modelamiento de tópicos y del k óptimo

Para desarrollar esto se escogió el método de muestreo VEM, método que es común entre los dos algoritmos. Estos dos métodos se compararon por medio de la medida de máxima verosimilitud (Loglikehood), en la figura 16, se grafica la variación de esta medida en cada uno de los algoritmos.

El límite de los tópicos para la construcción de esta grafica es de 32, debido a que las iteraciones del modelo CTM convergen en este punto, al hallar la medida para el siguiente valor el algoritmo en R se prolonga iterando de manera continua sin detenerse. El tiempo que tarda en obtener los resultados para el modelo LDA es de 12 horas y el de CTM se extiende a 96 horas.

En la gráfica, se puede observar que desde el tópico 12 al tópico 13 en el caso del modelo LDA el valor referente da un salto significativo, comparativamente, mientras que el modelo CTM permanece en crecimiento continuo y sin tanta variabilidad; no obstante, desde el tópico 13 hasta el 18 los valores que toma el modelo CTM tienden a ser más negativos que el modelo LDA, en los siguientes tópicos y hasta finalizar esta etapa sucede lo inverso.

Para el análisis de esta medida se prefiere el valor cuya medida de Loglikehood es más pequeña debido a que es ahí en donde la variabilidad del modelo es menor, pues a medida que este valor es mayor se considera que más variabilidad se deja explicar dentro del modelo en estudio (Field, Miles, & Field, 2012).

De esta forma y en concordancia con lo anteriormente estipulado hasta el tópico 12 el modelo LDA es el mejor, en el lapso de 13 y 18, lo es el modelo CTM y de este último en adelante continúa siendo LDA; de tal manera que se concluye que el mejor modelo dentro de este rango de tópicos es el LDA.

En continuidad, se compara los valores de esta medida en LDA para seleccionar el k optimo; si bien, entre el tópico 12 y el tópico 13, particularmente ocurre un salto significativo para el modelo LDA pues esta medida paso de un valor de -128,710 $*10^6$ (tópico 12) a -117,400 $*10^6$ (tópico 13) pasando de una tendencia al aumento de la negatividad a un lapso en donde ocurre lo contrario para volver a retomar en el tópico 18 su comportamiento anterior; situación que no ocurre con CTM, pues durante todo el trascurso de la gráfica su tendencia es a disminuir su negatividad, razón por la cual también se estudia la variabilidad entre ambos modelos en este punto de medida de Loglikehood, es ahí en donde para LDA esta es de -128,710 $*10^6$ y para CTM, es -119,498 $*10^6$, siendo sobresaliente nuevamente LDA.

Otro aspecto a considerar es que el procesamiento de los datos en el software R; con el modelo CTM fue más inestable que para LDA, como se muestra en la tabla 7, pues luego de 32 ya no se generaba el análisis de agrupación de los tópicos, situaciones significativas que permiten ratificar

la decisión de escoger el modelo LDA como el mejor para el desarrollo del presente caso de estudio, con un k óptimo de 12 tópicos.

Figura 16

Loglikehood del modelo LDA y CTM. Adaptado de Rstudio, versión 3.5

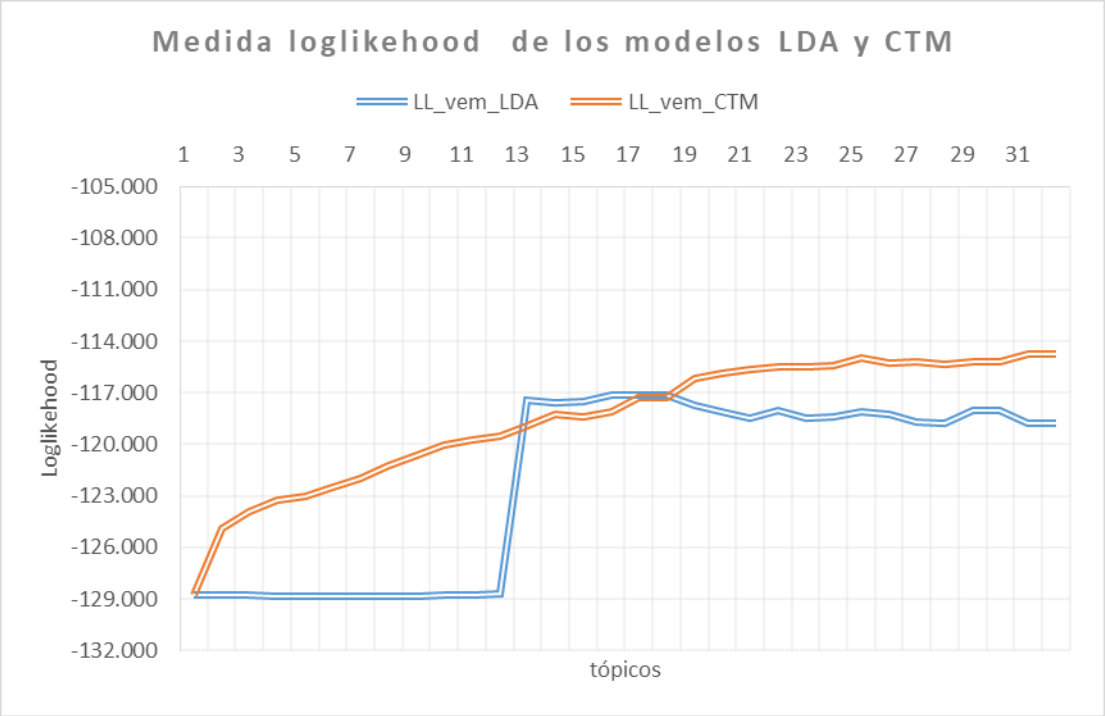


Tabla 7

Número óptimo de tópicos para el modelo LDA

Modelo	Loglikehood	Tiempo de procesamiento

Vem_LDA	-128,710 $*10^6$	12 horas
Vem_CTM	-119,498 $*10^6$	96 horas (4 días)

6.7 Analisis de los resultados

El mejor algoritmo para el modelamiento de los tópicos de este caso ha sido el modelo LDA, de este modelo se obtiene como salida una matriz en donde se describe cada una de las probabilidades a-posteriori de que un término sea seleccionado dentro de un tópico en específico ($\beta_{1:k,1:V}$).

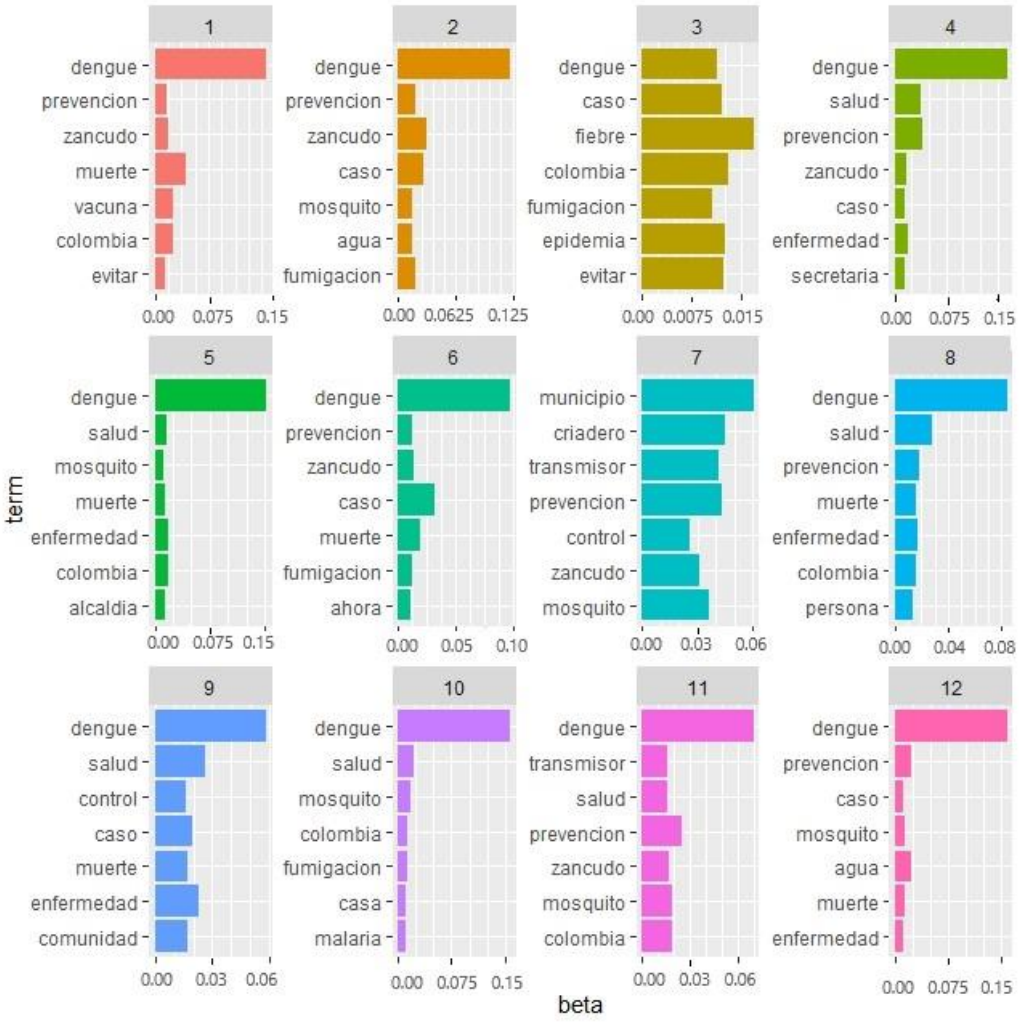
A continuación, se analizan los tópicos obtenidos y los tweets más característicos que se relacionan con cada uno.

Etapa de asignación de tópicos

En esta etapa, se agrupan los 7 términos que presentan mayor probabilidad de pertenecer a cada uno de los 12 tópicos; en total se tienen en cuenta para esta representación 42 términos de los 4248. En la figura 17, se pueden visualizar las agrupaciones respectivas, en donde “beta” es cada uno de los tópicos escogidos, y en los mismos se detalla la probabilidad de que cada término pertenezca a cada tópico “beta”, estas probabilidades oscilan entre 0,59% y 11,3%.

Figura 17

Tópicos con su integración de términos con mayor probabilidad



En continuidad, en la tabla 8, se determina cual es el tema o cuales son los temas que más predominan en cada uno de los tópicos y posteriormente se etiqueta cada uno.

Tabla 8

Tema general de cada tópico

Tópico	Etiqueta
1	Programas de prevención y vacunación contra el dengue en Colombia.
2	Fumigación y el control para la prevención de la picadura por mosquitos del dengue.
3	Casos de dengue en Colombia con presencia particular de síntomas de dolor y fiebre.
4	Acciones de la secretaria de salud para la prevención de casos y muerte por dengue.

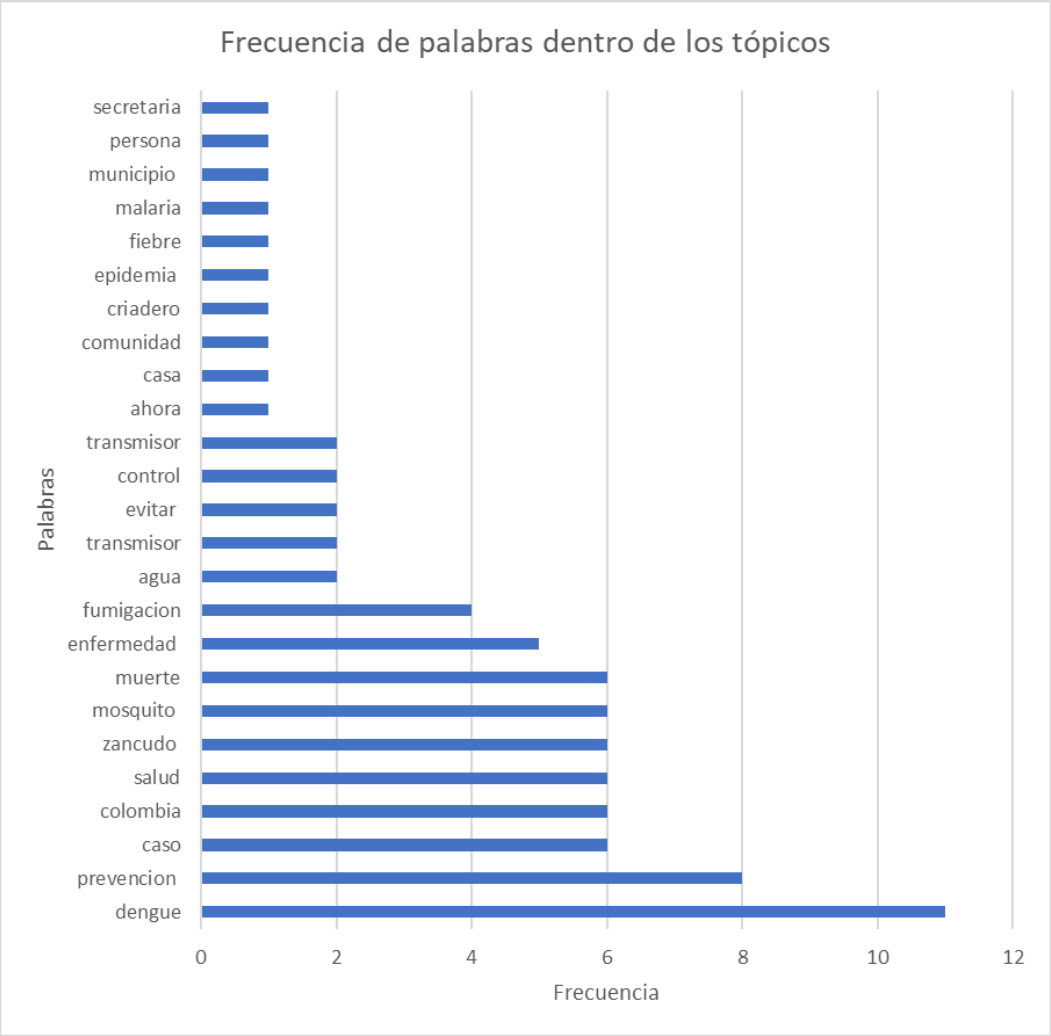
5	Actividades adelantadas por la alcaldía para evitar enfermedad y/o muerte por picadura del mosquito del dengue en Colombia.
6	Jornadas de fumigación departamental para la prevención de los casos de muerte por dengue.
7	Jornadas de prevención/ control/ eliminación de criaderos del mosquito/zancudo transmisor del dengue.
8	Prevención en Colombia para los casos de muertes y/o enfermedades por picadura de mosquito del dengue.
9	Control en la comunidad y cuidado del agua para evitar enfermedades o muerte por casos de dengue por aedesaegypti.
10	La fumigación en las casas para evitar la transmisión del dengue por picadura de mosquitos.
11	Colombia previene el contagio por dengue transmitido por zancudos o mosquitos.
12	Prevención de los casos de dengue en Colombia.

De esto, se obtiene que más del 80 % de los tópicos están proporcionando información relacionada con los procesos de prevención y mitigación del contagio por este virus para el periodo comprendido entre 2019 y 2020, con divergencia entre estos, en aspectos como: el tipo de enfoque: prevencion o el control; la acción que se desarrolla: fumigación o vacunación ; el área que en la que se enfoca: casas, departamental, o a nivel nacional; el aspecto que se quiere mitigar: la picadura, el contagio o la muerte.

Una observación importante sobre las palabras que componen cada tópico, algunas de estas están presentes en más de dos tópicos del total de los analizados, por ejemplo, la palabra “dengue” está presente en 11 de los 12 tópicos a la vez; “prevencion” 8; 6 de estas hacen parte de la agrupación de 6 tópicos y así sucesivamente como se muestra en la figura 18; información que concuerda con los análisis descriptivos de frecuencia expuestos anteriormente. Si bien, esto es una particularidad predominante del modelo LDA respecto a otros modelos de análisis de tópicos, pues permite la coexistencia de palabras en más de un tópico (Blei, Ng, & Jordan, 2003).

Figura 18

Palabras más predominantes dentro de la agrupación de los tópicos



Analisis general de tópicos

En esta etapa, se analizó los tópicos que son más o menos twitteados a lo largo del periodo de estudio, la probabilidad de vinculo es superior o igual al 7%; en la figura 19, se muestra que los tópicos con la mayor y menor cantidad de tweets asociados son el tópico 6 con 1826 tweets y el tópico 4 con 1750 tweets, respectivamente.

Figura 19

Cantidad de tweets por tópico

Los tópicos a los que se les asocian mayor cantidad de tweets, en orden descendente, son: tópico 6, tópico 10, tópico 1 y tópico 12; los que están asociados con la menor cantidad son tópico 2, tópico 9 y Tópico 4. Para mayor contextualización. En la tabla 9, se puede observar un ejemplo, con las probabilidades asignadas a cada tweet que pertenece al tópico 6 (Jornadas de fumigación departamental para la prevención de los casos de muerte por dengue) y su respectiva probabilidad. En la misma se observa que independientemente de que los tweets sean publicados en fechas distintas, si los tweets presentan coincidencia de términos se les asignan el mismo valor de probabilidad.

Tabla 9

Probabilidades asociadas a los tweets relacionados con el tópico 6

Fecha de publicación	**Lugar de publicación**	**tweets asociado al tópico 6**	**Probabilidad**
13/03/2020 19:22	Girardot	@Carlos Correa @joco quintero Lo coronavirus siempre han existido. Lo q esta una mutación la bautizaron coronavirus it19.se llama coronavirus salen largo como la del coronavirus. toda una ternura asesina. Pero como hay cuidado .como el dengue	0,109573

		publicitado y su fumigación https://t.co/InoSaV7z0n	
10/06/2020 15:04	Palmira	@Cali La toma de temperatura un error. enfermedad la generan. como el dengue. malaria. la constitución corporal del hombre y la mujer son diferente. en caso. Especial la mujer de acuerdo su ciclo hormonal su temperatura cambia diariamente no por el dengue ese.	0,109054
14/01/2020 14:45	Cartagena	@elsanoguerabaq Atencion el dengue haciendo estragos. comentan una niña del municipio de Malambo se encuentra internada (UCI) en un centro asistencial. hay agilizar prevención y fumigación para evitar caso en Barranquilla y el Atlantico.	0,10262
14/01/2020 14:44	Cartagena	@jorgecura1070 Atencion el dengue haciendo estragos. comentan una niña del municipio de Malambo se encuentra internada (UCI) en un centro asistencial. hay agilizar prevencion para evitar caso en Barranquilla y el Atlántico.	0,10262
27/02/2020 11:03	Bogotá	Por 100 persona 2 muerte causa del coronavirus la tasa letal significativa baja. lo sonar escandaloso. en Colombia 23 persona por100 muerte causa del dengue .do. ahora. en nuestro país y se ha algo?	0,09930
11/03/2020 11:39	Bogotá	la coyuntura. ha do con la investigación científica frente al dengue en Colombia y en país tropical ??tañen añoña de 12mil caso. @MincienciasCo @MinSaludCol @maxplanckpres	0,09909

Relación entre tópicos y su publicación en los meses estudiados

Adicionalmente se diseñó la figura 20 con el fin de identificar la cantidad de veces por mes en las que cada uno de los 12 tópicos es tratado. Las personas durante estos meses en particular, publican contenido en su mayoría relacionado con los tópicos 1,3,7 y 8; Los meses en donde más cantidad de publicaciones se efectúan es en enero, febrero y marzo del año 2020. Particularmente estas publicaciones están relacionadas con:

Tópico 1: Programas de prevención y vacunación contra el dengue en Colombia.

Tópico 3: Casos de dengue en Colombia con presencia particular de síntomas de dolor y fiebre.

Tópico 7: Jornadas de prevención/ control/ eliminación de criaderos del mosquito/zancudo transmisor del dengue.

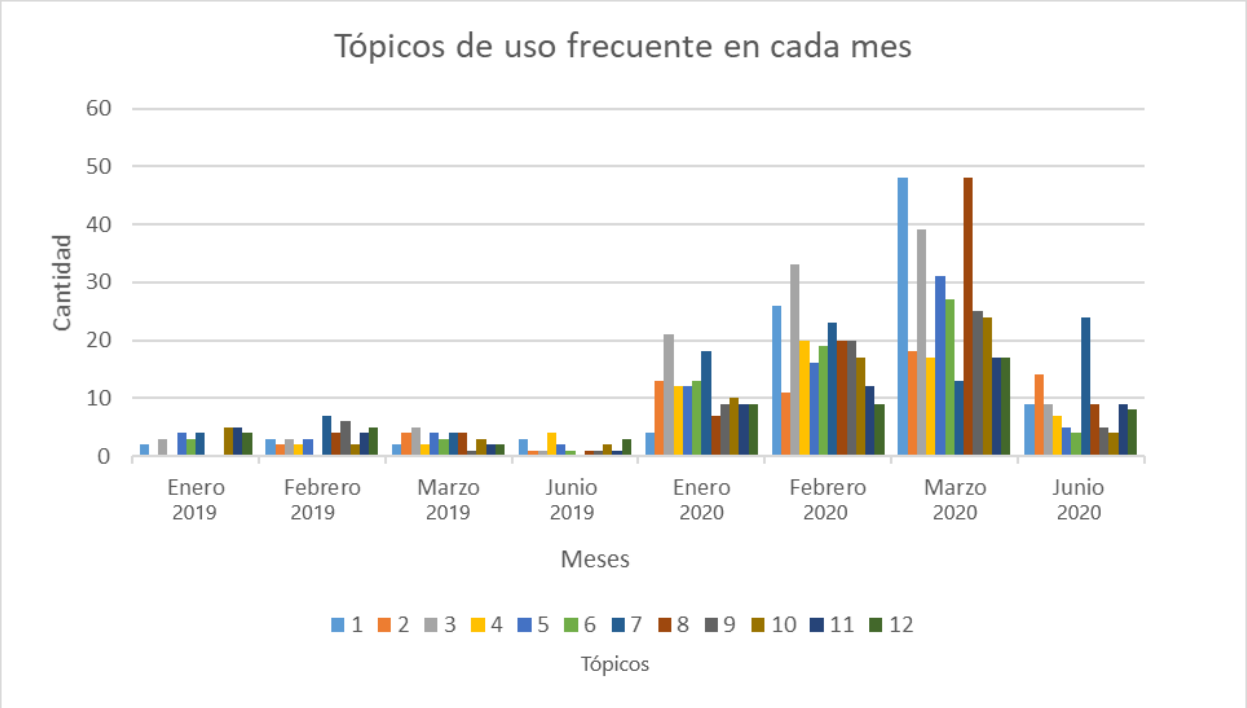
Tópico 8: Prevención en Colombia para los casos de muertes y/o enfermedades por picadura de mosquito del dengue.

Teniendo en cuenta los meses con mayor crecimiento de publicaciones y a su vez, atendiendo a lo consignado en dos informe presentado por el INS de Colombia sobre la principales epidemias y su comportamiento para los dos años estudiados, particularmente el dengue (Instituto Nacional de Salud, 2019)y (INS, 2020) en donde se señala que las semanas con mayor cantidad de casos son de la 1 a la 12, que integran los meses de enero, febrero y marzo tanto para un año como para el otro: se hará un análisis de tópicos más profundo únicamente para estos meses.

Si bien, marzo fue el mes en el que más se aumentó las veces en las que se abordaban los tópicos estudiados visualizado en la figura 20, del mismo modo que fue el mes en el que más cantidad de tweets se publicaron entorno a la temática estudiada.

Figura 20

Tópicos de uso frecuente por cada mes estudiado



Cantidad de tópicos por cada ciudad

Ahora bien, los temas entorno a cada uno de los tópicos en su mayoría fueron abordados en las ciudades de Bogotá, Barranquilla, Cartagena, Montería y Santa Marta; se encontró también que la cantidad de ciudades que hablaban de esto aumento más del 100 % del año 2019 al año 2020. En las figuras 21 y 22, se visualiza con mayor profundización, lo expuesto anteriormente.

Bogotá predomino tanto en los resultados de un año como en el otro, en 2020 Barranquilla, Medellín y Cali fueron las ciudades en donde se twitteo más, seguida, de esta; para el caso de 2019, fueron Santa Marta, Cartagena y Barranquilla.

Figura 21

Analisis de predominancia de tópicos por ciudades año 2019

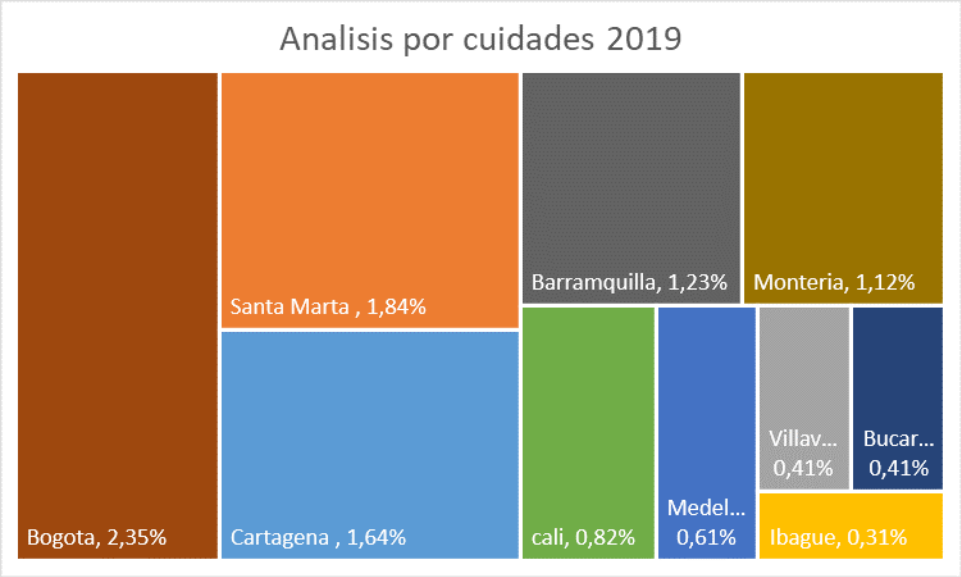
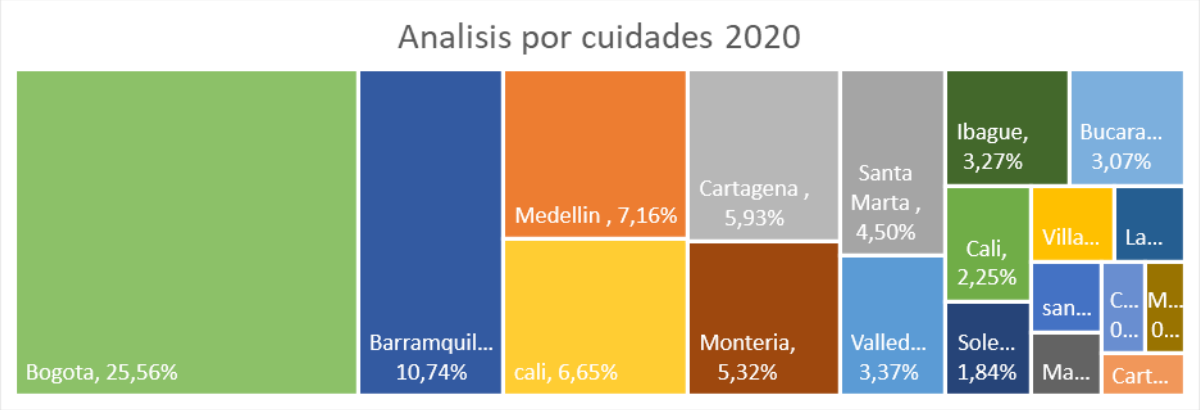


Figura 22

Análisis de predominancia de tópicos por ciudades año 2020



Según una publicación del Ministerio de Salud de Colombia con fines informativos sobre enfermedades transmisibles en Colombia, los mosquitos y zancudos transmisores del dengue

presentan un ciclo de vida amplio, en su mayoría, en altitudes que no superen los 2200 metros de altura sobre el nivel del mar (MinSalud, 2014).

De acuerdo con una investigación científica otro factor que, si bien puede ser una de las variables más significativas para que se transmita el dengue, relacionándolo a su vez con el ciclo de vida del transmisor, es la temperatura de las zonas donde habita. Los procesos biológicos de estos insectos poiquilotermos, se ven influenciados continuamente por la temperatura ambiente, las temperaturas entre los 78.8 °F y los 82.4 °F son las adecuadas para garantizar la supervivencia y el desarrollo de los mismos (Christofferson & Mores, 2016), situación que favorece la proliferación de los vectores; en temperaturas superiores o iguales a 50 °F se beneficia el proceso de ovoposición, encontrándose que en tiempos de lluvias esta actividad es mayor que en tiempos secos (Lega, Brown, & Barrera, 2017). De tal manera que para mayor profundización se recopilo información tanto de temperatura como de altitud para las ciudades mencionadas en la sección anterior.

Al analizar las altitudes y temperaturas de diez de las ciudades consignadas en las figuras 21, 22 y tabla 10 (información consignada allí fue extraída de Wather Spark) finalmente, se ratifica que en su mayoría se relaciona.

Tabla 10

Altitud y temperatura de las ciudades

Cuidad	Altitud (metros sobre el nivel del mar)	Oscilación de la temperatura (°F)	
Bogotá	2640	[44	66]
Barranquilla	18	[76	85]
Medellín	1495	[62	75]
Cali	1018	[66	85]
Cartagena	2	[76	88]
Montería	18	[74	96]
Santa Marta	6	[75	89]
Villavicencio	467	[67	89]
Bucaramanga	950	[68	81]
Ibague	1285	[63	83]

Nota: Adaptado de (Weather spark, 2008)

7. **Conclusiones**

A partir de los documentos científicos estudiados dentro de la revisión literaria se identificaron distintas herramientas para el análisis de tópicos dentro de colecciones de documentos particularmente de Twitter, siendo los más destacados el modelo LDA (Latent Dirichlet Allocation) y CTM (correlated Topic Models); para el caso del desarrollo de este proyecto de investigación se validaron ambos métodos mediante la medida de verosimilitud o también conocido como criterio de Loglikehood además de tener en cuenta el tiempo de procesamiento empleado por cada uno de los dos modelos, validándose que LDA tiene un mejor desempeño a partir de la medida de verosimilitud y desde la perspectiva del tiempo de procesamiento que se emplea para el hallazgo de la mejor cantidad de tópicos , este modelo requiere un tiempo significativamente menor(12 horas) comparado con el modelo CTM (96 horas). Resultando más pertinente para el estudio de esta temática el modelo LDA.

Por tanto, se considera método de modelamiento de tópicos una herramienta oportuna para ser adoptada por las entidades tomadoras de decisiones particularmente del área de la salud publica en Colombia independientemente de los eventos de interés en la vigilancia.

Respecto al Analisis del modelamiento de tópicos aplicado al contenido de tweets sobre el dengue en Colombia, se encuentra en gran parte que los temas publicados por los usuarios de esta red social están relacionados con las acciones de prevención, control y mitigación del contagio de dengue en Colombia. Si bien, es de gran importancia señalar que, aunque la campaña nacional “córtale las alas al dengue” fue lanzada en el año 2019 por el Ministerio de Salud, dentro de la información analizada se encontró que la actividad de vinculación a la misma por parte de la

comunidad aumento significativamente no durante este año sino hasta los inicios del 2020 en tras un proceso difusión y relacionamiento en mayor proporción.

Finalmente, se puede concluir que tanto el comportamiento de las publicaciones como el de los casos de dengue registrados difieren significativamente; mientras que para los casos de 2019 su tendencia todas las semanas es al crecimiento, para el 2020 sucede lo contrario; y si se analiza la cantidad de publicaciones de ambos años continuamente está en un comportamiento fluctuante, siendo hasta la semana 37 en donde se cumple un crecimiento superior de 2020 respecto a 2019, de ese tiempo hasta última semana en su mayoría es inferior. No obstante, a partir de un análisis de correlación efectuado se puede determinar que la relación entre el crecimiento de ambas variables es directa, es decir, a medida que la cantidad de casos crece, lo mismo sucede con las publicaciones, al mismo tiempo que la una decrece, la otra también lo hace.

Es relevante encontrar que con ayuda del modelamiento de tópicos se identificó que el mes de marzo es uno de los meses en donde más se publica tópicos relacionados con el dengue en Colombia, tanto en un año como en el otro, seguido de los meses de enero y Febrero; situación que concuerda con el análisis de los casos de dengue durante el mismo periodo de tiempo, de tal manera que se puede considerar viable y recomendable que entidades interesadas en salud publica puedan usar este tipo de analítica de datos para dar apoyo al desarrollo de sus proyectos.

8. Recomendaciones

Se recomienda en primera medida para profundizar en el análisis de modelamiento de tópicos realizar un estudio que considere la recolección de más cantidad de tweets, siendo esto posible a través la ampliación del periodo de búsqueda.

Al igual que se invita a contactar a la mayor cantidad de entidades de interés en salud pública con el fin de darles a conocer este mecanismo de vigilancia con el cual pueden apoyar sus procesos de toma de decisiones y/o generación de estrategias.

Referencias Bibliográficas

Aguirre, R. (1 de 02 de 2020). *El dengue en Colombia, con un aumento de casos del 93 % en enero*. Obtenido de El Colombiano: https://www.elcolombiano.com/colombia/el-dengue-al-alza-en-colombia-AI12381199

Al-garadi, M., Sadiq Khan, M., Varathan, K., Mujtaba, G., & Al-Kabsi, M. (08 de 2016). Uso de las redes sociales en línea para rastrear una pandemia: una revisión sistemática. *Revista de Informatica biomédica, 62,* 1-11. Obtenido de https://www.sciencedirect.com/science/article/pii/S1532046416300351?via%3Dihub

Atkins, B. (1992). *Tools computer aided corpus lexicography: The HEctor project* . Obtenido de https://doi.org/10.2307/44308286

Barrios, J. (15 de 06 de 2020). *Inteligencia Artificial y Machine Learning para todos*. Obtenido de https://www.juanbarrios.com/inteligencia-artificial-y-machine-learning-para-todos/

BBC News. (2020). *Por qué América Latina está registrando "la mayor epidemia de dengue de su historia"*. Obtenido de https://www.bbc.com/mundo/noticias-51496280

Beltrán, B. (2014). Mineria de datos. *Benemérita Universidad Autónoma de Puebla*, 67-70. 1. Obtenido de http://bbeltran.cs.buap.mx/NotasMD.pdf

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research 3,* 993-1022. Obtenido de https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf?fbclid=IwAR3RTl61Zlag4y4ubx7OCeunX0lCwbnT0O_hJbu5es9UqXw9K1vzmCxWNnI

Bouzillé, G., Poirier, C., Campillo-Giménez, B., Aubert, M.-L. ,., Chazard, E., Lavenu, A., & Cuggia, M. (02 de 2018). Aprovechar los macrodatos de los hospitales para monitorear las epidemias de gripe. *Biblioteca Nacional de Medicina* , 153.160. Obtenido de https://pubmed.ncbi.nlm.nih.gov/29249339/

Cairns, I., & Shetty, P. (16 de 07 de 2020). *Presentamos una nueva y mejorada API de Twitter.* Obtenido de Blog de Twitter: https://blog.twitter.com/es_la/topics/product/2020/Presentamos-una-nueva-y-mejorada-API-de-Twitter.html

Cenditel. (2016). *Guia teorica del proyecto modelado de topicos.* Obtenido de http://analisisdatos.cenditel.gob.ve/files/2018/09/LDA.pdf

Centro para el control y la prevención de la enfermedad. (10 de 12 de 2019). *CDC*. Obtenido de Dengue: https://www.cdc.gov/dengue/es/about/index.html

Chandia, B. (2016). *Aplicacion y Evaluacion LDA para asignacion de topicos en datos de twitter.*

Chandia, B. (2016). Aplicacion y Evaluacion para Asignacion de Topicos de datos de Twitter. *Pontificia Universidad Catolica de Valparaiso*, 55. Obtenido de http://opac.pucv.cl/pucv_txt/txt-5000/UCD5100_01.pdf

Choi, S., Lee, J., Kang, M., Min, H., Yoon, C., Yoon, & Sungroh. (01 de 10 de 2017). Large-scale machine learning of media outlets for understanding public reactions to nation-wide viral infection outbreaks. *Methods, 129*(1), 50-59. Obtenido de https://bibliotecavirtual.uis.edu.co:2191/science/article/pii/S1046202317300191#!

Christofferson, R., & Mores, C. (2016). Potencial for extrinsic incubation temperature to alter interplay betwwen transmission potential and mortality of dengue-infected Aedes aegyoti. *Environ Health Insights*, 10-119. Obtenido de https://pubmed.ncbi.nlm.nih.gov/27478382/

Collins, K. (2014). *Introduction to Supervised Machine Learning.* Obtenido de https://go.lnkam.com/link/r?u=https%3A%2F%2Fwww.coursera.org%2Flecture%2Fpython-machine-learning%2Fintroduction-to-supervised-machine-learning-EKQDv&campaign_id=b7YMMAqMdAL7wyzNe5m3wz&source=hl78ly46c2jdtn

Contreras, M. (2014). Minería de texto: una visión actual. *Biblioteca Universitaria, 17*(2), 129-138. Obtenido de https://brapci.inf.br/index.php/res/v/51778

Contreras, M. (29 de 03 de 2016). Minería de texto en la clasificación de material bibliográfico. *Biblios, 64*. Obtenido de https://www.redalyc.org/jatsRepo/161/16148511003/html/index.html

D, R., & single, J. (7 de 10 de 2018). *Text Mining with R*. Obtenido de https://www.tidytextmining.com/

Fidias G, A. (2012). *El proyecto de investigaciòn. Introducciòn a la metodologìa cientìfica* (6ta ed.). Obtenido de https://issuu.com/fidiasgerardoarias/docs/fidias_g._arias._el_proyecto_de_inv

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R.* (S. P. Ltd., Ed.) 1° edición. Obtenido de https://aedmoodle.ufpa.br/pluginfile.php/401852/mod_resource/content/5/Material_PDF/1.Discovering%20Statistics%20Using%20R.pdf

García, D. (2014). Tecnicas de clustering aplicadas al analisis de trending topics en conjunto de tweets. *Universidad Carlos III de Madrid*, 1-167. Obtenido de https://e-archivo.uc3m.es/bitstream/handle/10016/22233/PFC_adrian_garcia_dieguez_2014.pdf

Gomide, J., Veloso, A., Meira, W., Almeida, V., Benevenuto, F., Ferraz, F., & Teixeira, M. (2011). Vigilancia del dengue basada en un modelo computacional de localidad espacio-temporal de Twitter. *Proceedings of the ACM WebSci*, 14-17. Obtenido de https://dl.acm.org/doi/abs/10.1145/2527031.2527049

Griffiths, T., & Steyvers, M. (2004). *Finding scientific topics* (Vol. 101). Proceedings of the National academy of Sciences. Obtenido de https://www.pnas.org/content/101/suppl_1/5228.short

Guerra, C. (03 de 03 de 2015). *Hashtag: ¿qué es, para qué sirve y cómo usarlo?* Obtenido de https://carlosguerraterol.com/hashtag-que-es-para-que-sirve-como-usar/

Gupta, A., & Katarya, R. (10 de 2020). Social media based surveillance systems for healthcare using machine learning: A systematic review. *Journal of Medical Informatics, 18*. Obtenido de https://bibliotecavirtual.uis.edu.co:2191/science/article/pii/S1532046420301283#b0120

Hammoe, L. (2018). Deteccion de Topicos utilizando el modelo LDA. *Instituto Tecnologico de Buenos Aires*. Obtenido de

https://ri.itba.edu.ar/bitstream/handle/123456789/1250/TFI_Hammoe.pdf?sequence=1&isAllowed=y

Hammoe, L. (2018). *Detección de topicos Utilizando el modelo LDA.* Instituto Tecnologico de Buenos Aires. Obtenido de https://ri.itba.edu.ar/bitstream/handle/123456789/1250/TFI_Hammoe.pdf?sequence=1&isAllowed=y

Hartley, D., Nelson, N., Arthurd, R., Barboza, P., Collier, N., Lightfoot, N., . . . Brownstein, J. (11 de 2013). An overview of Internet biosurveillance. *Clinical Microbiology and Infection, 19*(11), 1006-1013. Obtenido de https://bibliotecavirtual.uis.edu.co:2191/science/article/pii/S1198743X14630020

Haupt, M., Diamant, A., Jiawei, L., Nali, M., & Mackey, T. (01 de 2021). Characterizing twitter user topics and communication network dynamics of the "Liberate" movement during COVID-19 using unsupervised machine learning and social network analysis. *Online Social Networks Media, 21*. Obtenido de https://bibliotecavirtual.uis.edu.co:2191/science/article/pii/S2468696420300550

Hirose, H., & Wang, L. (2012). Prediction of Infectious Disease Spread Using Twitter: A Case of Influenza. *IEEE*. Obtenido de https://ieeexplore.ieee.org/abstract/document/6424743/references#references

Huan, L., & Xia, H. (2013). Text Analytics in Social Media. *Springer Science,* 1-522. Obtenido de https://link.springer.com/chapter/10.1007/978-1-4614-3223-4_12

Ignacio Santiago. (26 de 12 de 2019). *Qué Es TWITTER, Para Qué Sirve y Cómo Funciona*. Obtenido de Santiago Ignacio: https://ignaciosantiago.com/twitter-que-es-como-funciona/

INS. (2020). *Boletin epidemiologico semana 53*. Obtenido de https://www.ins.gov.co/buscador-eventos/BoletinEpidemiologico/2020_Boletin_epidemiologico_semana_53.pdf

INS. (2020). Boletin Epidemiologico semana 53. Obtenido de https://www.ins.gov.co/buscador-eventos/BoletinEpidemiologico/2020_Boletin_epidemiologico_semana_53.pdf

Instituto Nacional de Salud. (2019). Boletin Epidemiologico de Salud semana 52. Obtenido de https://www.ins.gov.co/buscador-eventos/BoletinEpidemiologico/2019_Boletin_epidemiologico_semana_52.pdf

Instituto Nacional de Salud. (2019). *Boletin epidemiologico semana 52*. Obtenido de Colombia, Semanas Epidemiologicas: https://www.ins.gov.co/buscador-eventos/BoletinEpidemiologico/2019_Boletin_epidemiologico_semana_52.pdf

Kagashe, I., Zhijun, Y., & Mphil, I. (12 de 09 de 2017). Enhancing Seasonal Influenza Surveillance: Topic Analysis of Widely Used Medicinal Drugs Using Twitter Data. *Journal Of Medical Internet Research, 19*(9). Obtenido de https://www.jmir.org/2017/9/e315/

Kalyanam, J., Katsuki, T., Lanckriet, G., & Mackey, T. (02 de 2017). Exploring trends of nonmedical use of prescription drugs and polydrug abuse in the Twittersphere using unsupervised machine learning. *Addictive Behaviors*, 289-295. Obtenido de https://bibliotecavirtual.uis.edu.co:2191/science/article/pii/S0306460316302994#bb0150

Kraemer, M., Bisanzio, D., Reiner, R., Zakar, R., Hawkins, B. F., Smith, L., . . . T, P. (2018). Las inferencias sobre la variación espacio-temporal en la transmisión del virus del dengue son sensibles a las suposiciones sobre la movilidad humana: un estudio de caso que utiliza tweets geolocalizados de Lahore, Pakistán. *Biblioteca Nacional de Medicina de EE. UU, 1*(16), 18-114. Obtenido de https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6404370/

Kuz, A., Falco, M., & Giandini, R. (2016). Analisis de redes sociales: Un Caso practico. *Computacion y sistemas*. Obtenido de http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1405-55462016000100089

Lafferti, J., & Blei, D. (2007). Correlated Topic Model . *Euclid*. Obtenido de https://projecteuclid.org/journals/annals-of-applied-statistics/volume-1/issue-1/A-correlated-topic-model-of-Science/10.1214/07-AOAS114.full?tab=ArticleLink

Lampos, V., & Cristianini, N. (2010). Seguimiento de la pandemia de gripe mediante el seguimiento de la red social. *2do taller internacional sobre procesamiento de información cognitiva, CIP*, 411–416. Obtenido de https://ieeexplore.ieee.org/abstract/document/5604088/

Lauren, E., Kamil, T., & Gurkan, B. (2017). El modelo basado en redes de macrodatos de redes sociales predice la difusión de enfermedades contagiosas. *Biblioteca Nacional de Medicina*

de EE. UU, *45*(3), 110-120. Obtenido de https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6554721/

Lega, J., Brown, H., & Barrera, R. (2017). Aedes aegypti (Diptera:Culicidae) Abundance model improved with relative humidity and precipitation-driven edd hatching. *J. M Entomol*, 84-1375. Obtenido de https://pubmed.ncbi.nlm.nih.gov/28402546/

Mackey, T., Purushothaman, V., Li, J., Shan, N., Nali, M., Bardier, C., . . . Cuomo, R. (2020). Aprendizaje automático para detectar el autoinforme de síntomas, probar el acceso y la recuperación asociados con COVID-19 en Twitter: estudio retrospectivo de Big Data Infoveillance. *Biblioteca Nacional de Medicina*. Obtenido de https://pubmed.ncbi.nlm.nih.gov/32490846/

Manual MSD. (2020). *Generalidades sobre las infecciones por arbovirus, arenavirus y filovirus*. Obtenido de version para profesionales: https://www.msdmanuals.com/es/professional/enfermedades-infecciosas/arbovirus-arenavirus-y-filovirus/generalidades-sobre-las-infecciones-por-arbovirus-arenavirus-y-filovirus

Marqués, P., Gonzales, M., Vega , J., Pinto, A., & Quiroga, E. (2013). El analisis de las redes sociales. Un metodo para la mejora de la seguiridad en las organizaciones sanitarias. *Rev Esp Salud Pública*, 209-219. Obtenido de http://scielo.isciii.es/pdf/resp/v87n3/01_colaboracion_especial.pdf

Martinez, P., Martinez, J., Isabel, S., Julián, M., Luna, A., & Ricardo, R. (2016). Turning user generated health-related content into actionable knowledge through text analytics services. *Computers and Industry*, 43-56. Obtenido de https://bibliotecavirtual.uis.edu.co:2191/science/article/pii/S0166361515300518

McCreadie , R., Macdonald , C., Ounis , I., Osborne, M., & Petrovic, S. (23 de 12 de 2013). *IEEE Explore*. Obtenido de https://ieeexplore.ieee.org/document/6691620

Mesas, R. M. (2015). Analisis de Tendencias y Marcas Deportivas a traves de Twitter. *Tesis posdoctoral*, 68. Obtenido de https://repositorio.uam.es/bitstream/handle/10486/669057/Mesas_Javega_RusMaria_tfg.pdf

Minerva. (29 de 06 de 2017). *KDD: ¿Qué es el Knowledge Discovery in Databases o KDD?* Obtenido de Minerva: https://mnrva.io/kdd-platform.html

MinSalud . (11 de 01 de 2021). Guia para la implementacion del modelo de articulacion de redes "TO2ES: Por una colombia saludable". Bogota, Cundinamarca, Colombia. Obtenido de https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/VS/PP/ENT/implementacion-modelo-redes.pdf

MinSalud. (14 de 09 de 2014). *Dengue*. Obtenido de Prevención enfermedades transmisibles: https://www.minsalud.gov.co/salud/publica/PET/Paginas/dengue.aspx

MinSalud. (04 de 06 de 2014). *Ministerio de Salud y Proteccion Social* . Obtenido de Institucional: https://www.minsalud.gov.co/Ministerio/Institucional/Paginas/mision-vision-principios.aspx

MinSalud. (08 de 02 de 2019). *"Hay que cortarle las alas al dengue": ministro Juan Pablo Uribe*. Obtenido de Boletin de Prensa No 019 de 2019: https://www.minsalud.gov.co/Paginas/Hay-que-cortarle-las-alas-al-dengue-ministro-Juan-Pablo-Uribe.aspx

MinSalud. (2020). *Boletin Epidemiologico Semanal* . Obtenido de https://www.ins.gov.co/buscador-eventos/BoletinEpidemiologico/2020_Boletin_epidemiologico_semana_7.pdf

MinSalud. (2021). *Portal Web Sivigila*. Obtenido de Estado Epidemiologico de Colombia: http://portalsivigila.ins.gov.co/

Missier, P., Romanovsky, A., Miu, T., Pal, A., Da, M., & da Silva Sousa, L. (05 de 10 de 2016). Tracking Dengue Epidemics Using Twitter Content Classification and Topic Modelling. *International Conference on Web Engineering*, 80-92. Obtenido de https://bibliotecavirtual.uis.edu.co:2142/chapter/10.1007/978-3-319-46963-8_7

Missier, P., Sousa, L., Rafael, d. D., Cedrim, D., Garcia, A., Uchoa, A., . . . Romanovsky, A. (06 de 2018). VazaDengue: un sistema de información para prevenir y combatir las enfermedades transmitidas por mosquitos con las redes sociales. *Information Systems , 75*, 26-42. Obtenido de https://bibliotecavirtual.uis.edu.co:2191/science/article/pii/S030643791730618X

Odlum, M., & Summoo, Y. (1 de 6 de 2015). What can we learn about the Ebola outbreak from tweets? *American Journal of nfection Control* , 563-571. Obtenido de https://bibliotecavirtual.uis.edu.co:2191/science/article/pii/S0196655315001376

Olarte, D., & Ariza, N. (2020). Evaluacion de metodos de agrupamiento DBSCAN y LDA para el analisis de contenido de la red social Twitter. *Tesis pregrado*, 113. Bucaramanga. Obtenido de http://tangara.uis.edu.co/biblioweb/

OMS. (2005). 58° Asamblea Mundial de la salud. Ginebra. Obtenido de https://apps.who.int/gb/ebwha/pdf_files/WHA58-REC1/A58_2005_REC1-sp.pdf

OMS. (15 de 05 de 2013). *Zika*. Obtenido de Organizacion Mundial de la Salud: https://www.who.int/es/news-room/fact-sheets/detail/zika-virus

OMS. (29 de 04 de 2018). *Chagas*. Obtenido de Organizacion Mundial de la Salud: https://www.who.int/es/news-room/fact-sheets/detail/chagas-disease-(american-trypanosomiasis)

OMS. (28 de 04 de 2018). *Chikungunya*. Obtenido de Organizacion Mundial de la Salud: https://www.who.int/es/news-room/fact-sheets/detail/chikungunya

OMS. (30 de 04 de 2018). *Leishmaniasis* . Obtenido de Organizacion Mundial de la Salud: https://www.who.int/es/news-room/fact-sheets/detail/leishmaniasis

OMS. (2020). *Dengue y dengue grave.* Organizacion Mundial de la Salud . Obtenido de https://www.who.int/es/news-room/fact-sheets/detail/dengue-and-severe-dengue

OMS. (24 de 06 de 2020). *Dengue y dengue grave*. Obtenido de https://www.who.int/es/news-room/fact-sheets/detail/dengue-and-severe-dengue

OMS. (2 de 03 de 2020). Enfermedades transmitidas por vectores. Obtenido de https://www.who.int/es/news-room/fact-sheets/detail/vector-borne-diseases

Organizacion Mundial de la Salud . (2019). *Marco para el intercambio de virus gripales y el acceso a las vacunas y otros beneficios en el contexto de la preparación para una gripe pandémica.* Obtenido de https://apps.who.int/gb/ebwha/pdf_files/WHA72/A72(12)-sp.pdf

Ortiz, A. (11 de 01 de 2019). *Por qué es tan importante el análisis Big Data*. Obtenido de Hostdimeblog: https://blog.hostdime.com.co/por-que-es-tan-importante-el-analisis-big-data/

OSPS. (2017). *Analisis de situacion de salud de Santander* . Bucaramanga . Obtenido de http://web.observatorio.co/publicaciones/enos_indicadores_basicos_de_salud_2017.pdf

Padilla, J., Lizarazo, F., Murillo, O., Mendigaña, A., Pachón, E., & Vera, M. (2017). Epidemiología de las principales enfermedades trasmitidas por vectores en Colombia, 1990-2016. *Biomedica*, 27-40. Obtenido de http://www.scielo.org.co/pdf/bio/v37s2/0120-4157-bio-37-s2-00027.pdf

Pérez, M. L., & Calderòn, Z. (2011). *orientaciones prácticas para la elaboración exitosa de trabajos de grado en ingeniería* (primera ed.). Bucaramanga: Universidad Industrial de Santander.

PLoS Medicine. (2010). It's the network, stupid: why everything in me-dicine is connected. *PLoS Medicine Editors*. Obtenido de https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0050071

Ribas, E. (08 de 01 de 2018). *¿Qué es el Data Mining o minería de datos?* Obtenido de IEBS: https://www.iebschool.com/blog/data-mining-mineria-datos-big-data/

Rodriguez, C., & Rojas, J. (2019). Reglas de asociacion aplicadas al analisis de contenido de los tweets sobre enfermedades trnasmitidas por vectores en Santander, Colombia. *Universidad Industrial de Santander*, 1-115. Obtenido de http://tangara.uis.edu.co/biblioweb/

Rodríguez, J. (06 de 2013). Cómo utilizar el Análisis de Redes Sociales para temas de historia. *Tesis posdoctoral*. Obtenido de http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1665-44202013000100004#:~:text=El%20An%C3%A1lisis%20de%20Redes%20Sociales%20o%20an%C3%A1lisis%20estructural%20(ARS)%20ha,organizaci%C3%B3n%20empresarial%20y%20comunicaci%C3%B3n%20electr%C3%B3nica

Sánchez, T. (2014). Las redes sociales en Internet y su impacto en la Salud Pública. *Informacion en Ciencias de la salud , 25*(2). Obtenido de http://acimed.sld.cu/index.php/acimed/article/view/615/387#:~:text=Sin%20embargo%2C%20la%20utilidad%20de,comportamientos%20asociados%20a%20la%20b%C3%BAsqueda

Sancho, F. (14 de 12 de 2020). *Aprendizaje Supervisado y No Supervisado*. Obtenido de http://www.cs.us.es/~fsancho/?e=77

Santos, J. (13 de 06 de 2017). *Qués es el Análisis de Redes Sociales (#QuoMove)*. Obtenido de javiersantosbueno.com: https://javisantosbueno.com/2017/06/13/ques-es-el-analisis-de-redes-sociales-quomove/#:~:text=El%20an%C3%A1lisis%20de%20redes%20sociales%20es%20una%20aproximaci%C3%B3n%20metodol%C3%B3gica%20y,%2C%20organizaciones%2C%20pa%C3%ADses%20o%20cosas.

Schneider, J., McFadden, R., LAumann, E., Prem, S., Gandham, S., & Oruganti, G. (10 de 2012). Identificación del agente de cambio candidato entre hombres en riesgo de infección por VIH. *Biblioteca Nacional de Mediciona*, 1192-201. Obtenido de https://pubmed.ncbi.nlm.nih.gov/22762951/

Serri, M. (2018). Redes sociales y salud. *Revista chilena de infectologia*. Obtenido de https://scielo.conicyt.cl/scielo.php?script=sci_arttext&pid=S0716-10182018000600629

Sorin, M. (2011). Analyzing Social Media Networks with NodeXL: Insights from a Connected World by Derek Hansen, Ben Shneiderman, and Marc A. Smith. *International Journal of Human–Computer Interaction*, 405-408. Obtenido de https://www.tandfonline.com/doi/abs/10.1080/10447318.2011.544971?journalCode=hihc20

Spinak, E. (1996). Diccionario Enciclopédico de Bibliometría, Cienciometría e informática. Obtenido de UNESCO: https://unesdoc.unesco.org/ark:/48223/pf0000243329

Spreco, A., Eriksson, O., Dahlstrom, O., Cowling, B., & Timpka, T. (15 de 06 de 2017). Detección y predicción integradas de la actividad de la influenza para la vigilancia en tiempo real: diseño de algoritmos. *Biblioteca Nacional de Medicina*, 221. Obtenido de https://pubmed.ncbi.nlm.nih.gov/28619700/

Stieglitz, S., & Bruns, A. (03 de 2013). Insights from hashtag #supplychain and Twitter Analytics: Considering Twitter and Twitter data for supply chain practice and research. *International Journal of Social Research Methodology, 16*(2), 91-108. Obtenido de https://www.scopus.com/record/display.uri?eid=2-s2.0-84874484025&origin=inward&txGid=b960067d0b9554be42571fbf32a0431f

Timarán, S., Hernández, ,. I., Caicedo, S. J., Hidalgo, A., & Alvarado, J. (2016). El proceso de descubrimiento de conocimiento en bases de datos. En *En Descubrimiento de patrones de*

desempeño académico con árboles de decisión en las competencias genéricas (págs. 63-86). Bogota: Ediciones Universidad Cooperativa de Colombia. Obtenido de https://ediciones.ucc.edu.co/index.php/ucc/catalog/download/36/40/230-1?inline=1

Twitter. (15 de 05 de 2018). *Centro de ayuda* . Obtenido de Información sobre las API de Twitter: https://help.twitter.com/es/rules-and-policies/twitter-api#:~:text=Twitter%20permite%20acceder%20a%20partes,opini%C3%B3n%20del%20cliente%20en%20Twitter.

Twitter. (03 de 01 de 2018). *Glosario*. Obtenido de Centro de ayuda: https://help.twitter.com/es/glossary

Universidad Internacional de Valencia. (18 de 09 de 2018). *Promoción de la salud: definición, objetivos y ejemplos,* https://www.universidadviu.com/co/actualidad/nuestros-expertos/promocion-de-la-salud-definicion-objetivos-y-ejemplos. Obtenido de Ciencias de la salud: https://www.universidadviu.com/co/actualidad/nuestros-expertos/promocion-de-la-salud-definicion-objetivos-y-ejemplos#:~:text=La%20promoci%C3%B3n%20de%20la%20salud%20es%20un%20concepto%20mucho%20m%C3%A1s,de%20educaci%C3%B3n%20para%20la%20salud.

Valvuena, J., & Benitez, J. (2011). *Motores de Busqueda Web: Estado del arte en Probabilistic Latent Semantic Analysis y en Latent Dirichlet Allocation aplicado a problemas de acceso a la informacion en la Wem.* Obtenido de https://www.jabenitezz.com/personal/MASTER/MOTORES_DE_BUSQUEDA_WEB/TAREAS/MBW-TareaExtraFinal-JoseAlbertoBenitezAndrades-y-JuanAntonioValbuenaLopez.pdf

Van Manen, M. (1990). *nvestigar experiencia presencial: La ciencia humana para una pedagogía sensible a la acción de Londres.* . Ontario: Althouse.

Vasileios, L., & Nello, C. (14 de 10 de 2010). Tracking the flu pandemic by monitoring the social web. *IEEE Explore* . Obtenido de https://ieeexplore.ieee.org/abstract/document/5604088

Weather spark. (2008). *The Weather Year Round Anywhere on Earth* . Obtenido de https://weatherspark.com/

Zhai, C. x. (2013). *Mining Text data* . Springer. Obtenido de https://link.springer.com/book/10.1007%2F978-1-4614-3223-4#about

Zhou, L., Zhang, D., Christopher, Y., & Yu, W. (02 de 2018). Harnessing social media for health information management. *Electronic Commerce Research and Applications*, 139-151. Obtenido de https://bibliotecavirtual.uis.edu.co:2191/science/article/pii/S1567422317300960

Zhu, B., Zheng, X., Lui, H., Li, J., & Wang, W. (11 de 2020). Analysis of spatiotemporal characteristics of big data on social media sentiment with COVID-19 epidemic topics. *Chaos,Solitons & Fractals , 140*. Obtenido de https://bibliotecavirtual.uis.edu.co:2191/science/article/pii/S0960077920305208

Zikmund, W. G. (2003). *Business research methods* (septima ed.). Thomson South-Western: Ohio.