Análisis del impacto de las estrategias desarrolladas en redes sociales para la promoción de salud y prevención de enfermedades en Colombia.

Yanibe Franco Díaz
Martha Salcedo Parada

Director: Henry Lamos Díaz
Ph.D. Física-Matemática

Codirector:
Yuly Andrea Ramírez Sierra
Ingeniera industrial

Universidad Industrial de Santander
Facultad de Ingeniería Fisicomecánicas
Escuela de Estudios Industriales y Empresariales
Bucaramanga
2019

Agradecimientos

A Dios, por permitirnos llegar hasta aquí

A nuestras familias

Al profe Henry Lamos por su apoyo durante el proceso

A Yuly Ramírez; por su apoyo, dedicación y paciencia durante el desarrollo del proyecto

Al grupo de Investigación OPALO, por darnos la oportunidad de trabajar con ellos.

Yanibe Franco Diaz y Martha Salcedo Parada

Tabla de contenido

Lista de figuras

Lista de tablas

Listas de apéndices

Resumen

Título: Análisis del impacto de las estrategias desarrolladas en redes sociales para la promoción de salud y prevención de enfermedades en Colombia. *

Autores: Yanibe Franco Díaz, Martha Salcedo Parada **

Palabras Clave: Clasificación no supervisada, CTM, LDA, Rstudio, Tweet.

Descripción:
En este proyecto se evalúa el impacto de las estrategias difundidas a través de Twitter; referentes a la promoción de la salud y prevención de enfermedades en Colombia, utilizando minería de datos para caracterizar la información obtenida; Inicialmente se seleccionan las autoridades de salud en Colombia y sus campañas en Twitter luego se extraen los datos a través del API de Twitter con ayuda del software de programación Rstudio, el cual es utilizado en todo el proceso, estos datos son sometidos a un preprocesamiento con el fin de ajustar dos clasificadores no supervisados, *Latent Dirichlet Allocation* y *Correlated Topic Models*. En cada modelo se estima el parámetro k o número de tópicos el cual es fundamental para aplicar cada método, para ello se emplean dos métodos de muestreo Gibbs y VEM para LDA, y solo VEM para CTM, ya que en este caso no es posible utilizar Gibbs. Posteriormente para seleccionar el algoritmo que mejor se ajusta a los datos se utiliza la distancia de Hellinger, la cual arroja que LDA con VEM es el mejor modelo para este caso, por lo cual se determinan y etiquetan los tópicos predominantes en LDA VEM, de acuerdo a las matrices de probabilidad proporcionadas por el modelo LDA VEM y las estrategias de promoción de la salud y prevención de enfermedades que las entidades publican en sus páginas web institucionales. Finalmente se evalúa el impacto de acuerdo a la intensidad de los mensajes respecto a número de retweet y favoritos, encontrando un 71.3% de coincidencia con las estrategias publicadas en página web y que la cantidad de retweets y favoritos no es proporcional a la cantidad de tweet por cada estrategia.

---

* Proyecto de grado

** Facultad de Ingenieria Fisico-mecanicas. Escuela de estudiaos industriales y empresariales. Programa de Ingenieria Industrial. Director: Ph.D Henry Lamos Diaz. Codirector: Ing. Yuly Ramírez

Abstract

Project Title: Analysis of the impact of the strategies developed in social networks for health promotion and disease prevention in Colombia

Authores: Yanibe Franco Díaz, Martha Salcedo Parada

Keywords: CTM, LDA, Rstudio, Tweet, Unsupervised classification.

Description:

In this project the impact of the strategies spread through Twitter is evaluated; referring to health promotion and disease prevention in Colombia, using data mining to characterize the information obtained; Initially, the health authorities in Colombia and their campaigns on Twitter are selected, then the data is extracted through the Twitter API with the help of the Rstudio programming software, which is used throughout the process, this data is subjected to a preprocessing with In order to adjust two unsupervised classifiers, Latent Dirichlet Allocation and Correlated Topic Models. In each model is estimated the parameter k number of topics which is essential to apply each method, for these two sampling methods are used Gibbs and VEM for LDA, and only VEM for CTM, since in this case it is not possible to use Gibbs. Subsequently, to select the algorithm that best fits the data, the Hellinger distance is used, which shows that LDA with VEM is the best model for this case, Therefore, the predominant topics in LDA VEM are determined and labeled, according to the probability matrices provided by the LDA VEM models and the health promotion and disease prevention strategies that the entities publish in their institutional web pages. Finally, the impact is evaluated according to the intensity of the messages regarding retweet number and favorites, finding a 71.3% coincidence with the strategies published on the website and that the number of retweets and favorites is not proportional to the amount of tweet for each strategy.

---

* Graduation Project

** Facultad de Ingenieria Fisico-mecanicas. Escuela de estudiaos industriales y empresariales. Programa de Ingenieria Industrial. Director: Ph.D Henry Lamos Diaz. Codirector: Ing. Yuly Ramírez

Introducción

Actualmente se genera una enorme cantidad de datos diariamente (Aggarwal, Cheng Xiang, Crain, Zhou, Zha & Yang, 2012), sea en forma consciente o inconsciente por medio de la Internet, una de las principales herramientas de las que disponen los gobiernos para facilitar el acceso a la información, específicamente en sitios web institucionales y redes sociales. Estos medios están proporcionando datos oportunos, disponibles y rentables, como fotografías digitales, vídeos, blogs, entre otro tipo de datos, que permiten identificar enfermedades emergentes y así realizar vigilancia epidemiológica, con el fin de examinar tendencias, monitorear las prácticas de autocuidado de las personas con condiciones críticas e identificar patrones útiles, novedosos y comprensibles que apoyen la toma de decisiones e incentiven el cambio o la creación de políticas estratégicas.

En este sentido, la información se ha transformado en pieza clave para indagar en las aplicaciones médicas, como las estrategias de promoción de salud y prevención en las diferentes enfermedades que hoy en día están afectando a la sociedad, de manera que se han adoptado técnicas de análisis que han permitido explorar, codificar, identificar y caracterizar adecuadamente grandes volúmenes de datos recopilados de sitios de redes sociales en el contexto de problemáticas de salud pública (Kalyanam, Katsuki, Lanckriet, & Mackey, 2017). descubriendo patrones y tendencias que otorgan conocimiento sobre la situación analizada, según el objetivo y tipos de datos; este conjunto de técnicas está relacionadas con la minería de datos y el aprendizaje automático que se han convertido en una estrategia fundamental en la vigilancia digital teniendo un impacto positivo para disminuir la brecha entre la generación y el análisis de los datos.

La información generada en redes sociales se obtiene mediante la interfaz de programación de aplicaciones (por sus siglas en inglés API), la cual permite recopilar Tweets, *Retweets* o *Hashtag*

en Twitter; con esta información es posible realizar ciertos análisis tanto descriptivo, como análisis de contenido o análisis de redes, dado que se puede consultar un usuario o tema en específico (Finfgeld-Connett, 2015). Twitter se puede considerar una fuente de información abierta. Por lo tanto, esta fuente es una oportunidad para descubrir conocimiento de valor aplicando técnicas de minería de datos y de aprendizaje automático.

El presente trabajo pretende evaluar el impacto de las estrategias desarrolladas en redes sociales de promoción de la salud y prevención de enfermedades, mediante el análisis de la intensidad de los mensajes (número de veces que los mensajes han sido reenviados *Retweet* y número de *favoritos*) publicados por instituciones de salud en redes sociales como Twitter, y que hagan referencia específica a las temáticas de promoción en salud y prevención de enfermedades en Colombia. Esto permite caracterizar la información y descubrir patrones de comunicación y difusión de la misma, Así como analizar tópicos predominantes y comparar esta información descubierta con las campañas de promoción y prevención de salud de las respectivas autoridades de salud.

Este documento sigue una estructura que comprende el desarrollo de una revisión de literatura que soporta el tema de investigación, así como su correspondiente planteamiento del problema, objetivos, revisión de literatura, marco teórico, desarrollo de la metodología, resultados y conclusiones.

Tabla de cumplimiento de objetivos

Tabla 1
Cumplimiento de los Objetivos

Objetivos específicos	Ubicación
1. Realizar una revisión literaria para identificar el uso de clasificadores no supervisados en el sector salud utilizando datos de redes sociales.	Cap. 3
2. Construir clasificadores no supervisados aplicados al repositorio objeto de estudio y comparar la exactitud y precisión de los resultados obtenidos con cada modelo desarrollado, para identificar el modelo más representativo.	Cap. 6
3. Descubrir patrones o tendencias en las estrategias de promoción de salud y prevención de enfermedades en Colombia a partir de los datos obtenidos del API de Twitter, para caracterizar la información obtenida.	Cap. 7
4. Elaborar un artículo de carácter publicable, para dar a conocer los resultados de la investigación	Apéndice B

1. Planteamiento del problema

A nivel mundial, la promoción de la salud permite que la personas tengan control sobre su salud mediante la prevención sin enfocarse únicamente en el tratamiento y curación; abarcado las intervenciones sociales, ambientales, políticas y económicas, para proteger la salud y calidad de vida (Organización mundial de la salud [OMS], 2016).

En Colombia los programas de promoción y prevención son procesos que permiten a los colombianos mediante determinantes de la salud, los medios necesarios para mejorar y controlar su salud. Estos procesos se desarrollan principalmente en la formulación de políticas, creación de ambientes favorables, participación ciudadana; suponiendo una acción intersectorial que hace la movilización social para la transformación de las condiciones de salud (Ministerio de salud y protección solcial [MINSALUD], s.f.).

La OMS (2005) en la 58ª Asamblea Mundial de la Salud invita a que se considere "la posibilidad de establecer y aplicar sistemas electrónicos nacionales de información en materia de salud pública, y de mejorar, mediante la información, la capacidad de vigilancia y de respuesta rápida a las enfermedades y las emergencias de salud pública" (p.116). En este sentido, en Colombia, Torres Montero & Cabrera Chaves (2008) refieren que las campañas de promoción de la salud y prevención de enfermedades en los últimos años en su gran mayoría se realizan a partir de pautas publicitarias, cuñas, vallas unitarias, entre otros.

Esto deduce que hasta el 2008, en Colombia los medios más utilizados para la promoción de la salud y prevención de enfermedades no especificaban muy claramente el uso de redes sociales. Cabe resaltar que "la comunicación e interacción en las redes sociales a menudo reflejan los acontecimientos y la dinámica de la vida real, por lo que el continuo aumento de usuarios provoca

que las redes sociales se conviertan en precisos sensores de los acontecimientos del mundo real" (Garcia Diéguez, 2014, p5).

En la actualidad las páginas web o las redes sociales de las instituciones de salud son una herramienta de gran utilidad; por ejemplo: en los países de la región de américa, los ministerios de salud sin excepción cuentan con un espacio web institucional en Internet. Asimismo, muchos ministerios de salud también tienen presencia en al menos una red social, generalmente Twitter, sobre el uso que de las redes sociales hacen, principalmente pareciera que la prioridad es compartir mensajes de prevención y promoción de la salud, además de información clave sobre actividades relacionadas con las autoridades que dirigen esas instituciones (Novillo Ortiz y Hernández Pérez, 2015, p 63).

Para abordar esta realidad se plantea un estudio de investigación que analice el impacto de las estrategias desarrolladas en promoción de salud y prevención de enfermedades en Colombia, a partir de información obtenida de redes sociales, abordándolo desde la perspectiva de si realmente lo que reflejan las redes sociales de dichas instituciones en Colombia hace hincapié en las campañas promovidas en su página web, esto con el fin de apoyar a las instituciones en cuanto a la evaluación del impacto de este tipo de información.

2. Objetivos

2.1 Objetivo General

Analizar el impacto de las estrategias desarrolladas en redes sociales para la promoción de salud y prevención de enfermedades en Colombia, aplicando clasificadores no supervisados.

2.2 Objetivos específicos

- Realizar una revisión literaria para identificar el uso de clasificadores no supervisados en el sector salud utilizando datos de redes sociales.
- Construir clasificadores no supervisados aplicados al repositorio objeto de estudio y comparar la exactitud y precisión de los resultados obtenidos con cada modelo desarrollado, para identificar el modelo más representativo.
- Descubrir patrones o tendencias en las estrategias de promoción de salud y prevención de enfermedades en Colombia a partir de los datos obtenidos del API de Twitter, para caracterizar la información obtenida.
- Elaborar un artículo de carácter publicable, para dar a conocer los resultados de la investigación

3. Revisión de la literatura

El estudio de los programas de promoción de la salud y prevención de enfermedades se orienta según las diferentes etapas de la vida del ser humano y los diferentes niveles de prevención. Las etapas de la vida del ser humano se distinguen en: "in útero y nacimiento, primera infancia (0-5 años), infancia (6 - 11 años), adolescencia (12-18 años), juventud (14 - 26 años), adultez (27 - 59 años) y vejez (60 años y más) aunque no deben tomarse en forma absoluta y recordar que existe diversidad individual y cultural" (MINSALUD, s.f.); los niveles de prevención de la enfermedad se definen en prevención primaria, secundaria y terciaria.

> Según la OMS "La prevención primaria está dirigida a evitar la aparición inicial de una enfermedad o dolencia. La prevención secundaria y terciaria tiene por objeto detener o retardar la enfermedad ya presente y sus efectos mediante la detección precoz y el tratamiento adecuado o reducir los casos de recidivas y el establecimiento de la cronicidad" (OMS, 1998, p 13).

Adicionalmente las diferentes temáticas tratadas alrededor de la promoción de la salud y prevención de enfermedades proponen un enfoque basado en la vigilancia apoyado en la evidencia, utilizando fuentes convencionales o de redes, para su posterior análisis, que contribuya a los diferentes objetivos de salud pública; la primera fuente es la recolección de información por medio de encuestas y la segunda es aquella en la que se utilizan las redes sociales.

Dentro de algunos ejemplos que utilizan diferentes fuentes para el análisis del contenido de la información difundida, se destacan los siguientes propósitos: evaluar las estrategias de diseminación para aumentar la aceptación del conocimiento de salud pública y los programas y políticas basados en la evidencia (EBPP, por sus siglas en inglés) para el control de la diabetes entre los departamentos de salud locales (LHD, por sus siglas en inglés) (Parks et al., 2017),

cuantificar la comunicación de la vacuna contra el virus del papiloma humano (VPH) en Twitter y el desarrollo una metodología novedosa para mejorar la recopilación y el análisis de los datos de Twitter (Massey , Leader, Yom-Tov, Fisher & Klassen, 2016), evaluar las tendencias de opinión respecto a la vacunación contra el virus del papiloma humano utilizando datos de Twitter (Du, Xu, Song, & Tao, 2017); examinar las tendencias temporales, la distribución geográfica y los correlatos demográficos de creencias antivacunas con base en información capturada de Twitter (Tomeny, Vargo, & El-Toukhy, 2017), demostrar el uso de Twitter como un método de vigilancia de brotes de Ébola en tiempo real para monitorear la diseminación de la información, capturar la detección temprana de epidemias y examinar el contenido del conocimiento público y las actitudes de una población en específico (Odlum & Yoon, 2015),e identificar temas y patrones de un corpus de Tweets con el fin de obtener una comprensión más amplia de los comportamientos del uso no médico de medicamentos formulados (NMUPD, por sus siglas en inglés) (Kalyanam, et al. 2017).

Por consiguiente, estos estudios se enfocan en analizar problemas de salud pública como gripes o influenzas, creencias anti-vacunación, control de diabetes, uso de medicamentos, mortalidad por cáncer, transmisión del VIH entre personas que se inyectan drogas (PWID, por sus siglas en ingleés), búsqueda de información entorno a la salud pública, vigilancia de brotes de Ébola, epidemias y calidad del aire, con el fin de mejorar los sistemas de vigilancia, mejorar la salud de la población, y el conocimiento de la información por parte de la población.

El análisis de contenido se enmarca en un proceso de descubrimiento de conocimiento por medio de la minería de datos; siendo el descubrimiento de conocimiento (KDD, por sus siglas en inglés**),** según Fayyad, Piatetsky-Shapiro, & Smyth (1996): “el proceso no trivial de identificar patrones valiosos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos”, y como se observa en la Figura 1 contiene como una de sus fases la minería de datos,

definida por Witten & Frank (2000) como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos.

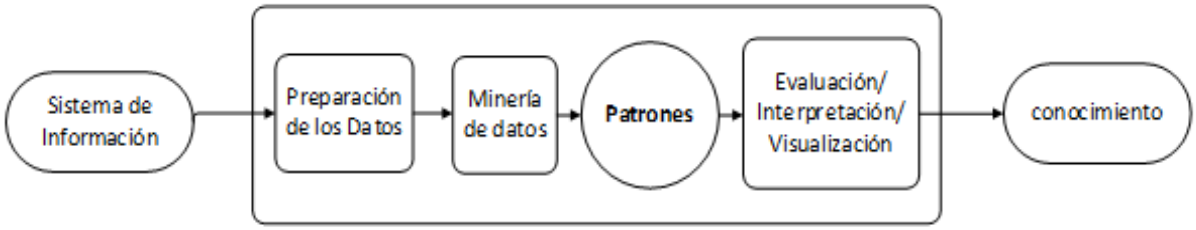


Figura *1*: Proceso de KDD. Adaptado de "Introducción a la minería de datos" , de Hernández Orallo et al., p.13, 2004, Madrid, España. Editorial Pearson

Así, el proceso de KDD permite: la selección, limpieza, transformación y proyección de los datos, analizando los datos para extraer patrones y modelos adecuados, evaluando los patrones para convertirlos en conocimiento que apoye la toma de decisiones, y hacer que el conocimiento esté disponible para su uso. Esta definición establece la relación entre el KDD y la minería de datos (Hernández Orallo et al, 2004)

Los datos son obtenidos de las redes sociales, por medio de diferentes herramientas, como por ejemplo: Massey, et al. (2016) utiliza dos métodos para recopilar datos de Twitter relacionados con la vacunación contra el virus del papiloma humano (VPH), el primero es una recolección prospectiva de datos con ayuda del API de Twitter y la segunda es una recolección retrospectiva a través de Microsoft Research; del mismo modo, (Kagashe, Yan, & Suheryani, 2017)usan datos históricos de Twitter en su estudio, por medio del API de Twitter y la palabra clave "gripe". En china, (Wang, Paul, & Dredze, 2015) usaron el API público de Weibo para extraer mensajes de este microblog con el fin de investigar si los datos de las redes sociales se pueden utilizar tanto para identificar las tendencias de la calidad del aire como analizar la respuesta del público en China respecto a este tema.

La preparación de los datos tiene como objetivo hacer que los datos del corpus sean más consistentes para facilitar la representación del texto (Hu & Liu, n.d.); por ejemplo, la información extraída de Twitter, contiene gran cantidad de tweets que no son relevantes (ruido), por un lado se tienen mensajes con poca información textual y, por otro, mensajes con baja calidad que comprende errores ortográficos, uso de signos de exclamación, entre otros (Diéguez, 2014), por lo tanto se deben usar técnicas de filtrado que permitan obtener de datos relevantes.

La representación consiste en el tratamiento preliminar de los datos, es decir su transformación. En este proceso se realizan operaciones de agregación o normalización, consolidando los datos para descubrir conocimiento en el texto; por ejemplo, Kagashe, Yan. & Suheryani (2017) usan un clasificador de aprendizaje automático, máquinas de soporte vectorial (SVM, por sus siglas en inglés) con una función de núcleo lineal ajustando los parámetros C y gamma, con el fin de extraer identificar las tres drogas más consumidas durante una influenza estacionaria; este clasificador fue comparado con un clasificador basado en la frecuencia de término de unigrama (TF) y un clasificador basado en clave (palabra clave o léxico), estos puntos de referencia son elegidos porque se han utilizado ampliamente para separar tweets de individuos con experiencia real (personal o válida). Por otro lado, Odlum & Yoon (2015), transforman el texto en forma de vector y N-gram (unigrama, bigrama y trigrama) para agruparlos según las similitudes de percepciones públicas de Enfermedad del virus del Ébola (EVD, por sus siglas en inglés). También (Mesas Jávega, 2015), utiliza los mensajes escritos por la cuenta que se desea estudiar para encontrar los términos más relevantes obtenidos de una matriz de termino-documento para almacenar la frecuencia de cada termino en cada documento, eliminando los términos que tenían una frecuencia total menor al 5%; posteriormente, con los términos más relevantes se crea una matriz de distancias euclidianas entre términos, a la cual se aplican algoritmos de agrupamiento.

Como fases importantes del proceso de descubrimiento de conocimiento se encuentran la minería de datos y el aprendizaje automático. Especialmente los modelos no supervisados, son eficientes para encontrar automáticamente patrones en los datos y resumir el contenido del corpus de texto, dado que los algoritmos son computacionalmente eficientes y los resultados del análisis de contenido tienen la ventaja de no ser afectados por el sesgo del juicio humano (Kalyanam et al., 2017). Como técnicas para el descubrimiento del conocimiento se encuentran: la asignación latente de Dirichlet (LDA, por sus siglas en inglés), usada por Kagashe, Yan & Suheryani (2017), para encontrar temas de tendencia ajustados de los tweets de cada medicamento relacionados con el consumo de drogas durante una influenza estacionaria, se destaca que esta técnica permite descubrir automáticamente temas ocultos de una colección de documentos de texto representados como una bolsa de palabras; otro uso de esta técnica fue dado por Wang, Paul, & Dredze (2015), quienes la usan con el fin de filtrar los mensajes que pertenecen a temas relevantes para la calidad del aire o la contaminación, resaltando que al estimar los parámetros de este modelo cada documento tiene una distribución discreta sobre "temas" y cada tema tiene una distribución discreta sobre las palabras, las distribuciones de palabras específicas del tema suelen dar una alta probabilidad de ocurrencia a las palabras que tienden a aparecer juntas en los documentos. Por lo tanto, cada tema puede interpretarse como un grupo de palabras tópicamente o semánticamente coherentes. Estos parámetros se derivan por completo de un corpus de texto en bruto, lo que permite que el modelo aprenda temas específicos de los datos de interés.

Otro método es el llamado Modelo Tema Bi-término (BTM, por sus siglas en inglés), usado por Kalyanam,et.al (2017), para identificar k-temas subyacentes en el corpus, y para descomponer cada tweet en términos de los temas identificados, permitiendo recuperar los tweets más correlacionados con un tema en particular; la verificación de la calidad de los temas se realiza

mediante dos evaluaciones, la primera una evaluación supervisada la cual implica etiquetar cada tweet como relevante o irrelevante ya que varios tweets son considerados falsos positivos (Retweet), la segunda comprende una evaluación no supervisada utilizando la métrica llamada "pureza del agrupamiento", que cuantifica la coherencia de un tema, es decir, si los tweets de un tema son muy diversos en términos de contenido, ese tema es considerado irrelevante. También Mesas Jávega (2015), utiliza algoritmos de agrupamiento (UPGMA, DIANA, K-means, PAM, CLARA y Modelado de Mixturas de distribuciones normales) considerando entre 2 y 8 grupos, evaluados por métricas internas que generan una evaluación numérica para identificar el algoritmo que mejor representa los datos analizados; la asociación obtenida a partir de los algoritmos de agrupamiento permite analizar qué y cómo se está propagando el mensaje y si alguno de los grupos está formado por las palabras claves de una campaña publicitaria.

4. Marco de Referencias

(Sanabria Ruiz, 2017), en su trabajo de grado titulado "aplicación de técnicas de agrupamiento (clustering) para el análisis estadístico de tendencias en Twitter basado en el lenguaje de programación r", hace referencia en la revisión de literatura sobre la búsqueda de técnicas de agrupamiento utilizadas en el aprendizaje automático; además, propone evaluar las diferentes combinaciones de las variantes del algoritmo k-means (algoritmo de *clustering*) y las distancias (Euclidiana, Máximum, Manhattan, Minkowski) con una base de datos del benchmark obtenidas de un conjunto de entrenamiento para la recopilación de spam de SMS de teléfonos móviles, para seleccionar la variante que tenga los resultados más eficientes para su posterior aplicación en un caso de estudio real, para lo cual se extraen datos utilizando el API de Twitter por medio del lenguaje de programación R, permitiendo obtener información relevante como por ejemplo los

temas más frecuentes en el caso de estudio.

Se destaca el desarrollo de esta investigación porque se realiza una comparación entre algoritmos correspondientes a técnicas de clustering aplicadas a un repositorio objeto de estudio extraído de una red social (Twitter), con el fin de seleccionar el algoritmo que proporciona resultados eficientes y relevantes.

Mesas Jávega (2015), en su trabajo de grado titulado "análisis de tendencias y marcas deportivas a través de Twitter", explica la creación de una herramienta para el análisis de campañas publicitarias lanzadas en Twitter, que consiste en el desarrollo de tres módulos: el módulo de extracción, el módulo de preprocesamiento y el módulo de análisis, que a su vez se divide en el módulo de *clustering* y módulo de análisis de campañas.

En el módulo de extracción, se extraen tweets escritos o que mencionan cuentas de deportistas y marcas seleccionadas por medio del REST API de Twitter y son almacenados utilizando MongoDB.

En el módulo de preproceso, usan los tweets almacenados en el módulo anterior con el fin de eliminar términos no relevantes como lo son: caracteres especiales, signos de puntuación, números, URLs, espacios, tabulaciones, entre otros.

El módulo de análisis, se divide en dos partes: la primera parte, en el módulo de *clustering*, utilizan los mensajes escritos por la cuenta que se desea estudiar para encontrar los términos más relevantes obtenidos de una matriz de termino-documento usando la librería de texto de R llamada *text mining* (*tm)* para almacenar la frecuencia de cada termino en cada documento, eliminando los términos que tienen frecuencia total menor al 5%. Posteriormente, con los términos más relevantes crean una matriz de distancias euclidianas entre términos, a la cual aplicaron algoritmos de clúster

(UPGMA, DIANA, K-means, PAM,CLARA y Modelado de Mixturas de Normales) a entre 2 y 8 clúster, evaluados por métricas internas que generaron una evaluación numérica para identificar el algoritmo que mejor se ajusta a los datos analizados, de la asociación obtenida a partir de los clúster se analiza qué y cómo se está propagando su mensaje y si alguno de los clúster está formado por las palabras claves de una campaña publicitaria. Si se detecta que algún clúster representa a una campaña publicitaria, se usan las palabras que la componen para obtener tweets que hablen de ella y poder analizar donde y con quien se está propagando, entrando así a la segunda parte del módulo de análisis, el módulo de análisis de campañas, en particular, se supone que los tweets que hablan de la campana son aquellos que mencionan a la cuenta en estudio y a alguna de las palabras clave de esta. De este modo, dichos tweets son geolocalizados y se crean mapas de impacto para estudiar el alcance de dicha campaña y el lugar en dónde está teniendo más relevancia, también se estudió la red que se forma entorno a la campaña para poder así detectar aquellos usuarios que han sido determinantes en su propagación, por medio de la creación de un grafo que generan las menciones de los tweets que hablan de la campana. Sobre dicho grafo se aplican medidas de centralidad, en particular el algoritmo HITS.

De este trabajo se destaca la importancia del desarrollo de dicha herramienta para el presente proyecto, ya que contiene de manera detallada la creación de la herramienta que permitirá conocer la relevancia y vitalización (impactos) de las campañas deportivas a través de Twitter.

García Diéguez (2014), en su proyecto de fin de carrera titulado "técnicas de clustering aplicadas al análisis de *trending topics* en conjunto de tweets" presentan y evalúan simultáneamente dos mecanismos distintos.

En el primero proponen he implementan un conjunto de algoritmos mediante los que obtienen una serie de puntuaciones que los ayudan en la tarea de determinar el carácter de trading tópicos

de un término. Para esta primera parte utilizan un conjunto de tweet ya dado, pasando por una etapa de preprocesado, tratamiento, parametrización y representación. En la etapa de preprocesamiento ofrece dos opciones un filtrado por tweets individuales o por fragmento de texto eliminando características no relevantes. En la etapa de tratamiento se obtiene un vector de términos representativos detectando hashtags, palabras o ambas a la vez; según el usuario que interactúa con el mecanismo. En la etapa de parametrización emplean 2 métodos (N-grams y elección del formato fecha) y 4 algoritmos (el algoritmo de frecuencia de término i "*Term frequency",* el algoritmo de frecuencia de término ij, el algoritmo TF –IDF, El algoritmo NTF "*Normalized Term Frequency*") con el fin de evaluar los términos representativos recibidos. En la última etapa, la etapa de representación, se dispone de manera visual todos esos datos, para posteriormente realizar un análisis exportando automáticamente todos los resultados a ficheros de Excel, para realizar la representación gráfica.

En el segundo plantean diversas estrategias que, junto con la ayuda de los algoritmos, les permite contemplar distintos escenarios y, por lo tanto, detectar temas más generales o más específicos según sus intereses. Esto implica que además de detectar *trading topics*, investigan con más detalle los temas sobre los que se hablan. En este mecanismo, utilizan la detección de hashtags y/o la detección de palabras. Se crean diferentes n-gramas obtenidos según el comportamiento de los mismos y se evalúan por medio de ventana deslizante, combinaciones sin repetición, analizando cada uno de los algoritmos implementados (el algoritmo de frecuencia de término ij, algoritmo TF –IDF, El algoritmo NTF.

Se cita este trabajo porque permite evaluar qué mecanismos ofrecen mejores resultados y por lo tanto, deducir cuáles son los que más se asemejan al uso real que hace Twitter en la detección de *Trading Topics*, mostrando así que la frecuencia es uno de los principales requisitos para que

un término sea *Trading Topics* y se trata de una característica necesaria, pero no suficiente y por ello se evalúan diferentes algoritmos de frecuencia, por otro lado concluye que la detección de hashtags es el método más eficaz para detectar posibles *Trading Topics*.

5. Marco Teórico

5.1 Salud Pública.

De acuerdo con la Ley 1122 de 2007 la salud pública está constituida por un conjunto de políticas que busca garantizar de manera integrada, la salud de la población por medio de acciones dirigidas tanto de manera individual como colectiva ya que sus resultados se constituyen en indicadores de las condiciones de vida, bienestar y desarrollo. Dichas acciones se realizarán bajo la rectoría del Estado y deberán promover la participación responsable de todos los sectores de la comunidad. (Ministerios de salud y protección social, s.f.)

5.2 Programas de promoción y prevención.

Proceso para proporcionar a las poblaciones los medios necesarios para mejorar la salud y ejercer un mayor control sobre la misma, mediante la intervención de los determinantes de la salud y la reducción de la inequidad. Esto se desarrolla fundamentalmente a través de los siguientes campos: formulación de política pública, creación de ambientes favorables a la salud, fortalecimiento de la acción y participación comunitaria, desarrollo de actitudes personales saludables y la reorientación de los servicios de salud; por sus características la promoción de la salud supone una acción intersectorial sólida que hace posible la movilización social requerida para la transformación de las condiciones de salud. (MINSALUD, s.f)

5.3 Impacto en redes sociales

El impacto en la red social se enmarca dentro del monitoreo de la red social de acuerdo a la campaña, publicación que se realice (Chona, Londoño y Gross Beltrán, 2013), de lo anterior de debe tener en cuenta que existen diferentes métricas para el monitoreo de la red social , para ello (Chona et al., 2013) , definen que las métricas que existen son:

- Métricas de exposición calificada: Son aquellas que miden las conexiones que se han creado atreves de las plataformas de redes sociales; por ejemplo, cantidad de fans, seguidores, o contactos que tienen
- Métricas Estratégicas: Son aquellas que miden la forma como las personas valoran una publicación; por ejemplo, el número de menciones de usuarios, número de tuits, el número de me gustas, número de veces que comparten la publicación, entre otros.
- Métricas financieras: permiten medir cuanto cuesta cada publicación mirando el costo por cada interacción o conexión, costo por cada punto porcentual en que la marca, publicidad consiga mejorar su posicionamiento

5.4 Redes Sociales

Las redes sociales se enmarcan dentro de la Web 2.0 en donde el usuario participa como contribuidor activo y no solo como espectador de los contenidos de la Web (usuario pasivo).

5.4.1 Twitter. Es una Red social fundada en la ciudad de San francisco en el año 2006, una de sus características es que permite twittear mensajes de texto plano de corta longitud, con un máximo de 280 caracteres, llamados tweets (Romero Salamanca, 2018). También permite la interacción con otros usuarios por medio de hashtags, menciones y retweets.

5.4.2 API de Twitter. Twitter comparte información, y proporciona a las empresas, los

desarrolladores y los usuarios acceso programático a los datos de Twitter mediante sus API (interfaces de programación de aplicaciones), a través de ventanas de usuario o de aplicación. Las APIs de Twitter cuentan con cinco puntos de conexión y tres niveles de acceso a datos para satisfacer las necesidades de las aplicaciones en todas las etapas de crecimiento; de acuerdo al punto final de la API se enmarca en alguna de las 5 conexiones y a un nivel de acceso como se observa en la Tabla 2.

Tabla 2
Ejemplo de puntos finales para la dimensión "Tweets y respuestas" en los tres niveles de acceso

	Tweets y respuestas	Nivel de acceso		
		Estándar	Premium	Enterprise
Puntos finales	Publicar y participar	☒		
	Tweets de búsqueda: 7 días	☒		
	Tweets de búsqueda: 30 días		☒	☒
	Tweets de búsqueda: archivo completo		☒	☒
	Filtrar tweets	☒		☒
	Tweets de muestra	☒		☒
	Tweets por lotes			☒

Nota: Adaptado de "Pricing — Twitter Developers", de Twitter INC, 2018.

Los cinco puntos de conexión que incluyen las API`s de Twitter son (Twitter INC, 2018):

- Cuentas y usuarios: permite a los desarrolladores administrar de forma programática el perfil y la configuración de una cuenta, silenciar o bloquear usuarios, administrar usuarios y seguidores, solicitar información sobre actividad de una cuenta autorizada, entre otros.
- Tweets y respuestas: Los desarrolladores pueden acceder a los Tweets al buscar palabras clave específicas, solicitar una muestra de Tweets de cuentas específicas y publicar tweets a través de su API.
- Mensajes directos: Brinda acceso a las conversaciones por mensajes directos (DM, por sus siglas en inglés) de usuarios que han otorgado permiso de forma explícita a una aplicación específica.

- Anuncios: Ofrece un conjunto de API que permite a los desarrolladores, ayudar a las empresas a crear y administrar automáticamente campañas de anuncios en Twitter. Los desarrolladores pueden usar Tweets públicos para identificar temas e intereses, y así brindar a las empresas herramientas para ejecutar campañas publicitarias a fin de alcanzar las diversas audiencias en Twitter.
- Herramientas y SDK del editor: Proporciona herramientas para desarrolladores y editores de software que permiten integrar las cronologías de Twitter, el botón para compartir y otros contenidos de Twitter en las páginas web.

Los tres niveles de acceso a datos que permite Twitter son (Twitter INC, 2018):

- Standard APIs: Son APIs de acceso gratuito y con una complejidad de consulta básica
- Premium APIs: Son APIs de acceso escalable a los datos de Twitter para aquellos que buscan crecer, experimentar e innovar, con contratos flexibles mes a mes.
- Enterprise APIs: Son APIs de acceso pago, ofrecen el más alto nivel de acceso y confiabilidad para quienes dependen de los datos de Twitter

Obtener el acceso al API de Twitter. Para tener acceso al Api de Twitter se hace mediante la autenticación por usuario o por aplicación, a continuación, se describen los pasos para la creación de la aplicación.

- Ingresar a https://apps.twitter.com/
- Iniciar sesión en Twitter
- Dar click en “create new app”
- En “*Aplication Details*”:
    - *Name*: Colocar el nombre de la aplicación
    - *Description*: Colocar una descripción de la aplicación

- *Website*: La dirección URL de su cuenta de Twitter
- *Callback URLs*: La dirección URL de su cuenta de Twitter

- Aceptar las "*Developer Agreement*"

Cabe resaltar que, a partir del 24 de julio del 2018, Twitter no permite la creación de aplicaciones de Twitter a través de apps.twitter.com. Ahora se le redirigirá a su cuenta del portal para desarrolladores en developer.twitter.com (Twitter INC, 2018).

Obtención de las llaves de acceso al API de Twitter. A continuación, se detalla el procedimiento para obtener las llaves de acceso:

- Ingresar a la aplicación creada en el numeral 4.3.2.1 por medio de https://apps.twitter.com/
- En la pestaña *Keys and Access Tokens* encuentra:
  - *Consumer Key (API Key)*
  - *Consumer secret (API secret)*
  - *Acceess Token*
  - *Access token Secret*

Impacto en redes sociales

5.5 Representación de la información

5.5.1 Corpus de texto. Un corpus puede ser cualquier colección que contenga más de un documento, o bien como un corpus lingüístico donde el número de documentos no es relevante, pero si otras connotaciones como pueden ser la autoría de los documentos (por ejemplo, dos de los documentos estén escritos por el mismo autor) o el tipo de documentos que alberga la colección (Hernández, 2002)

Según Atkins (1992) existen cuatro tipos genéricos de colecciones de texto:

- Archivos: se denomina así al conjunto de textos sin relación aparente entre sí, que se almacenan en formato magnético en un lugar común.
- Bibliotecas de texto en formato magnético (ETL o *Electronic Text Library*): se trata del conjunto de archivos en formato magnético que presenta conatos de estandarización, pero sin criterios de selección definidos.
- Corpus: es aquel subconjunto de una ETL que se ha creado con unos criterios definidos y con un propósito determinado.
- Subcorpus**:** cualquier subconjunto de un corpus cuya selección se realiza de forma dinámica durante las consultas ha dicho corpus.

5.5.2 Transformación de un archivo textual a un archivo numérico. La transformación de un documento, que contiene palabras, a un vector de números con el cual los algoritmos pueden trabajar. Por ejemplo, se tienen estos tres documentos (cada frase es un documento):

Primer documento: Juega al futbol.

Segundo documento: Le gusta el baloncesto.

Tercer documento: Mira un partido de futbol.

La siguiente matriz tiene como filas a los documentos, y una columna por cada palabra diferente que hay en el total de documentos (vocabulario). La idea es colocar la frecuencia de la palabra en el documento por cada palabra. De esta manera el documento 1 tiene las palabras “juega”, “al” y “futbol” por lo que la fila uno tiene valor 1 para esas palabras y 0 para el resto.

	juega	al	futbol	le	gusta	el	Baloncesto	Mira	Un	partido	de
1	1	1	1	0	0	0	0	1	1	1	1
2	0	0	0	1	1	1	1	0	0	0	0
3	0	0	1	0	0	0	0	1	1	1	1

Figura 2: Ejemplo matriz Tf. Recuperado de "Algoritmos de Clustering y Aprendizaje Automático Aplicados a Twitter", De Blanco Eric-Joel and Sanz Hermida, 2016, pág. 20. Cataluña

Los documentos 1 y el 3 son similares dado que ambos hablan de fútbol. Sin embargo, si se aplica la similitud coseno la cual es una medida de semejanza que evalúa el valor del coseno del ángulo comprendido entre dos vectores (documentos), si no se parecen el valor es 0 y 1 si los documentos son idénticos, de esta manera para los documentos 1 y 3 se obtiene un valor de 0.258, para los documentos 1 y 2 y los documentos 2 y 3 el resultado de la similitud da cero (0), por lo que en este ejemplo funcionan bien estas técnicas (Blanco and Sanz , 2016).

5.5.3 Documento termino matriz o un término documento matriz. Una manera rápida de explorar los datos textuales es simplemente contar las ocurrencias de cada palabra o término. El número de veces que una palabra en particular aparece en un documento dado se denomina frecuencia de término (*tf*). En estadística *tf* se puede resumir en una matriz de términos de documentos, que es una matriz rectangular con filas que representan documentos y columnas que representan términos únicos. El elemento (i, j) de esta matriz proporciona los conteos del término j-ésima (columna) en el i-ésimo documento (fila). También es posible transponer la matriz convirtiéndola en una matriz de términos -documentos donde las filas y columnas representan términos y documentos respectivamente. (Imai, 2017)).

5.5.4 Frecuencia de término (o por sus siglas en ingles TF). El método calcula el número de repeticiones de una palabra (término) en el documento (Trstenjak, Mikac, and Donko, 2014):

$$a_{ij} = f_{ij} = frecuencia\ del\ termino\ i\ en\ el\ documento\ j \quad (1)$$

5.5.5 Frecuencia de término. Frecuencia inversa del documento del término i (por sus siglas en ingles TF-IDF). El método *Tf-Idf* determina la frecuencia relativa de las palabras en un documento específico a través de una proporción inversa de la palabra en todo el cuerpo del documento. En la determinación del valor, el método incluye dos elementos (Trstenjak, et.al, 2014):

- TF: término de frecuencia del término i en el documento j
- IDF: frecuencia inversa de documento del término i.

TfIdf se puede calcular como (Trstenjak et al., 2014):

$$a_{ij} = tf_{ij}idf_i = tf_{ij} \times log_2(N|df_i) \quad (2)$$

Dónde:

a_{ij} , es el peso del término i en el documento *j*

N, es el número de documentos en la colección

tf_{ij} , es la frecuencia del término i en el documento *j*

df_i . Es la frecuencia del documento i en la colección

Esta fórmula arroja buenos resultados cuando es utilizada en documentos con igual longitud, sin embargo para obtener mejores resultados con documentos de diferente longitud se usa la siguiente ecuación modificada (Trstenjak et al., 2014):

$$a_{ij} = tf_{ij}idf_i = \frac{f_{ij}}{\sqrt{\sum_{s=1}^{N}(tfidf(a_{sj}))^2}} \times log_2(N|df_i) \quad (3)$$

5.6 Aprendizaje Automático (*Machine Learning*)

Según Mitchell (1997), “un programa de computadora aprende de la experiencia *Experience* (E) con respecto a alguna tarea *Task* (T) y alguna medida de rendimiento *Performance* (P), si el rendimiento en T, medido por P, mejora con experiencia E.” (p.3). Por ejemplo, un programa de computadora que aprende a jugar damas podría mejorar su rendimiento medido por su habilidad para ganar en la clase de tareas que involucran jugando a los juegos de damas, a través de la experiencia obtenida jugando juegos contra sí mismo

5.6.1 Clasificación Supervisada y no supervisada. La clasificación es el proceso mediante el cual se agrupa un conjunto de elementos en función de una representación de los mismos en un espacio n-dimensional. En el caso de que sea supervisada, se tiene información acerca de etiquetas previas, de manera que se sabe a qué grupo pertenece cada clase, entonces lo que se desea es encontrar un conjunto de “criterios”, probablemente reglas, que permita, dado un nuevo elemento, situarlo en un grupo. En el caso de la clasificación no supervisada no se tiene información sobre las características de los grupos que se podrían conformar, además no se conoce la cantidad de agrupaciones a conformar, entonces el fin consiste en encontrar un agrupamiento que reúna en un mismo grupo los elementos más “parecidos” y coloque en grupos diferentes a los elementos “dispares”. Por lo tanto, hay una gran diferencia entre ambos métodos de clasificación, de hecho, en muchos casos el agrupamiento puede verse como el primer paso para proceder a realizar la clasificación supervisada (Gonzales et al., 2008, p.693.).

5.6.2 Técnicas de clasificación no supervisada.

5.6.2.1 Técnicas de Agrupamiento (Clustering). Pertenece a una técnica de clasificación no supervisada, y no se tiene ninguna información acerca de la organización de los elementos en

grupos o clases, por lo tanto, el objetivo es encontrar dicha organización con base en la proximidad de los elementos. No existe información previa acerca de esta estructura, y la interpretación de las clases o grupos obtenidos es una labor a realizar "a posteriori" por parte del analista. Habitualmente estos elementos están representados por un vector de valores de atributos, es decir, son puntos de algún espacio multidimensional. Estos valores también se suelen denominar atributos, componentes o simplemente variables. Intuitivamente, dos elementos pertenecientes a un agrupamiento válido deben ser más parecidos entre sí que a aquellos elementos clasificados en grupos distintos. Partiendo de esta idea se desarrollan las técnicas de agrupamiento. Obviamente, estas técnicas dependen de cómo sean los datos de partida, de que medidas de semejanza se estén utilizando y de qué clase de problemas se estén resolviendo. Un ejemplo de clasificación sería encontrar los elementos que determinan la aparición de cáncer de pulmón analizando datos de edad, calidad de vida, nivel económico, etc., tanto de personas enfermas como sanas (Gonzales et al., 2008, p.691).

5.6.3 Agrupamiento particional y agrupamiento jerárquico. El objetivo final del proceso de agrupamiento es obtener un conjunto de clases o grupos. Cuando estos grupos son disjuntos y cubren todo el conjunto de elementos se dice que el agrupamiento es particional. En algunos casos lo que se desea no es exactamente un agrupamiento particional sino una jerarquía de agrupamientos "anidados", de tal manera que cada grupo de un nivel se divide en varios en el nivel siguiente; esta estructura se denomina agrupamiento jerárquico y tiene una representación gráfica muy intuitiva denominada dendograma.

5.6.4 Métodos de clasificación no supervisada.

5.6.4.1 K medias. El algoritmo de K medias (*k means*) se trata de un método de agrupamiento por vecindad en el que se parte de un número determinado de prototipos y de un conjunto de

ejemplos a agrupar, sin etiquetar. La idea del k-medias es situar a los prototipos o centros en el espacio, de forma que los datos pertenecientes al mismo prototipo tengan características similares [Moody & Darken 1989, MacQueen 1967] Citado en (Gonzales et al., 2008, p.704.).

5.6.4.2 Latent Dirichlet Allocation (LDA). Es un modelo generativo que permite que conjuntos de observaciones puedan ser explicados por grupos inadvertidos que describen por qué algunas partes de los datos son similares. Por ejemplo, si las observaciones son palabras en documentos, cada documento es una mezcla de categorías y la aparición de cada palabra en un documento se debe a una de las categorías a las que el documento pertenece. Entonces, la intuición detrás de la técnica LDA es que cada documento siempre exhibe múltiples temas en su cuerpo.

5.6.4.3 Partición Alrededor de Medoids (PAM). Según Kaufmann (1990), el algoritmo PAM es uno de los más antiguos y más sencillos. Este algoritmo se aplica según el siguiente procedimiento:

- Se parte de una selección inicial de k-medianas. Como en el caso de las k medias, estas medianas se utilizan para generar una partición en k grupos del conjunto total de elementos.
- Para cada grupo obtenido se intenta reemplazar la mediana asociada a él por algún punto del mismo grupo que sea más idóneo. Esto se hace considerando cada punto del grupo como candidato y calculando la distancia total del resto de elementos del grupo a dicho punto. Si hay mejora con respecto a la mediana actual, este es sustituido por el punto en cuestión, en caso contrario se mantiene.
- Se recalculan los grupos con base en las nuevas medianas seleccionadas y se vuelve con ellos al paso anterior.
- El proceso termina cuando no se cambian las medianas en una iteración.

Este método es más robusto que el de las k-medias en presencia de datos atípicos. Citado en (Gonzales et al., 2008, p.712.)

5.6.4.4 CLARA. Según Kaufmann (1990), el agrupamiento para grandes aplicaciones (por sus siglas en inglés CLARA), es utilizado cuando se trabaja con grandes bases de datos, a diferencia del algoritmo PAM que se aplica a datos pequeños. CLARA se basa en un proceso de muestreo. La idea básica es que, si se toma una muestra suficientemente aleatoria y representativa, las medianas obtenidas con ella serán los mismos que los del conjunto local. Con esta idea, CLARA realiza sucesivos muestreos y aplica PAM a cada uno de ellos. Los conjuntos de medianas obtenidos se analizan sobre el conjunto total, seleccionando el que permite obtener menor distancia global. Con estas medianas se genera un nuevo proceso de muestreo y una nueva iteración.

5.6.4.5 DBSCAN. Método que se basa en el análisis de densidad. Los métodos de agrupamiento basados en la densidad intentan encontrar las regiones del espacio con una alta densidad de elementos, que estén separados por otras regiones de baja densidad. Obviamente todos los métodos de esta categoría se basan en el concepto de "densidad" de una región, habiéndose desarrollado diversas formas de medirla, algunas de las cuales tienen fundamento estadístico. En este caso, se parte del conocimiento de la distribución de probabilidad conjunta entre elementos y grupos para obtener la distribución de probabilidad asociada a cada grupo y, a partir de ella, se desarrolla un procedimiento de asignación de elementos a los grupos.

5.7 Minería de datos a partir de datos de Twitter

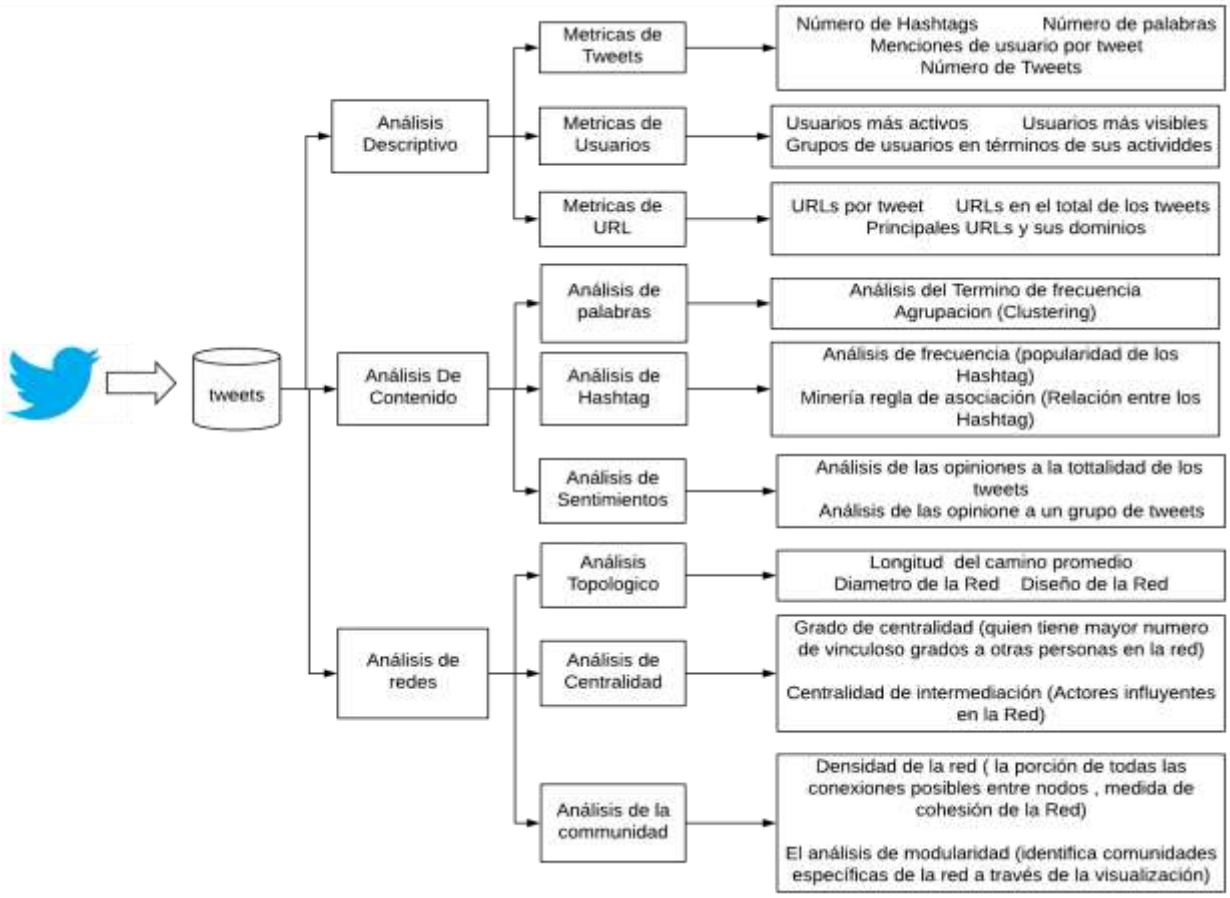


Figura 3: Un marco propuesto de extraer inteligencia de Twitter (Twitter Analytics). Adaptado de "Insights from hashtag #supply chain and Twitter Analytics: Considering Twitter and Twitter data for supply chain practice and research", de Chae , 2015, p 250

El potencial de las técnicas de la Minería de Datos está comenzando a ser reconocido y aprovechado para el análisis de datos producidos en redes sociales, generando así una disciplina llamada Minería de Redes Sociales *(Social Media Mining).* Dentro de esta minería de redes sociales y de acuerdo a los datos que la red social proporciona, su análisis se enmarca en tres grandes grupos (ver Figura 3).

5.7.1 Análisis Descriptivo. Se enfoca en métricas cuantitativas de los usuarios, tweets y localizador uniforme de recursos (URL, por sus siglas en ingles), de los tweets por medio del

número de veces que un usuario ha sido mencionado (@) en los tweets, la cantidad de hashtags que se usan en el tweet, el número de tweets que hace el usuario; los usuarios por medio del número de menciones que ha tenido el usuario y finalmente las URLs , dado que estás aportan información relevante al tweet, y se mide por el dominio de la URLs , el número de URLs que hay en el Tweet y la información que este contiene.

5.7.2 Análisis de contenido. Dado los grandes volúmenes de datos proporcionados por los usuarios se hace necesario analizar los textos que estos publican en sus redes sociales mediante la detección de tendencia de tópicos en la búsqueda de temas emergentes y la evolución de su tendencia a partir del texto del tweet; otra forma es comprender mejor de lo que están hablando los usuarios intentando adjudicar un valor de polaridad para poder identificar si un texto es positivo, negativo o neutral mediante el análisis de sentimiento.

5.7.3 Análisis de Redes. Se enfoca en particular en representar las relaciones entre los usuarios, como quién sigue a quién o quién reenvía un mensaje de otro usuario, puede plantear cuestiones como quiénes son los usuarios más relevantes, quién responde a un producto, quién habla con quién o cómo se agrupan, entre otras.

Siendo así dos análisis importantes, la búsqueda de usuarios más relevantes y la búsqueda de comunidades dentro de la red. La búsqueda de usuarios parte de medir el grado de importancia de cada usuario (nodo) , su influencia sobre el resto de usuarios y en consecuencia la capacidad para propagar información dentro de la red, esto se hace mediante medidas de centralidad como: la centralidad del grado para medir la importancia de cada usuario en la red (Ejemplo: las veces que retwittea y las veces que ha sido retwitteados), la centralidad de intermediación para medir la capacidad de propagar información (Ejemplo: que tan rápido puede trasmitir información en la red, distancias cortas entre usurarios), y por último la centralidad de cercanía para medir la

influencia sobre los demás usuarios (Ejemplo: Tener conexión con la mayor parte de los nodos o usuarios). Finalmente, la búsqueda de comunidades, una comunidad en una red se suele considerar como un grupo de nodos con más y/o mejores interacciones entre sus miembros que entre sus miembros y el resto de la red.

5.8 Latent Dirichlet Allocation (LDA)

El LDA es un modelo probabilístico generativo para colecciones de datos discretos, como el cuerpo de un texto (Blei, Ng, & Jordan, 2003), proporcionando el mecanismo para encontrar patrones de coocurrencia de términos y el uso de esos patrones para identificar temas coherentes (Aggarwal et al., 2012). La idea básica del modelo LDA es que los documentos en el corpus se representan como una mezcla al azar sobre los temas latentes y cada tema se caracteriza por una distribución de las palabras en los documentos.

Se debe tener siempre en cuenta que el algoritmo no tiene información externa sobre los temas encontrados y los documentos no están etiquetados con los temas o palabras claves (es un proceso de aprendizaje no-supervisado). La distribución de temas encontrados surge mediante el cálculo de la estructura de temas ocultos que probablemente generan la colección de documentos observados (Chandía Sepúlveda, 2016).

5.8.1 Conceptos preliminares al LDA.

5.8.1.1 Distribución de Dirichlet. Es una familia de distribuciones de probabilidad continuas multivariantes parametrizadas por un vector α de reales positivos. Es una generalización multivariada de la distribución beta.

$$Dir\ (\vartheta\ /\ \alpha) = \frac{1}{B(\propto)} \prod_{K=1}^{K} \theta_k^{\alpha_k - 1} \tag{4}$$

Dónde:

$$\sum_{k=1}^{k} \theta_k = 1\ ;\ \theta_k\ \geq 0 \tag{5}$$

5.8.2 Proceso generativo de latent Dirichlet allocation. LDA constituye el siguiente proceso generativo para cada documento en un corpus (Silvestre Gómez, 2018).

- Para cada tópico:

  a. Definir una distribución de las palabras.

$$\vec{\beta}_k \sim Dir(\eta) \tag{6}$$

$$p\ (\beta_k\ /\ \eta\ ) = \frac{1}{B(\eta)} \prod_{v=1}^{W} \beta_{kv}^{\eta - 1} \tag{7}$$

$$\sum_{v=1}^{w} \beta_{kv} = 1\ ;\ \beta_{iv}\ \geq 0 \tag{8}$$

- Para cada documento:

  a. Definir un vector de probabilidades de tema para cada documento.

$$\vec{\theta}_{\mathrm{d}} \sim \mathrm{Dir}(\propto) \tag{9}$$

$$p\ (\theta_d\ /\ \propto\ ) = \frac{1}{B(\propto)} \prod_{i=1}^{K} \theta_{di}^{\propto - 1} \tag{10}$$

$$\sum_{i=1}^{k} \theta_{di} = 1\ ;\ \theta_{di}\ \geq 0 \tag{11}$$

  b. Para cada palabra:

i. Asignar un tópico. Para ello se sigue una distribución multinomial.

$$Z_{d,n} \sim Mult(\vec{\theta}_d) \quad , Z_{d,n} \epsilon \{1, \dots, k\} \tag{12}$$

ii. Muestrear una palabra del vocabulario. Se escoge siguiendo una distribución multinomial.

$$W_{d,n} \sim Mult(\vec{\beta}_{z_{d,n}}) \quad , W_{d,n} \epsilon \{1, \dots, V\} \tag{13}$$

Tabla 3
Parámetros del LDA

Parámetro	Definición
η (**eta**)	Define la distribución de cada tópico en función de las palabras
α (Alpha)	Vector de tamaño *k* que se emplea como parámetro de la distribución de Dirichlet que define la probabilidad de que un tema ocurra en un documento, los valores pequeños de α indican que los documentos están conformados por un pequeño número de temas, caso contrario un valor alto de α indica que los documentos están conformados por un alto número de temas.
K	Número de tópicos o temas
D	Conjunto de documentos en el corpus
V	Conjunto de términos o palabras en el corpus
$\vec{\boldsymbol{\beta}}_{\boldsymbol{k}}$	Representa la estructura del tema *k*, es decir la probabilidad de que cada palabra ocurra en el tema k
$\vec{\boldsymbol{\theta}}_{\mathbf{d}}$	Es la proporción de los temas para el d-ésimo documento, es decir la probabilidad de que cada tema ocurra en el documento d
$\boldsymbol{Z_{d,n}}$	Se refiere a la asignación de un tema a la n-ésima palabra del d-ésimo documento.
$\boldsymbol{W_{d,n}}$	Es la n-ésima palabra del documento d-ésimo.

Nota: Adaptado de "Implementación de Asignación Jerárquica Latente de Dirichlet Para Modelado de Temas", de Silvestre Gómez, María, 2018, Universidad de Sevilla.

Como resultado el modelo de tema probabilístico estimado por LDA, se tienen dos tablas (matrices). La primera tabla describe las probabilidades a-posteriori de seleccionar un término cuando se muestrea un tema en particular $(\beta_{1:k,1:V})$, la segunda tabla describe la probabilidad a-posteriori de seleccionar un tema en particular al muestrear un documento $(\theta_{1:d,1:k})$.

5.9 Modelo de temas correlacionados (CTM)

Es un modelo jerárquico de colecciones de documentos. El CTM modela las palabras de cada

documento a partir de un modelo de mezcla. Los componentes de la mezcla son compartidos por todos los documentos en la colección. Las proporciones de la mezcla son variables aleatorias específicas de los documentos. El CTM permite que cada documento muestre múltiples temas con diferentes proporciones. Por lo tanto, puede capturar la heterogeneidad en datos agrupados que exhiben múltiples patrones latentes.

A continuación, se explica la terminología y notación utilizada para describir los datos, las variables latentes y los parámetros en el CTM:

- Palabras y documentos: las únicas variables aleatorias observables que se consideran son palabras que se organizan en documentos. Sea $w_{d,n}$ denote la palabra n-ésima en el documento d-ésima, que es un elemento en un V-vocabulario de termino. Sea w_d el vector de N_d palabras asociadas con el documento.
- Temas: un tema β es una distribución sobre el vocabulario, un punto en la V – 1 simplex. El modelo contendrá K temas $\beta_{1:k}$
- Asignaciones tema: cada palabra se extrae de uno de los K temas. La asignación de tema $z_{d,n}$ está asociada con la enésima palabra d del documento.
- Proporciones temáticas: finalmente, cada documento está asociado a un conjunto de proporciones de temas θ_d, que es un punto en el simplex k-1. por lo tanto, θ_d es una distribución sobre índices de temas y refleja las probabilidades con las que las palabras se extraen de cada tema de la colección, normalmente se considera una parametrización natural de esta multinomial $\eta = log(\theta_i|\theta_k)$.

El modelo de tema correlacionado asume que un documento de N palabras surge del siguiente proceso generativo. Temas dados $\beta_{1:k}$, un K-vector μ y una K×K matriz de covarianza ∑

- Extraer $\eta_d|\{\mu|\Sigma\} \sim \mathcal{N}(\mu, \Sigma)$.
- Para $n \epsilon \{1, \dots, N_d\}$:

(a) Extraer el tópico asignado $Z_{d,n}|\eta_d$ para Mult$(f(\eta_d))$

(b) Extraer la palabra $w_{d,n}|\{z_{d,n}, \beta_{1:k}\}$ para Mult$(\beta^z_{d,n})$,

Donde $f(\eta)$ mapea una parametrización natural de las proporciones de tema a la parametrización media,

$$\theta = f(\eta) = \frac{exp\{\eta\}}{\Sigma_i exp\{\eta_i\}} \tag{14}$$

Tener en cuenta que η no indexa una familia exponencial mínima. La adición de cualquier constante a η dará como resultado un parámetro de medio de idénticos. Este proceso se ilustra como un modelo gráfico probabilístico de tema correlacionado en la Figura 4.

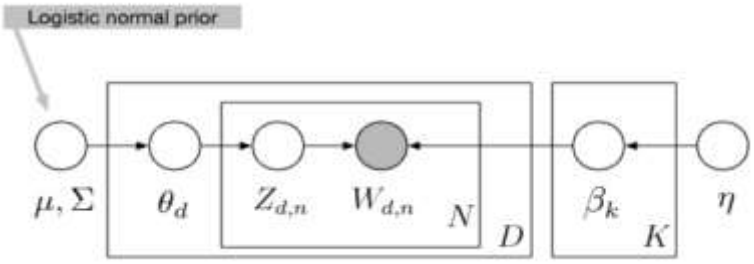


Figura 4: Modelo grafico probabilístico de Correlated Topic Models. Adaptado de Blei and Lafferty (2005). Graphical model representation of the correlated topic model. Recuperado de http://people.ee.duke.edu/~lcarin/Blei2005CTM.pdf

Según Blei and Lafferty (2005), un modelo gráfico probabilístico es una representación gráfica de una familia de distribuciones conjuntas con un gráfico, los nodos denotan variables aleatorias, los bordes denotan posibles dependencias entre ellos. En el caso de CTM la familia de distribuciones conjuntas es la distribución normal logística, utilizada para modelar las proporciones de temas latentes de un documento.

El CTM se basa en el modelo de asignación latente de Dirichlet (LDA). Sin embargo, una limitación de LDA es la incapacidad de modelar la correlación del tema; está limitación proviene del uso de la distribución de Dirichlet para modelar la variabilidad entre temas, por lo cual el CTM ofrece un mejor ajuste que LDA en una colección de artículos. Además, la CTM proporciona un camino natural de visualizar y explorar conjuntos de datos no estructurados (Blei and Lafferty, 2005)

5.10 Estimación de los parámetros

En general, la estimación de los parámetros del LDA y CTM se lleva a cabo mediante la maximización de la probabilidad de todos los documentos. En particular, dado un corpus de documentos D = {1,…, d}, se hace una estimación de k y los parámetros de las distribuciones de Dirichlet y normal que maximizan la probabilidad de registro de los datos (Benítez Andrades and Valbuena López, 2011).

Para ajustar el modelo de LDA o la CTM a las matrices de probabilidad, el número de temas (k) debe fijarse a priori, este se puede estimar mediante muestreo de Gibbs o de VEM (explicado más adelante), sin embargo la estimación utilizando el muestreo de Gibbs requiere la especificación de valores para los parámetros de las distribuciones anteriores $(\propto, \mu)$; Griffiths y Steyvers (2004) sugieren un valor de 50 / k para los parámetros de las distribuciones en LDA y CTM. Debido a que el número de temas en óptimo se desconoce, se ajustan los modelos con diferentes números de temas y el número óptimo se determina de acuerdo a medidas como por ejemplo la de máxima Verosimilitud

5.10.1 Criterios de estimación de k. Máxima verosimilitud (ML). Para la estimación de ML del modelo LDA, la probabilidad de registro de los datos, la suma de las probabilidades de registro

de todos los documentos, se maximiza con respecto a los parámetros del modelo α y β (Grun and Hornik, 2011).

$$\ell(\propto, \beta) = \log(p(w_i| \propto_i , \beta)) \tag{15}$$

$$\ell(\propto, \beta) = \log \int \left\{ \sum_z \left[\prod_{i=1}^{N} p(w_i|z_i , \beta) p(z_i|\theta) \right] \right\} p(\theta| \propto) \, d\theta \tag{16}$$

Para CTM se maximiza respecto a los parámetros μ, Σ and β (Grün and Hornik, 2011) así:

$$\ell(\propto, \beta) = \log(p(w_i|\mu_i , \Sigma_i , \beta)) \tag{17}$$

$$\ell(\propto, \beta) = \log \int \left\{ \sum_z \left[\prod_{i=1}^{N} p(w_i|z_i , \beta) p(z_i|\theta) \right] \right\} p(\theta| \mu, \Sigma) \, d\theta \tag{18}$$

5.10.2 Métodos de estimación de los parámetros. El cálculo directo de los parámetros de las distribuciones de Dirichlet es intratable debido a la naturaleza de la computación. La solución a esto es el uso de diversos métodos alternativos de estimación aproximada. Algunos de ellos son el algoritmo esperanza-maximización (EM) y el muestreo de Gibbs.

5.10.2.1 Algoritmo esperanza-maximización o algoritmo EM. Se basa en encontrar estimadores de máxima verosimilitud de parámetros en modelos probabilísticos que dependen de variables no observables, de acuerdo a la iteración de dos pasos: un paso de esperanza (E) donde las probabilidades posteriores son calculadas con base en las actuales estimaciones de los parámetros; y un paso de maximización (M) en el que los parámetros son actualizados en función de la minimización de los criterios y la dependencia de las probabilidades posteriores calculadas en el paso E (Chandía Sepúlveda, 2016).

Para estimar los parámetros que maximizan la probabilidad total del corpus, se basa en que para

cada valor k existe un valor α, es decir que k, hace parte de un conjunto de distribuciones K. El algoritmo variacional EM se describe brevemente como sigue:

- Paso E: Para cada documento, encontrar los valores de los parámetros de optimización variacional.
- Paso M: Maximizar la banda baja del resultado de la probabilidad con respecto a los parámetros del modelo. Esto corresponde a la búsqueda de la estimación de máxima verosimilitud con la posterior aproximada que se calcula en el paso E.
- El paso E y el paso M se ejecutan iterativamente hasta alcanzar un valor de máxima verosimilitud.

Mientras tanto, los parámetros de estimación calculada pueden utilizarse para inferir la distribución del tema de un nuevo documento mediante la realización de la inferencia variacional.

El método de la máxima verosimilitud estima θ_0 buscando el valor de θ que maximiza $\hat{\ell}\ (\theta\ /\text{x})$. Este es el llamado estimador de máxima verosimilitud (MLE) de θ_0:

$$\hat{\theta}_{mle} = \arg max\ \hat{\ell}\ (\theta / x_1\ ,\ldots,\ x_n) \tag{19}$$

5.10.2.2 Muestreo de Gibbs. Según Migon et al. (2008), se asume un vector de probabilidades aleatorias $\boldsymbol{\theta}$ y $\boldsymbol{p}(\boldsymbol{\theta})$ es la función de densidad conjunta y $\boldsymbol{p}(\boldsymbol{\theta_i}\ /\boldsymbol{\theta_{-1}}\ )$ es la distribución condicional de $\boldsymbol{\theta}$ dado los otros parámetros. Donde $\boldsymbol{\theta_{-j}}$ es:

$$\theta_{-j} = (\ \theta_1, \theta_2, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p\ ) \tag{20}$$

El muestreo de Gibbs es un caso particular del algoritmo de Metrópolis Hastings, donde la distribución propuesta $q\ (\theta^{'}/\ \theta\ )$ son las distribuciones condicionales $p(\theta_i\ /\theta_{-1}\ )$ del parámetro θ :

- Se establecen unos valores iniciales ($\theta^0 = \theta_1^0, \dots , \theta_p^0$).
- Para cada iteración se realiza:
  - Se genera θ_1^j de $p(\theta_1 / \theta_2^{j-1}, \dots , \theta_p^{j-1})$.
  - Se genera θ_2^j de $p(\theta_2 / \theta_1^{j-1}, \theta_3^{j-1} \dots , \theta_p^{j-1})$.
  - Se genera θ_3^j de $p(\theta_1 / \theta_2^{j-1}, \theta_4^{j-1} \dots , \theta_p^{j-1})$.
  - Por último, se genera θ_p^j de $p(\theta_1 / \theta_2^{j-1}, \theta_4^{j-1} \dots , \theta_{p-1}^{j-1})$.

Finalmente se encuentra la distribución a posteriori del parámetro θ por medio de las distribuciones condicionales.

5.11 Medidas de validación de clúster

Las medidas numéricas que se aplican para juzgar varios aspectos de la validez de clúster como distancias y similitud, se clasifican en los siguientes tres tipos:

- Medidas Internas: se utiliza para definir la medida en que las etiquetas de clúster coinciden con las etiquetas de clase suministradas externamente.
- Medias Externas: se utiliza para medir la bondad de una estructura de agrupación sin tener en cuenta la información externa.
- Medidas Relativas: se utiliza para comparar dos agrupaciones o agrupamientos diferentes. A menudo se usa un índice externo o interno para esta función

Las primaras son propias de la clasificación supervisada ya que se cuenta con información a priori que permite validar la información que el algoritmo desarrolla, la segunda se usa tanto en la clasificación supervisada como no supervisada; siendo de gran importancia en la clasificación no supervisada ya que no se cuenta con información externa de la clasificación y los datos se deben

comparar y validad con la misma información ofrecida por el modelo no supervisado.

5.11.1 Medidas internas.

5.11.1.1 Cohesión. Mide qué tan estrechamente relacionados están los objetos en un clúster. Como por ejemplo la suma interna de cuadrados (SSW, por sus siglas en inglés) (Guzmán, s.f).

5.11.1.2 Separación. Mide que los clústeres estén ampliamente separados entre ellos. Existen varios enfoques para medir esta distancia entre clúster: distancia entre el miembro más cercano, distancia entre los miembros más distantes o la distancia entre los centroides.

5.11.1.3 Índice de Dunn. Mide separación entre *clusters*. Es la relación entre la menor distancia entre las observaciones que no están en el mismo grupo, y la mayor distancia entre los grupos. Tiene un valor entre cero e infinito, y debe ser lo más alto posible. Un valor más alto indica un mejor rendimiento del algoritmo de *clustering* (SCHIATTI SISÓ , 2017).

5.11.1.4 Índice de Silhoutte. Este índice corresponde al promedio del valor *Silhouette* de cada observación y mide el grado de confianza en la asignación de clusters de una observación en particular. Los valores próximos a 1 gozarán de una mayor confianza en la agrupación realizada, por el contrario, los valores próximos a -1 significan que la agrupación no es fiable (SCHIATTI SISÓ, 2017).

5.11.1.5 Distancia Hellinger. Es usada para cuantificar la similaridad entre dos distribuciones de probabilidad la P y Q que puede definirse para un espacio muestral arbitrario (Ω, A).

La distancia de Hellinger se define en distribuciones discretas, siendo $P = (p_1, p_2, \dots, p_k)$ y $Q = (q_1, q_2, \dots, q_k)$ como (Coronoda Matutti et al.,2015):

$$\mathrm{H(P,Q)} = \frac{1}{\sqrt{2}} \sqrt{\sum_{\mathrm{i=1}}^{\mathrm{k}} (\sqrt{\mathrm{p_i}} - \sqrt{\mathrm{q_i}})^2} \tag{21}$$

El cual está relacionado directamente con la norma euclidiana de la diferencia de las raíces cuadradas de los vectores, es decir

$$\mathrm{H(P,Q)} = \frac{1}{\sqrt{2}} \left\| \sqrt{P} - \sqrt{Q} \right\|_2 \tag{22}$$

En distribuciones continuas se define como:

$$\mathrm{H(P,Q)} = \frac{1}{2} \int (\sqrt{dP} - \sqrt{dQ})^2 \tag{23}$$

Donde $\sqrt{dP}$, respectivamente, $\sqrt{dQ}$ denota la raíz cuadrada de las densidades.

Un valor de 1 significa que la distancia es máxima y, por lo tanto, las distribuciones son muy diferentes. Un valor de 0 significa que las distribuciones son muy similares (Rus, Niraula, and Banjade, 2013).

5.11.1.6 Divergencia Kullback-Leibler. La divergencia de Kullback-Leibler (también llamada entropía relativa) es una medida de cómo una distribución de probabilidad es diferente de una segunda, la distribución de probabilidad de referencia. La divergencia de Kullback-Leibler se define para distribuciones de probabilidad $\boldsymbol{P} = (\boldsymbol{p_1}, \boldsymbol{p_2}, \dots, \boldsymbol{p_k})$ y $\boldsymbol{Q} = (\boldsymbol{q_1}, \boldsymbol{q_2}, \dots, \boldsymbol{q_k})$ como (León Boyajian and Lamberti, 2011):

$$d_{KL} = \sum_{i=1}^{d} P_i \, ln \frac{P_i}{Q_i} \tag{24}$$

6. Proceso de descubrimiento de conocimiento

La metodología que es utilizada para el desarrollo de las diferentes etapas que dan cumplimiento a los objetivos planteados comprende: el establecimiento de herramientas ofimáticas necesarias para la generación de los modelos, la definición de las autoridades de salud a considerar en el análisis, la extracción de los tweets, la limpieza y representación de los tweets, el ajuste de modelos no supervisados, la selección del mejor modelo y el análisis de resultados.

6.1 Herramientas utilizadas

- Computador Portátil
  - Marca: Acer
  - Procesador: Intel Core i5
  - Memoria RAM: 4 GB
  - Sistema operativo: 64 bits
  - Velocidad del procesador: 2.30 GHz
  - Núcleos: 4
  - Procesadores Lógicos: 8
- Software estadístico R Studio
  - Versión: 3.5.0
  - Lenguaje: español
  - Fecha: 2018

6.2 Sistemas de información

Se realiza una revisión de las autoridades de salud en Colombia, a través de sus funciones, y principales campañas de promoción y prevención, con el fin de identificar las entidades con las a tener en cuenta en el presente proyecto.

6.2.1 Entidades nacionales de salud. La ley 100, expedida el 23 de diciembre de 1993 (Congreso de la república, 1993) reglamenta el sistema general de seguridad social de Colombia integral, compuesto por 4 libros entre los cuales se encuentra en el segundo libro, el Sistema General de Seguridad Social en Salud (SGSSS)

En el artículo 155 de la ley 100 de 1993 se definen los integrantes del sistema general de seguridad social en salud (Congreso de la república, 1993):

- Organismos de dirección, vigilancia y control: El Ministerio de Salud y de Trabajo, el consejo nacional de seguridad social en salud, la superintendencia nacional en salud; entre otros.
- Organismos de administración y financiación: Las entidades promotoras de salud, las direcciones seccionales, distritales y locales de salud y el fondo de solidaridad y garantía.
- Instituciones prestadoras de servicios de salud, públicas, mixtas o privadas.
- Demás entidades de salud que, al entrar en vigencia la presente ley, estén adscritas a los Ministerios de Salud y Trabajo.
- Empleadores, trabajadores y sus organizaciones y trabajadores independientes que cotizan al sistema contributivo y los pensionados.
- Beneficiarios del sistema general de seguridad social en salud en todas sus modalidades.
- Comités de participación comunitaria "Copacos" creados por la Ley 10 de 1990 y las organizaciones comunales que participen en los subsidios de salud.

De esta manera en el desarrollo del presente proyecto se seleccionan los organismos de dirección, vigilancia y control, y las demás entidades de salud que estén adscritas a los ministerios de salud y trabajo; para ello, se usa la lista de entidades nacionales y territoriales generada por el

departamento administrativo de la función pública y que reposa el portal de datos abiertos del estado colombiano, filtrada por las entidades de naturaleza jurídica de orden nacional pertenecientes al sector de salud y protección social. (Ministerio de Tecnologías de la Información y las Comunicaciones, 2018)

Además, se incluye a la organización panamericana de la salud y organización mundial de la salud, con presencia en Colombia, ya que trabaja en coordinación con el Ministerio de Salud y Protección Social, sus entidades adscritas, otras entidades del gobierno nacional, de los gobiernos territoriales, y organizaciones de la sociedad civil colombiana.

A continuación, una breve descripción de cada autoridad a tener en cuenta:

- Empresa Social del estado Centro Dermatológico Federico Lleras Acosta: tiene como misión brindar, con calidad humana y seguridad, servicios especializados en dermatología. Realizar formación, educación e investigación en las áreas de su competencia. Asesorar al gobierno nacional en la planeación y ejecución de estrategias para la promoción de la salud, la prevención y el control de las patologías cutáneas, en el marco de la responsabilidad social (Centro Dermatologico Federico Lleras Acosta, s.f.).
- Fondo pasivo Social de Ferrocarriles Nacionales de Colombia: es una entidad adaptada (EAS) que presta servicios de salud a los pensionados de los ferrocarriles nacionales de Colombia, puertos de Colombia y a sus respectivos beneficiarios.
- Fondo de prevención Social del congreso de la República: tiene por objetivo primordial el reconocimiento y pago de las pensiones y cesantías de los Congresistas, de los empleados del congreso y del mismo Fondo.
- Instituto Nacional de Cancerología (INC): es una institución del estado colombiano en su

orden nacional, que trabaja por el control integral del cáncer a través de la atención y el cuidado de pacientes, la investigación, la formación de talento humano y el desarrollo de acciones en salud pública (Instituto Nacional de Cancerología, 2018).

- Instituto Nacional de Salud (INS): es una entidad pública de carácter científico-técnico en salud pública, de cobertura nacional, que contribuye a la protección de la salud en Colombia mediante la gestión de conocimiento, el seguimiento al estado de la salud de la población y la provisión de bienes y servicios de interés en salud pública (Ospina, 2018).

  LA INS tiene como objeto: (i) el desarrollo y la gestión del conocimiento científico en salud y biomedicina para contribuir a mejorar las condiciones de salud de las personas; (ii) realizar investigación científica básica y aplicada en salud y biomedicina; (iii) la promoción de la investigación científica, la innovación y la formulación de estudios de acuerdo con las prioridades de salud pública de conocimiento del instituto; (iv) la vigilancia y seguridad sanitaria en los temas de su competencia; la producción de insumos biológicos; y (v) actuar como laboratorio nacional de referencia y coordinador de las redes especiales, en el marco del sistema general de seguridad social en salud y del sistema de ciencia, tecnología e innovación.

- Instituto Nacional de Vigilancia de Medicamentos y Alimentos (INVIMA): proteger y promover la salud de la población, mediante la gestión del riesgo asociada al consumo y uso de alimentos, medicamentos, dispositivos médicos y otros productos objeto de vigilancia sanitaria.

- Ministerio de Salud y Protección Social (MINSALUD) : tiene como misión dirigir el sistema de salud y protección social en salud, a través de políticas de promoción de la salud, la prevención, el tratamiento y la rehabilitación de la enfermedad y el aseguramiento, así como

la coordinación intersectorial para el desarrollo de políticas sobre los determinantes en salud, bajo los principios de eficiencia, universalidad, solidaridad, equidad, sostenibilidad y calidad, con el fin de contribuir al mejoramiento de la salud de los habitantes de Colombia.

- Sanatorio de Agua de Dios, Empresa Social del Estado: el Sanatorio de Agua de Dios Empresa Social del Estado, de orden nacional, es una institución prestadora de servicios de salud de baja complejidad a pacientes Hansen y demás población, la cual ejecuta actividades de docencia, investigación y capacitación en enfermedades de salud pública, con un recurso humano que brinda seguridad, humanización, calidad y calidez en el proceso de atención al paciente y su familia.
- Sanatorio de contratación, Empresa social del Estado: especializada en el manejo integral de los pacientes de Hansen, está orientada a la prestación de servicios de salud con calidad técnico–científica
- Superintendencia Nacional de Salud (SUPERSALUD): proteger los derechos de los usuarios del Sistema General de Seguridad Social en Salud mediante la inspección, vigilancia, control y el ejercicio de la función jurisdiccional y de conciliación, de manera transparente y oportuna.
- Organización Panamericana de Salud / Organización Mundial de la salud – Colombia (OPS/OMS COLOMBIA): trabaja en coordinación con el Ministerio de Salud y Protección Social, sus entidades adscritas, otras entidades del gobierno nacional, de los gobiernos territoriales, y organizaciones de la sociedad civil colombiana, brindando cooperación técnica en protección social y salud, fortaleciendo la articulación intersectorial, para el mejoramiento de las condiciones de salud y bienestar de la población, y el logro de los objetivos de desarrollo del milenio, en especial de las poblaciones vulnerables, con un

enfoque de equidad, género, diversidad etnocultural y derechos humanos. alcanzar una vida saludable, mediante la promoción de estrategias orientadas a la reducción y manejo de riesgos, a la promoción de factores protectores de la salud, y del acceso con calidad y equidad a los servicios de salud son los grandes retos en los que la OPS/OMS basa su compromiso por la salud individual y colectiva.

6.2.2 Twitter. De acuerdo con la tabla 4, en la cual se describen las páginas web y la cuenta de Twitter de las autoridades anteriormente descritas y que para el presente proyecto se obtiene la información a través de las autoridades que tengan cuenta de Twitter, como lo son:

- Empresa Social del estado Centro Dermatológico Federico Lleras Acosta
- Instituto Nacional De Cancerología
- Instituto Nacional De Salud
- Instituto Nacional De Vigilancia De Medicamentos Y Alimentos
- Ministerio De Salud Y Protección Social
- Sanatorio de contratación, Empresa social del Estado
- Superintendencia Nacional De Salud
- Organización Panamericana De La Salud/Organización Mundial De La Salud – Colombia

No se incluye dentro de las entidades seleccionadas al Sanatorio de contratación, Empresa social del Estado, dado que esta entidad solo ha publicado un tweet el día 29 de enero del 2018, lo cual no es relevante para considerarlo dentro de los análisis del presente proyecto.

Tabla 4
Lista de Autoridades de Salud en Colombia

Autoridad de salud	Sigla	Página web	Twitter
Empresa Social del estado Centro Dermatológico Federico Lleras Acosta	----------	www.dermatologia.gov.co	@dermatologico
Fondo pasivo Social de Ferrocarriles Nacionales de Colombia	----------	www.fps.gov.co	NO TIENE
Fondo de prevención Social del congreso de la Republica	----------	www.fonprecon.gov.co	NO TIENE
Instituto Nacional De Cancerología	INC	www.cancer.gov.co	@INCancerologia
Instituto Nacional De Salud	INS	www.ins.gov.co	@INSColombia
Instituto Nacional De Vigilancia De Medicamentos Y Alimentos	INVIMA	www.invima.gov.co	@invimacolombia
Ministerio De Salud Y Protección Social	MINSALUD	www.minsalud.gov.co	@MinSaludCol
Sanatorio de Agua de Dios, Empresa Social del Estado	----------	www.sanatorioaguadedios.gov.co	NO TIENE
Sanatorio de contratación, Empresa social del Estado	----------	www.sanatoriocontratacion.gov.co	@SanatorioESE
Superintendencia Nacional De Salud	SUPERSALUD	www.supersalud.gov.co	@Supersalud
Organización Panamericana De La Salud/Organización Mundial De La Salud – Colombia	OPS/OMS COLOMBIA	www.paho.org/col	@OPSOMS_Col

6.2.3 Promoción de la Salud y prevención de Enfermedades. La promoción de la salud y prevención de enfermedades en las diferentes entidades nacionales de salud en Colombia, están enmarcadas en proyectos, direcciones o campañas diseñadas. A continuación, se describe para

cada entidad nacional de salud su forma de participación en la promoción de la salud y prevención de enfermedades.

6.2.3.1 Centro Dermatológico Federico Lleras Acosta. El centro dermatológico maneja diferentes campañas sujetas a dos proyectos:

- Sobre Cáncer de Piel
- Autoexamen de Piel

6.2.3.2 Instituto nacional de cancerología. El instituto maneja dos programas de salud pública relacionados con:

- Vigilancia de la leucemia aufuadas pediátricas
- Recomendaciones sobre la aplicación de la vacuna contra el virus del papiloma humano

6.2.3.3 Instituto Nacional De Salud. El instituto maneja 5 direcciones entorno a la promoción de salud y prevención de enfermedades

- Redes de salud publica
- Vigilancia y análisis del riesgo en salud publica
- Investigación de salud pública
- Observatorio nacional de salud

6.2.3.4 Instituto nacional de vigilancia de medicamentos y alimentos (INVIMA). El INVIMA maneja cuatro direcciones relacionadas con el control y vigilancia de la salud pública:

- Dirección de Alimentos y Bebidas
- Dirección de Medicamentos y Productos Biológicos
- Dirección de Dispositivos Médicos y Otras Tecnologías

- Dirección de Cosméticos, Aseo, Plaguicidas y Productos de Higiene Doméstica

6.2.3.5 Ministerio de Salud y Protección Social. El ministerio basa sus estrategias en salud pública en 10 programas.

- Estilos saludables: relacionados con temáticas como: (1) actividad física (2) Nutrición y alimentación saludable (3) prevención y consumo del tabaco (4) peso saludable (5) Lavado de manos (6) salud bucal, visual y aditiva
- Poblaciones vulnerables: relacionada con temáticas como (1) enfoque de curso de vida (2) lactancia materna, muerte subirá, vacunación y salud bucal; en primera infancia
- Enfermedades trasmisibles: son las enfermedades como (1) zika, dengue y chikunguña (2) Infección respiratoria Aguda –IRA (3) tuberculosis (4) enfermedades infecciones desatendidas, oncocercosis, tracoma, gohelmintiasis, enfermedad de chagas, tétanos neonatal, sífilis congénita, malaria, rabia trasmitida por perros, lepra.
- Enfermedades no trasmisibles (ENT): los cuatro tipos principales son: (1) las enfermedades cardiovasculares, como los infartos de miocardio, el ataque cerebrovascular (acv), la falla cardíaca, la hipertensión arterial, entre otras. (2) los diferentes tipos de cáncer. (3) las enfermedades respiratorias crónicas, como la neumopatía obstructiva crónica o el asma. (4) la diabetes
- Salud sexual y reproductiva: relacionados con temáticas como: (1) la sexualidad y sus derechos (2) violencia de Genero (3) salud materna (4) anticoncepción (5) Cánceres relacionados con sexualidad y reproducción; Cáncer de cuello uterino, cáncer de mama, cáncer de próstata (6) infecciones de transmisión sexual – ITS; virus de la inmunodeficiencia humana (VIH) y sindrome de la inmunodeficiencia adquirida (SIDA) (6) prevención del aborto inseguro e interrupción voluntaria del embarazo (IVE) (7) preventivo de embarazo en

jóvenes, derechos sexuales y reproductivos de los jóvenes y adolescentes.

- Epidemiologia y demografía:
- Salud ambiental: relacionada con temáticas como (1) agua y saneamiento básico (2) inspección vigilancia y control sanitario (3) sustancias y productos químicos (4) entornos saludables (5) aire y salud (6) minería y salud (7) vecindad y fronteras (8) zoonosis (9) cambio climático
- Salud mental: relacionada con temáticas como (1) sustancias psicoactivas (2) depresión y prevención del suicidio
- Vacunación: relacionada con temáticas como (1) esquemas de vacunación (2) puntos de vacunación (3) vacunación del viajero (4) vacuna contra el virus del papiloma humano
- Salud nutricional: relacionada con temas como: (1) alimentación y nutrición: lactancia materna (2) control deficiencia de micronutrientes (3) alimentación saludable (4) inocuidad y calidad de alimentos (5) atención integral a la desnutrición aguda (6) política de seguridad alimentaria y nutricional.

6.2.3.6 Superintendencia Nacional De Salud. Por medio de la dirección de atención del usuario la Supersalud desarrollo dos programas enfocados en la salud, como lo son:

- Dirección de atención al usuario: se encarga de ejercer la inspección y vigilancia sobre el cumplimiento de los derechos en salud y la debida protección al usuario, verificando que los vigilados suministren información útil, suficiente, veraz y oportuna que les permitan ejercer eficazmente sus derechos. De igual manera, la dirección de atención al usuario se encarga de responder las peticiones, quejas y reclamos que formulan ante la superintendencia, los usuarios del sistema general de seguridad social en salud; para atender las reclamaciones que involucren una urgencia vital o la integridad del usuario la dirección de atención al usuario

cuenta a su cargo con el grupo de soluciones inmediatas en salud (Supersalud ,2017).

- Dirección de participación ciudadana: se encarga de promover entre los usuarios del sistema general de seguridad social en salud la conformación de espacios para la promoción de mecanismos de control social en el sector, como (Supersalud ,2017):
    - Dialoguemos
    - Seminarios
    - Supersalud en las regiones
    - Socialización de Estrategias de Participación Ciudadana
    - Videoconferencias
    - Participación en población infantil de 8 a 12 años de edad

6.2.3.7 Organización Panamericana de la Salud/Organización Mundial De La Salud – Colombia. La OMS/OPS Colombia, participa en la promoción de la salud por medio de proyectos de cooperación como lo son

- Atención Integral a la Mujer y al Recién Nacido
- Desarrollo de Sistemas y Servicios de Salud
- Evidencia en Salud y Control de Enfermedades (Cólera, dengue, Lepra, Malaria, Tuberculosis, tracoma, Geohelmintiasis, Oncocercosis)
- Salud Ambiental y Entornos Saludables
- Salud en Emergencias y Desastres

6.3 Extracción de datos de Twitter

La extracción de los datos se hace utilizando el API estándar de Twitter y generando la conexión con R a través de la función *get timeline* del paquete *Retweet* desarrollado por Kearney (2018).

La selección de los datos de Twitter que soportan el tema de investigación, se realiza de acuerdo con los tweets publicados por las autoridades seleccionadas (Tabla 4) en el periodo comprendido entre el 01/01/2018 desde las 00:00:00 y el 30/04/2018 hasta las 00:00:00; periodo que se elige teniendo en cuenta los tiempos para el desarrollo del proyecto y con el fin de obtener información suficiente para incluir en el proceso de análisis.

La base de datos por cada usuario está compuesta por 42 variables que representan la información de cada tweet extraído en formato de codificación UTF-8 y se muestran disgregados en la Tabla 5 como variables se encuentran: el texto del tweet, la fecha de publicación, el ID del usuario que lo publica, las URLs usadas, entre otros. En la Tabla 5 y se destaca que la mayor cantidad de información se obtiene del INVIMA, de MINSALUD y de Supersalud.

Tabla 5
Número de tweets extraídos y nombre el archivo.csv por cada Entidad

Entidad	Número de Tweets extraídos	Nombre de archivo.csv
Dermatológico	236	dermatológico.csv
INCancerologia	196	INCancerologia.csv
INSColombia	787	INSColombia.cvs
Invimacolombia	1943	Invimacolombia.csv
MinSaludCol	1925	MinSaludCol.csv
Supersalud	1083	Supersalud.csv
OPSOMS_Col	165	OPSOMS_Col.csv
TOTAL	6.335	Total.csv

Nota: Adaptado de Rstudio versión 3.5.0 (2018)

6.4 Preprocesamiento de la información

En la presente etapa se lleva a cabo la limpieza y la transformación de los datos, lo cual representa un paso importante para la generación de los modelos de minería de datos.

6.4.1 Limpieza de los datos. La entrada de esta etapa son los archivos .csv de cada autoridad (x) obtenidos en la etapa de captura de información. Las operaciones que se realizan se describen en la Figura 5 utilizando el paquete *text mining (tm)* desarrollado por Feinerer, Hornik, & Meyer

(2008) y *base* desarrollado por R Core Team (2018), y como resultado de este proceso de limpieza se obtiene el archivo de datos el cual se encuentra en el mismo formato csv.

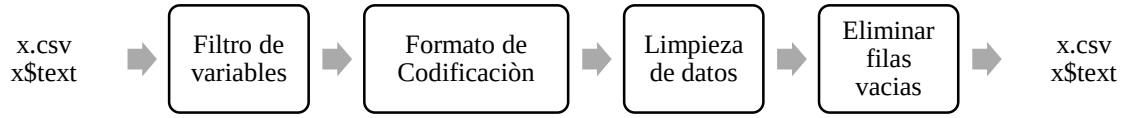


Figura 5: Proceso de Limpieza de Datos

En esta etapa se realizan las siguientes actividades:

- Filtro de variables: en cada archivo x.csv, se filtran 6 de las 42 variables que contiene cada archivo, seleccionando las variables más importantes (Ver Tabla 6) para proceder con el proceso de limpieza y transformación de esta información capturada:

Tabla 6
Variables seleccionadas

Variable	Definición
created_at	Devuelve le fecha y hora en que fue creado el tweet
screen_name	Devuelve el nombre del usuario que realizo el tweet
Text	Devuelve el contenido del tweet
favorite_count	Devuelve la fuente de donde se realizó el tweet
retweet_count	Devuelve el número de veces que se retwiteo el tweet
Hashtags	Devuelve los *hashtags* del Tweet
is_retweet	Devuelve TRUE, si el Tweet es un Retweet y FALSE en caso contrario

Nota: Adaptado de Rstudio versión 3.5.0 (2018)

- Formato de codificación: se cambia el formato de codificación de UTF-8 a un formato de codificación Latin1.

- Limpieza de datos: utilizando el paquete *tm y base*, se elimina información no relevante como:

  - Extender abreviaturas.
  - Transformar mayúsculas a minúsculas.

- Eliminación de URLs: la mayoría de tweets contienen URLs de redireccionamiento a enlaces con mayor cantidad de información sobre lo que se habla en el tweet.
- Eliminación de correos institucionales: los tweets incluyen direcciones de correo electrónico como medio de contacto con los usuarios.
- Eliminación de menciones de usuarios: algo común en Twitter es mencionar a algún usuario dentro del contenido del tweet.
- Eliminación de imágenes, caracteres y emoticones: algo común es incluir en sus tweets imágenes, emoticones o caracteres, que permitan una mejor visualización del tweet o llamar más su atención.
- Eliminación de símbolos de puntuación: son símbolos que generan ruido y no aportan sentido al análisis del contenido del tweet.
- Eliminación de numeraciones: la información de fechas u horas no es relevante para el análisis de este proyecto.
- Eliminación de palabras vacías (*Stop-Words*): las palabras vacías son comúnmente conjunciones, preposiciones o artículos que hacen referencia a aquellos términos que por un lado carecen de significado (su contenido semántico es escaso) y por otro suelen ser muy frecuentes, por lo que su aporte al contenido del texto es casi nulo. Por ejemplo, se eliminan palabras como: el, la, los, nosotros, con, conque, si no, mientras, entre otras.
- Eliminar filas vacías: Después de realizar a limpieza, se procede a eliminar los tweets que no tienen información, es decir aquellos tweets en donde sus caracteres o palabras eran emoticones o URLs o palabras vacías

Al finalizar el proceso de limpieza el repositorio de datos objeto de estudio se conforma por 5830 tweets disgregados como se muestra en la Tabla 7.

Tabla 7
Número de tweets usados en el análisis

Nombre del usuario	Número de tweets
dermatológicoFe	235
INCancerologia	196
INSColombia	777
invimacolombia	1468
MinSaludCol	1913
Supersalud	1083
OPSOMS_Col	158
TOTAL	5.830

Nota: Adaptado de Rstudio versión 3.5.0 (2018)

6.4.2 Representación. La información de entrada para esta etapa es la variable texto obtenida después del proceso de limpieza. Las operaciones que se realizan se muestran en la Figura 6, en donde la salida es la información transformada en el espacio vectorial adicionalmente en esta etapa se muestran los términos o palabras con mayor frecuencia tanto para cada autoridad como para el total de los tweets.

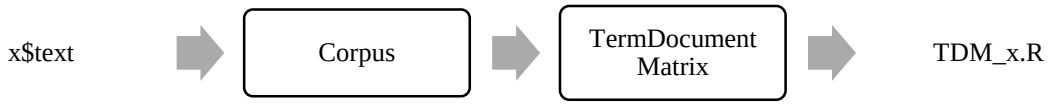


Figura 6: Proceso de Representación

En esta etapa de representación se realizan las siguientes actividades:

- Corpus: la variable texto de los tweets es procesada por la función "VectorSource" del paquete *tm*, la cual permite que cada tweet sea interpretado como un documento y sea representado en el espacio vectorial en donde cada palabra es separada como una variable, para posteriormente conformar un corpus o colección de documentos, en este caso colección de tweets.
- Matriz término documento: cada corpus es procesado con la función "TermDocumentMatrix" del paquete *tm*, la cual convierte cada archivo corpus a un formato

de matriz término documento (Tabla 8), Esta transformación es requerida como información de entrada de los modelos de minería de datos y de aprendizaje automático.

A partir de la matriz término documento que cuenta con un peso de término frecuencia (número de veces que aparece el termino en el documento) se procede a calcular el *Tf-Idf*, de tal forma que si un término aparece más veces en un documento tiene más importancia y menos importancia si se presenta lo contrario (aparecer menos veces en un documento).

Tabla 8
Nombre archivos de tipo TermDocuemntMatrix

Username	Nombre de archivo. R
dermatológicoFe	TDM_dermatológicoFe.R
INCancerologia	TDM_INCancerologia.R
INSColombia	TDM_INSColombia.R
Invimacolombia	TDM_invimacolombia. R
MinSaludCol	TDM_MinSaludCol.R
Supersalud	TDM_Supersalud.R
OPSOMS_Col	TDM_OPSOMS_Col.R
TOTAL	TDM_Total.R

Nota: Adaptado de Rstudio versión 3.5.0 (2018)

En la figura 7 se observa para el total de documentos (5830 tweets) que los términos más frecuentes son “salud” e “información”. En el Apéndice A, se observa para cada entidad la nube de palabras que representa el *Tf-Idf* de las palabras de mayor ocurrencia en los tweets que la conforman; para el caso de cada autoridad se indican las 4 palabras de mayor importancia:

- Centro dermatológico Federico lleras: piel, cáncer, esencial y protegemos
- Instituto Nacional de Cancerología: programa, cáncer, asbesto y hoy
- Instituto Nacional de salud: salud, Colombia, INSs y director
- Invima: esencial, protegemos, información, vida y alertas
- Ministerios de Salud: Salud, evite, enfermedades e informativo
- Organización mundial de la salud: salud, interesa, “mensaje y conversatorio

- Superintendencia de salud: Salud, datos, favor y numero

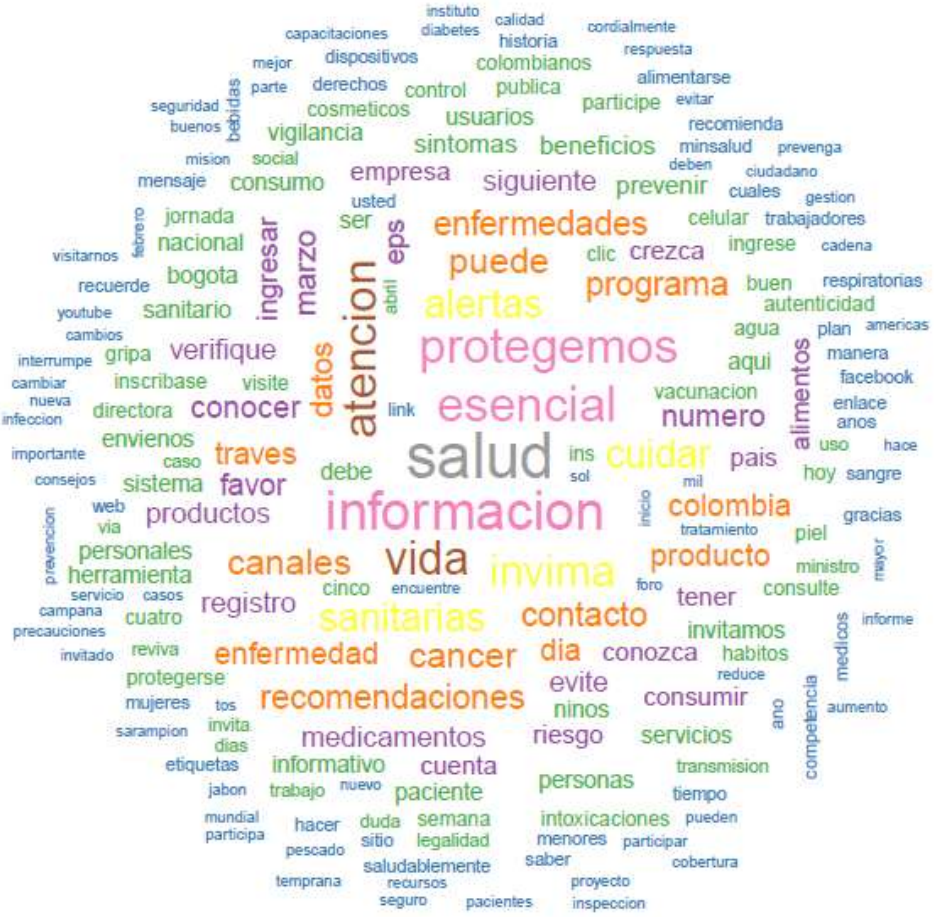


Figura 7: Nube de palabra del Total Tweets. Adaptado de Rstudio versión 3.5.0 (2018)

6.5 Minería de texto. De acuerdo a la revisión de literatura en cuanto a los métodos utilizados para el modelamiento de tópicos y teniendo en cuenta los objetivos de este estudio, los algoritmos usados son *LDA* y *CTM*, a los cuales se les determina el K óptimo (número de tópicos) en cada uno de los modelos y el mejor modelo es aquel que presenta las mayores distancias de *Hellinger* entre tópicos. Para el ajuste de los dos algoritmos se sigue el proceso general de la Figura 8 y este se describe en la Tabla 9:

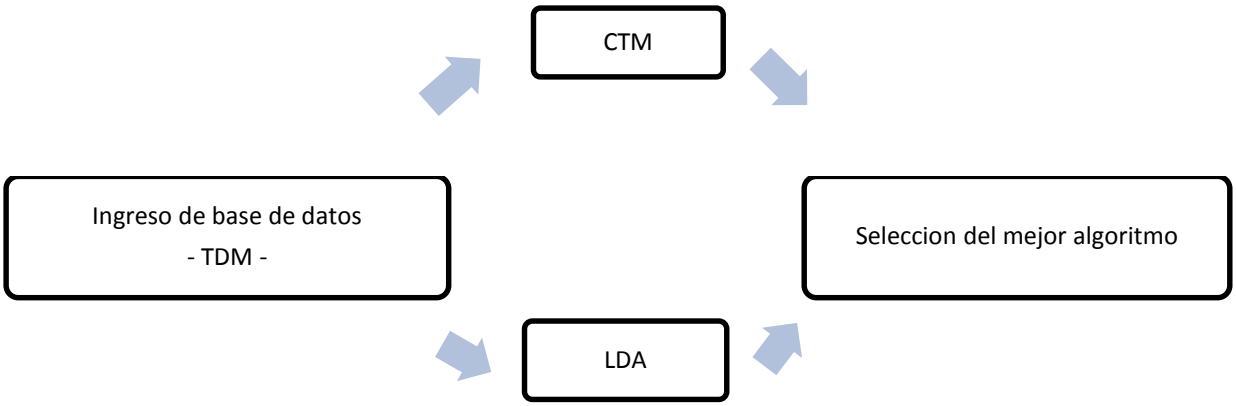


Figura 8: Proceso de minería de Texto

Tabla 9
Procedimiento de ajuste de los algoritmos

Paso	Nombre	Descripción
Paso 1	Ingresar Base de datos	Se usa como base de datos una Matriz termino-Documento de 5830 documentos y 6599 términos obtenida en la etapa de preprocesamiento de los datos. Es necesaria una transformación a matriz de documento – término, dado que la función LDA y CTM del paquete *topicsmpdels solo* permite como dato de entrada esta matriz.
Paso 2	Determinar K para cada algoritmo	Se seleccionan dos métodos de muestreo Gibbs y VEM, con el objetivo de verificar cuál de los dos métodos se ajusta mejor al modelo y obtiene el mejor K (tópicos). En el caso de CTM, este no define el método de GIBBS como muestro, solo se usa VEM.
Paso 3	Seleccionar K optimo	El número de temas óptimos se selecciona de acuerdo con la medida de máxima verosimilitud. Se selecciona el método que obtuvo la medida de máxima verosimilitud más alta para K y se escoge el K que se ajusta a cada uno de los tres modelos resultantes.
Paso 4	Selección mejor algoritmo	1. Se ejecutan nuevamente los tres modelos con los K óptimos para cada uno. 2. Se crean las matrices de distancias *Hellinger* entre tópicos de cada modelo

Se ejecutan tres modelos LDA con VEM, LDA con GIBBS y CTM con VEM; la selección del número óptimo de temas para cada modelo se muestra en la tabla 10 y en la Figura 9, se muestra la variación de las medidas de máxima verosimilitud (*loglikelihood*) por cada modelo respecto al número de tópicos. Así, para cada modelo los tópicos que mejor se ajustan a los parámetros

definidos, son 66 temas para el modelo de LDA con VEM, 45 temas para el modelo LDA con GIBBS, y finalmente para el modelo CTM con VEM, 69 temas.

Tabla 10
Número óptimo de K para cada modelo

	Loglikelihood	Tiempo de procesamiento	Tópicos
vem_LDA	-338.755.650.914.659	1.54 Horas	66
gibbs_LDA	-336.594.903.015.539	30.34 Min	45
vem_CTM	-338.113.273.452.270	3.85 Días	69

Nota: Adaptado de Rstudio versión 3.5.0 (2018)

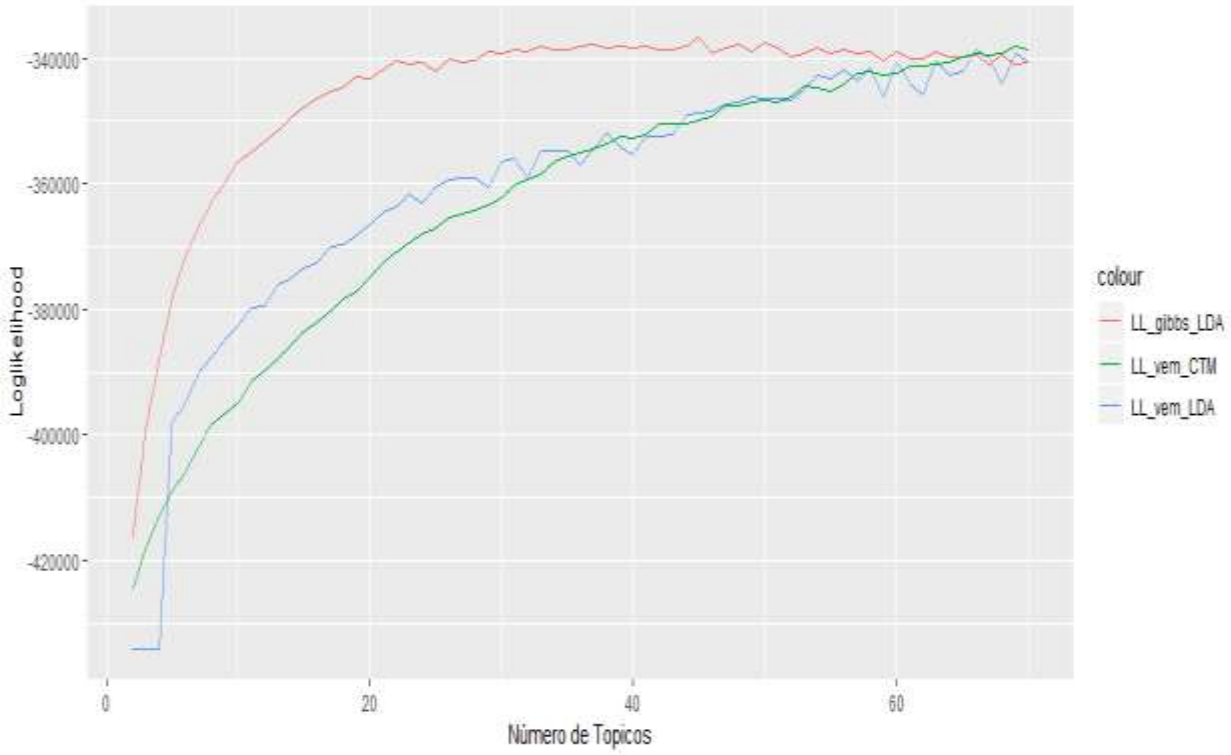
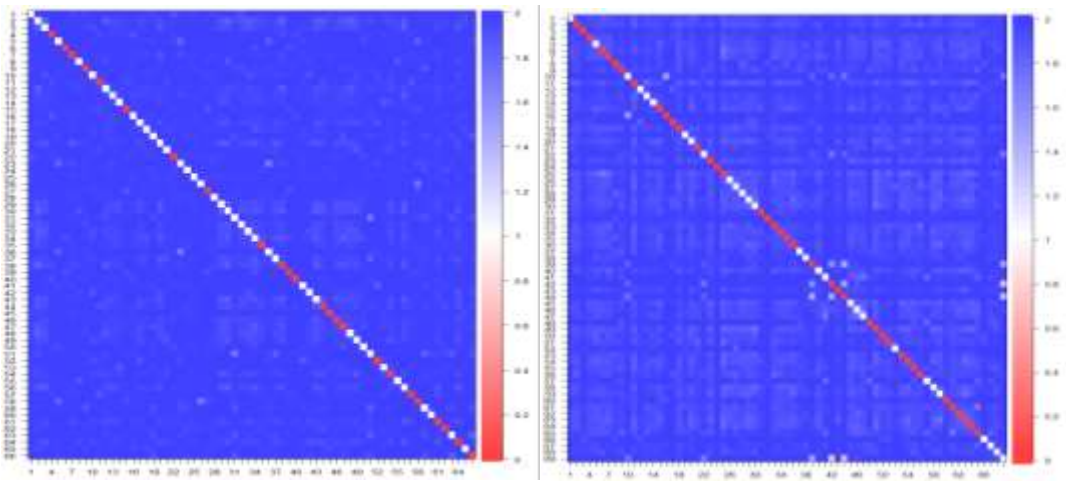
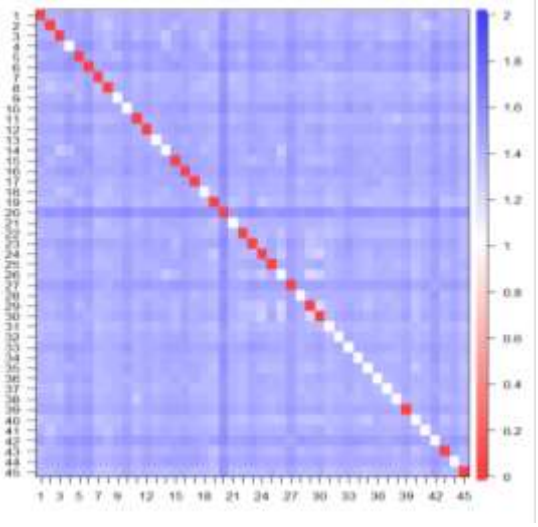


Figura 9: *LogLikelihood* de los Modelos. Adaptado de Rstudio versión 3.5.0 (2018)

Posteriormente en Figura 10, se muestran las matrices de distancias *Hellinger* de los tres modelos; el modelo LDA con el muestreo VEM es el que presenta mayor proporción de tópicos caracterizados por altas distancias, siendo de esta forma el modelo que presenta más diferencia entre los tópicos, en comparación con los otros dos modelos, de los cuales en el modelo LDA con GIIBS no se presenta diferencia significativa entre los tópicos



(a) (b)



(c)

Figura 10: Distancias Hrllinger . (a) LDA con VEM (b) CTM con VEM (c) LDA con GIIBS

7. Análisis de resultados

Considerando que los mejores resultados son obtenidos del modelo LDA con el muestreo de VEM, de este modelo se tienen dos tablas (matrices): la primera tabla de 6.599 términos por 66 tópicos, describe las probabilidades a-posteriori de seleccionar un término cuando se muestrea un tema en particular ($\beta_{1:k,1:V}$), la segunda tabla de 66 tópicos por 5.830 documentos, describe la probabilidad a-posteriori de seleccionar un tema en particular al muestrear un documento ($\theta_{1:d,1:k}$). A

continuación, se realiza el análisis de los tópicos obtenidos y de los documentos (tweets) que corresponden a los tópicos más representativos.

7.1 Asignación de Tópicos

En la Tabla 11, se muestran los nueve términos con más probabilidad de ser asignados al tópico, estas probabilidades se encuentran entre el 0.5 % y 0.2%

Tabla 11
Términos con mayor probabilidad según Tópico

	Tópico 1	Tópico 2	Tópico 3	Tópico 4	Tópico 5	Tópico 6	Tópico 7
1	ins	Servicios	Mayor	Puede	Registro	Consumo	casos
2	directora	Salud	Cuidado	Anos	Consumir	Reducir	sarampión
3	donación	Tecnologías	Cobertura	ninos	Sanitario	Cinco	caso
4	menciona	Ser	Salud	menores	Través	Realizar	factores
5	organos	Toda	Departamentos	historia	Celular	Minutos	pais
6	pais	Medicamentos	Cooperación	adolescentes	Verifique	Minimo	segundo
7	trasplantes	Usted	Tres	ninas	Producto	Riesgo	santa
8	mortalidad	Consultar	Nuevos	tiempo	Autenticidad	Peso	principales
9	red	Públicos	Urgencias	cambiar	Evite	Ejercicio	edad
	Tópico 8	Tópico 9	Tópico 10	Tópico 11	Tópico 12	Tópico 13	Tópico 14
1	marzo	Sol	Atención	enfermedad	pacientes	Tener	cancer
2	bogota	Mujeres	Sanitarias	sintomas	febrero	Población	internacional
3	inscribase	Importante	Alertas	tratamiento	exclusiones	habitos	cali
4	participe	Ospina	Canales	enfermedades	medicamentos	cuatro	lucha
5	invitado	Informe	Personas	explica	participar	nacional	salud
6	publica	Martha	Evite	Jose	aqui	cauca	ministro
7	cordialmente	Resultados	Gripa	transmisibles	sistema	gobierno	ciudades
8	habla	Protegerse	Contacto	mojica	construccion	salud	paises
9	audiencia	Recomienda	Tos	alejandro	recursos	Niños	ciudad
	Tópico 15	Tópico 16	Tópico 17	Tópico 18	Tópico 19	Tópico 20	Tópico 21
1	esencial	Hoy	Dia	deben	web	salud	cuenta
2	protegemos	Niños	calidad	Enero	sitio	encuentro	programa
3	cuidar	Apoyo	mundial	cuales	invima	lideres	parte
4	vida	Investigación	prevenir	adoptar	nuevo	bogota	enfermedades
5	empresa	Participa	salud	entidades	quejas	edad	Director
6	crezca	Zika	cuales	siguientes	Polvora	medellin	Hace
7	enfermedad	Comercio	servicio	especialista	Reclamos	control	Entidad
8	renal	Dudas	demuestra	medidas	Especial	compromiso	Prevención
9	diferentes	Huérfanas	conmemoracion	indica	lesionados	Evento	Mejor
	Tópico 22	Tópico 23	Tópico 24	Tópico 25	Tópico 26	Tópico 27	Tópico 28
1	hacer	Debe	invitamos	departamento	informacion	vida	Están
2	ins	Producto	enlace	equipo	Reportar	mision	Momento
3	lugar	Tener	siguiente	participa	Obtener	labor	Campo
4	debes	Revisar	tema	hora	Pasos	proteger	País
5	traves	Nombre	esperamos	inicio	Solicitudes	apoya	Destaca
6	punto	País	manana	autoridades	Hurto	cumplir	Importancia
7	consultar	Numero	seguir	reunion	Sugerencias	dejarlos	Americas
8	aumenta	Registro	recuerde	empresas	Único	respeta	Epidemiologia
9	votacion	Caso	campana	Siga	Generan	Compartir	Cartagena

Continuación tabla 11:

	Tópico 29	Tópico 30	Tópico 31	Tópico 32	Tópico 33	Tópico 34	Tópico 35
1	salud	Salud	medicos	directora	Conozca	eps	Proyecto
2	ministro	Publica	dispositivos	salud	Hospital	salud	Ley
3							
4	foro	Aportes	invima	curable	Atención	riesgo	Marco
5	visita	Juntos	medicamentos	cuello	Nuevas	acciones	Asbesto
6	consejos	Pension	aquí	uterino	Cuidados	solo	Paz
7	intervencion	Legal	conozcas	farmaceutica	Servicios	experiencia	Ana
8	andres	trabajemos	cosmeticos	debemos	trabajando	sistema	Niño
9	futuro	Justo	alimentos	politica	putumayo	Personas	Fortaleciendo
	Tópico 36	Tópico 37	Tópico 38	Tópico 39	Tópico 40	Tópico 41	Tópico 42
1	revise	vigilancia	pais	riesgo	trabajo	dias	Jornada
2	fecha	Control	salud	tipo	evitar	mensaje	Atención
3	vencimiento	inspeccion	recursos	fisica	tipos	gracias	Ciudadano
4	fabricante	Salud	victimas	actividad	infantil	buenos	Centro
5	presente	secretaria	soat	vida	temprana	favor	Usuario
6	lote	productos	sistema	diabetes	deteccion	privado	Bogota
7	registro	Siempre	hospitales	reduce	exposicion	envienos	Usuarios
8	empaque	Hace	millones	saludable	varios	queja	Continua
9	Numero	buscando	clinicas	alimentacion	alternativas	Través	Viernes
	Tópico 43	Tópico 44	Tópico 45	Tópico 46	Tópico 47	Tópico 48	Tópico 49
1	salud	conocer	mundo	luis	consulte	consulte	Vacunación
2	social	plan	presentar	correa	eps	eps	Semana
3	ano	nueva	frente	fernando	derechos	derechos	Americas
4	sector	beneficios	cada	vicesalud	salud	salud	Salud
5	proteccion	salud	ser	serna	usuarios	usuarios	Vacunas
6	seguridad	app	plazo	salud	afiliados	afiliados	Enfermedades
7	ministerio	respiratoria	situacion	vacuna	sanciono	sanciono	Atención
8	plan	infeccion	milagrosas	viceministro	ser	ser	Modelo
9	Generar	ira	constante	virus	seguimos	Seguimos	Integral
	Tópico 50	Tópico 51	Tópico 52	Tópico 53	Tópico 54	Tópico 55	Tópico 56
1	eps	ejemplo	datos	sangre	enfermedades	marzo	Colombia
2	ciudadanos	arrugas	numero	donantes	familiar	trabajadores	Nacional
3	informativo	colageno	contacto	cancer	transmision	independientes	Salud
4	traves	contiene	favor	aire	agua	pago	Instituto
5	preguntas	fresas	paciente	ayudar	medicina	partir	Saber
6	logros	antioxidantes	personales	podemos	cadena	vigencia	Pierdas
7	puedes	frescas	buen	esperanza	interrumpe	cambios	Expertos
8	contamos	citricos	dia	paciente	jabon	owlyyjinovf	Sistema
9	Internet	encontrar	envienos	colombiana	cumbre	Mayor	Cuentan
	Tópico 57	Tópico 58	Tópico 59	Tópico 60	Tópico 61	Tópico 62	Tópico 63
1	minsalud	informacion	siguiente	recomendaciones	aumento	ingresar	Productos
2	niveles	ingrese	link	uso	narino	eventos	Competencia
3	encuentran	legalidad	verificar	cosmeticos	efectos	reporte	Invima
4	mil	duda	visite	manual	unidad	gracias	Aquí
5	azucar	producto	etiquetas	encontrara	productos	contactarnos	Denuncia
6	enfermedad	verifique	recomienda	protectores	abril	invitamos	Corrupción
7	sangre	hecho	invima	solares	colombia	adversos	Contrabando
8	diabetes	conoces	nombre	correcto	tabaco	ventanas	Ilegalidad
9	Aprenda	vida	informe	visite	funcional	Actualización	Detene

	Tópico 64	Tópico 65	Tópico 66
1	atencion	respiratorias	invima
2	manera	primer	piel
3	profesionales	via	colombianos
4	Regional	protegerse	herramienta
5	implementacion	precauciones	usuarios
6	Colombia	infecciones	cuida
7	Será	ser	invita
8	Virtual	epidemiologico	encuentre

Nota: Adaptado de Rstudio versión 3.5.0 (2018)

En la Tabla 12 se asigna un nombre a cada uno de los tópicos de acuerdo a su tema o temas más predominantes, teniendo en cuenta los términos y documentos que se clasifican en cada tópico.

Algunos tópicos se denominan como no informativo, porque la información que contienen no es relevante en cuanto a campañas de promoción de la salud y prevención de enfermedades, ya que son tweets referentes a respuestas de preguntas, quejas, reclamos, entre otros; como es el caso de los tweets: *"@valen20011 por favor envíe los datos de la paciente vía mensaje directo con copia a @Supersalud", "@eleones89 Cordial saludo, por favor envíenos sus datos por DM para remitirlos al área correspondiente e iniciar el trámite*" , *"@jennicitac2 Buen día por favor envíenos los datos personales de la paciente y un número de contacto.*". Similarmente sucede para los tópicos 41, y 52 con los tweets: *"@hiperconectado @SaludAntioquia Buenas tardes nos puede colaborar con un número de contacto de algún familiar de la menor por DM", "@Retornovital06 @INCancerologia Déjenos por favor los datos personales del paciente y un número de contacto"*.

Por tanto, los tópicos *no informativos* representan el 19,7% del total de tópicos, teniendo en cuenta que solo el 71,3% de los tópicos proporcionan información sobre campañas y/o estrategias de promoción de salud y prevención de enfermedades en Colombia durante el primer cuatrimestre del año 2018.

Tabla 12
Etiquetas de cada tópico

Tema	Nombre	Tema	Nombre
Tópico 2	Financiación de medicamentos y procedimientos con recursos públicos	Tópico 35	Colombia sin asbesto
Tópico 3	Incremento de cobertura al servicio de salud	Tópico 36	registro sanitario en alimentos y productos
Tópico 4	Importancia de detectar a tiempo cáncer en niños y adolescentes	Tópico 37	Inspección vigilancia y control en alimentos como pescado
Tópico 5	Verificar registro sanitario y prevenir intoxicaciones	Tópico 38	no informativo
Tópico 6	Hábitos saludables para prevenir la diabetes	Tópico 39	Alimentación balanceada y actividad física
Tópico 7	Casos de sarampión prevención y control	Tópico 40	Colombia en contra del trabajo infantil
Tópico 8	No informativo	Tópico 41	no informativo
Tópico 9	Evitar propagación de virus y correcta conservación de los alimentos	Tópico 42	no informativo

Continuación tabla 12

Tema	Nombre	Tema	Nombre
Tópico 10	Formas para prevenir un resfriado-gripa	Tópico 43	no informativo
Tópico 11	Síntomas y tratamiento de enfermedades transmisibles	Tópico 44	Aplicación POSPópuli (permite conocer de manera fácil y rápida cuales medicamentos, tratamientos y procedimientos cubre el plan de beneficios en salud)
Tópico 12	Uso adecuado de los recursos proporcionados por el sistema	Tópico 45	Alerta sanitaria medicina milagrosa, leucemia síntomas y tratamientos
Tópico 13	Atención a la primera infancia y estilo de vida saludable	Tópico 46	Riesgos vacuna VPH
Tópico 14	Lucha contra el cáncer y prevención del consumo de alcohol	Tópico 47	Infórmese sobre sus derechos
Tópico 15	Regulación de precios de medicamentos y cobertura total para enfermedad renal crónica	Tópico 48	Atención integral a personas con Cáncer
Tópico 16	Estudio Zika en embarazadas y niños, modelo IVC (inspección, vigilancia y control)	Tópico 49	Modelo integral de atención de salud (MIAS)
Tópico 17	Enfermedades huérfanas, importancia de informar/se sobre la calidad y disponibilidad del servicio	Tópico 50	Cambio de EPS a través de portal web
Tópico 18	Medidas para controlar el sarampión, factores de riesgo de pérdida auditiva	Tópico 51	Productos regulados por el invima
Tópico 19	Canales oficiales para quejas, denuncias y reclamos, y lesionados con pólvora	Tópico 52	no informativo
Tópico 20	no informativo	Tópico 53	Dona sangre
Tópico 21	Vacunación VPH (virus del papiloma humano)	Tópico 54	Promoción medicina familiar
Tópico 22	no informativo	Tópico 55	No informativo
Tópico 23	Información que deben tener las etiquetas de los productos	Tópico 56	Sarampión
Tópico 24	Facebook, plataformas y páginas web donde se encuentran temas referentes a salud	Tópico 57	Diabetes, alimentación saludable, jornada vacunación
Tópico 25	no informativo	Tópico 58	Cálculos riñones, legalidad e irregularidades en productos
Tópico 26	Canales para denunciar irregularidades en el etiquetado de productos empacados	Tópico 59	Revisar etiquetas de medicamentos y alimentos, problemas de salud producidos por la manipulación de asbesto,
Tópico 27	Control de padres sobre redes sociales, respetar y apoyar la misión medica	Tópico 60	Manual de uso de cosméticos
Tópico 28	no informativo	Tópico 61	Efectos en la salud producidos por el mercurio, impuesto al tabaco
Tópico 29	no informativo	Tópico 62	Eventos adversos a medicamentos foream, prevenga enfermedades respiratorias
Tópico 30	Fortalecer sistema de salud a través de la legalidad	Tópico 63	No a la corrupción e ilegalidad
Tópico 31	Informe acerca de irregularidades en cosméticos o medicamentos	Tópico 64	Promoción y mantenimiento y materno perinatal.
Tópico 32	Cáncer de cuello uterino	Tópico 65	Prevenga enfermedades respiratorias
Tópico 33	no informativo	Tópico 66	Herramientas de georreferenciación invima

7.2 Análisis del impacto

Para el análisis general de los documentos se tienen en cuenta los tweets, en donde los tópicos ocurren con probabilidades superiores al 50% al muestrear el documento (tweet), así, en total se consideran 3.815 tweets. Cabe resaltar que cada tópico puede estar conformado por 10, 12 o 15

tweets, que fueron repetidos a lo largo del cuatrimestre, cada uno con diferentes interacciones ya sea como favoritos y/o retweets, teniendo en cuenta que a los tweets de igual contenido les asigna por igual la misma probabilidad.

Por ejemplo; para el tópico 14- Lucha contra el cáncer y prevención del consumo de alcohol, en la Tabla 13 se observa la probabilidad de ser asignado a sus tweets, de igual manera se observa que para los tweets iguales en contenido pero publicados en diferentes fechas, el tópico fue se asignado con la misma probabilidad.

Tabla 13
Probabilidad del tópico 14 en ser asignado a los documentos

Fecha	Tweet	Probabilidad
25 de abril de 2018	Cali será la primera ciudad del mundo donde se implementará el C/Can 2025: Desafío de Ciudades Contra el Cáncer. La princesa de Jordania, @DinaMired, presidenta de la Unión Internacional Contra el Cáncer, @UICC, explicó al ministro @agaviriau los pormenores de la iniciativa. https://t.co/2XDjjVPDU8	92.55%
26 de abril de 2018	Cali, la primera ciudad del mundo donde se implementará el C/Can 2025: Desafío de Ciudades Contra el Cáncer. La princesa de Jordania, @DinaMired, presidenta de la @UICC, explicó al ministro @agaviriau los pormenores de la iniciativa. https://t.co/2XDjjVPDU8	90.86%
25 de abril de 2018	Cali, la primera ciudad del mundo donde se implementará el C/Can 2025: Desafío de Ciudades Contra el Cáncer. La princesa de Jordania, @DinaMired, presidenta de la @UICC, explicó al ministro @agaviriau los pormenores de la iniciativa. https://t.co/2XDjjVPDU8	90.86%
25 de abril de 2018	El ministro @agaviriau se reunió con la princesa de Jordania, @DinaMired, presidenta de la Unión Internacional Contra el Cáncer, @UICC, para conversar sobre la iniciativa C/Can 2025: Desafío de Ciudades Contra el Cáncer. @INCancerologia https://t.co/vCNBtHiX0l	89.69%
25 de abril de 2018	El ministro @agaviriau se reúne con la princesa de Jordania, @DinaMired, presidenta de la Unión Internacional Contra el Cáncer, @UICC, para conversar sobre la iniciativa C/Can 2025: Desafío de Ciudades Contra el Cáncer. @AlexDuran76 @SaludCali @WiesnerCarolina @INCancerologia	89.69%
9 de abril de 2018	Si no consumo alcohol prevengo futuros daños en hígado y riñones. #MiSaludEsPrimero #EstilosDeVidaSaludables	69.76%
6 de abril de 2018	Si no consumo alcohol prevengo futuros daños en hígado y riñones. #MiSaludEsPrimero #EstilosDeVidaSaludables	69.76%
16 de febrero de 2018	Existen muchas creencias sobre lo que genera o produce el #cancer. Conozca de primera mano cuales son mito y cuales realidad. #mejortempranoconCPreD	65.34%
6 de febrero de 2018	Existen muchas creencias sobre lo que genera o produce el #cancer. Conozca de primera mano cuales son mito y cuales realidad. #mejortempranoconCPreD	65.34%

Continuación tabla 13:

Fecha	Tweet	Probabilidad
4 de febrero de 2018	Hoy, únete #DíaMundialContraElCáncer. existen muchas creencias sobre lo que genera o produce el #cancer. Conozca de primera mano cuales son mito y cuales realidad. #	63.05%
13 de marzo de 2018	Si no consumo alcohol reduzco el riesgo de cáncer y mejoro la salud de mi hígado. #EstilosDeVidaSaludables. https://t.co/q0cacWxHdu https://t.co/cVQH5BTkJq	62.52%
12 de marzo de 2018	Si no consumo alcohol reduzco el riesgo de cáncer y mejoro la salud de mi hígado. #EstilosDeVidaSaludables. https://t.co/q0cacWxHdu https://t.co/cVQH5BTkJq	62.52%
9 de marzo de 2018	Si no consumo alcohol reduzco el riesgo de cáncer y mejoro la salud de mi hígado. #EstilosDeVidaSaludables. https://t.co/q0cacWxHdu https://t.co/cVQH5BTkJq	62.52%
8 de marzo de 2018	Si no consumo alcohol reduzco el riesgo de cáncer y mejoro la salud de mi hígado. #EstilosDeVidaSaludables. https://t.co/q0cacWxHdu https://t.co/cVQH5BTkJq	62.52%
7 de marzo de 2018	Si no consumo alcohol reduzco el riesgo de cáncer y mejoro la salud de mi hígado. #EstilosDeVidaSaludables. https://t.co/q0cacWxHdu https://t.co/cVQH5BTkJq	62.52%
6 de marzo de 2018	Si no consumo alcohol reduzco el riesgo de cáncer y mejoro la salud de mi hígado. #EstilosDeVidaSaludables. https://t.co/q0cacWxHdu https://t.co/cVQH5BTkJq	62.52%
4 de febrero de 2018	#DíaMundialcontraelCáncer la lucha contra el cáncer la hacemos compartiendo derechos en salud	62.47%
21 de febrero de 2018	En la @FNDCol, el ministro @agaviriau se reúne con gobernadores para cerrar las modificaciones al decreto de operación de @AdresCol. https://t.co/1WdeNh3AOi	58.38%
6 de febrero de 2018	Tendencias de mortalidad por cáncer tanto en hombres como mujeres en Argentina, EE.UU y Colombia varía en éstos países. #ControlCáncer en @UniJaveriana https://t.co/Grm6srFVv3	51.77%

Nota: Adaptado de Rstudio versión 3.5.0 (2018)

7.2.1 Análisis General

7.2.1.1 Análisis de impacto por tweets. Al analizar el porcentaje de tweets por cada uno de los tópicos, en la Figura 11, se muestra el porcentaje de documentos en los cuales el tópico tiene una probabilidad superior al 50% de ser asignado al documento. Por ejemplo, el tópico 15 (Regulación de precios de medicamentos y cobertura total para enfermedad renal crónica) con 183 tweets (4.8 %), y el tópico 14 (Lucha contra el cáncer y prevención del consumo de alcohol) con 22 tweets (0.58%), son los tópicos con el mayor y el menor número de documentos, respectivamente, durante el primer cuatrimestre del año 2018.

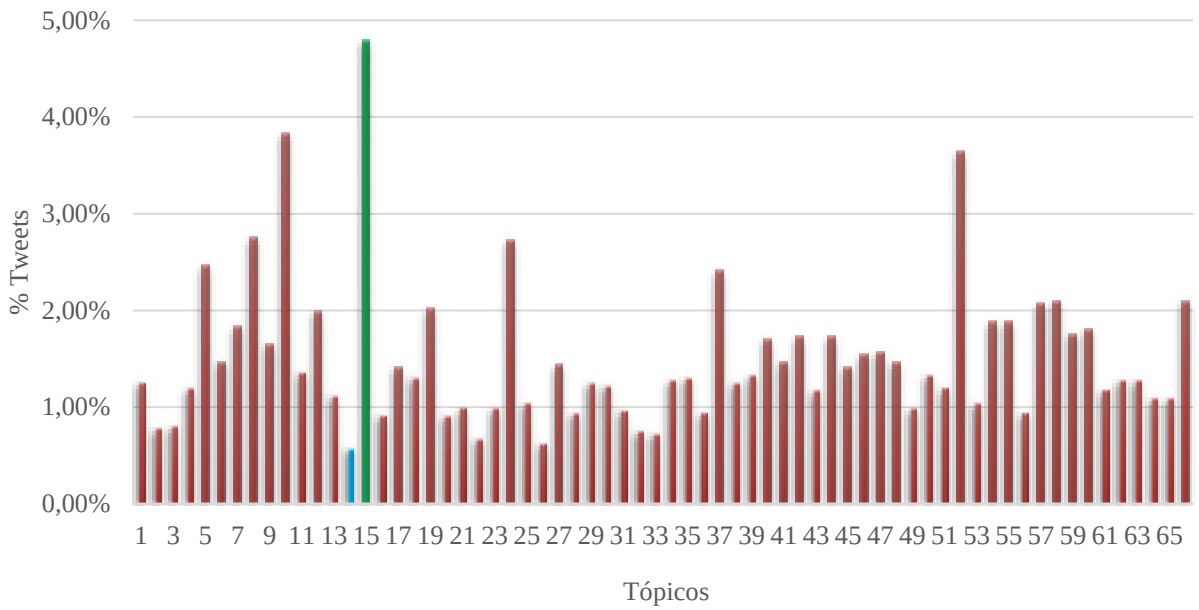


Figura 11: Porcentaje de tweets por tópico

Adicionalmente, teniendo en cuenta el porcentaje de tweets que son publicados por mes, en la figura 12 se muestra que en los meses de enero y febrero el flujo de tweets fue bajo, con 428 y 855 tweets respectivamente, siendo marzo (1279 tweets) y abril (1253 tweets), los meses en los que las entidades realizan más publicaciones. Así, en la Figura 13, Figura 14, Figura 15 y Figura 16, se muestra para cada mes el porcentaje de tweet correspondientes a cada tópico

Adicionalmente, con el fin de identificar cuáles tópicos son más o menos twitteados en cada uno de los primeros cuatro meses del año 2018, en la Figura 12 se muestra que en el mes de enero al ser el mes con menos publicaciones de Tweets también tuvo pocos temas tratados. En el mes de febrero se evidencia que los tweets que más fueron publicados no se enfocan en temas de promoción de la salud y prevención de enfermedades dadas por las entidades de salud a nivel nacional en sus páginas web.

Tabla 14
Tópicos más y menos twitteados por mes

Mes	Tópicos más twitteados	Tópicos menos twitteados
Enero	Regulación de precios de medicamentos y cobertura total para enfermedad renal crónica (Tópico 15)	Financiación de medicamentos y procedimientos con recursos públicos (Tópico 2)
	Canales oficiales para quejas, denuncias y reclamos, y lesionados con pólvora (Tópico 19)	Importancia de detectar a tiempo cáncer en niños y adolescentes (Tópico 4)
	Herramientas de georreferenciación Invima (Tópico 66)	Síntomas y tratamiento de enfermedades transmisibles (Tópico 11)
Febrero	Regulación de precios de medicamentos y cobertura total para enfermedad renal crónica (Tópico 15)	Sarampión (Tópico 56)
	No informativo - Atención al cliente (Tópico 52)	Promoción y mantenimiento y materno perinatal (Tópico 64)
	No informativo - Afiliación de empleados al sistema de seguridad social (Tópico 55)	Colombia sin asbesto (Tópico 35)
Marzo	No informativo (tópico 8)	Síntomas y tratamiento de enfermedades transmisibles (Tópico 11)
	Promoción medicina familiar (tópico 54)	Enfermedades huérfanas, importancia de informar/se sobre la calidad y disponibilidad del servicio (Tópico 17)
	Formas para prevenir un resfriado-gripa (Tópico 10)	no informativo (Tópico 41)
Abril	Aplicación POSPópuli (permite conocer de manera fácil y rápida cuales medicamentos, tratamientos y procedimientos cubre el plan de beneficios en salud)	Importancia de detectar a tiempo cáncer en niños y adolescentes (Tópico 4)
	Síntomas y tratamiento de enfermedades transmisibles	Control de padres sobre redes sociales, respetar y apoyar la misión médica (tópico 27)
	Regulación de precios de medicamentos y cobertura total para enfermedad renal crónica (Tópico 15)	Incremento de cobertura al servicio de salud (Tópico 3)

Nota: Adaptado de Rstudio versión 3.5.0 (2018)

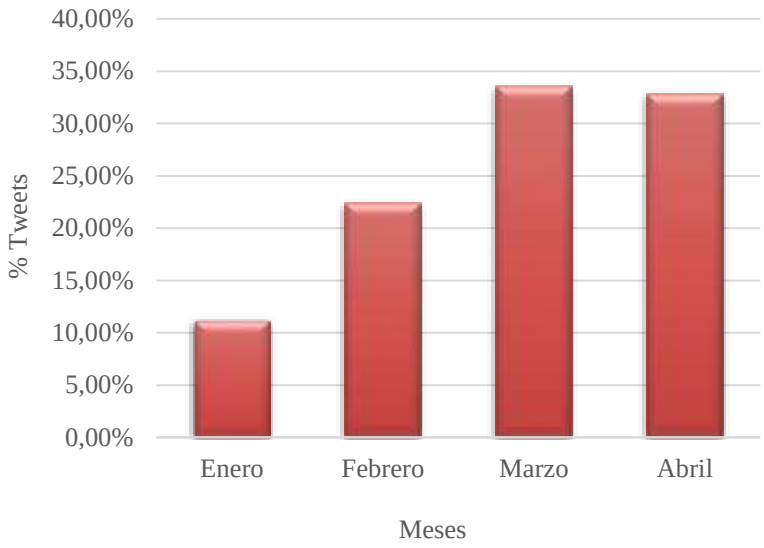


Figura 12: Porcentaje de Tweets por mes

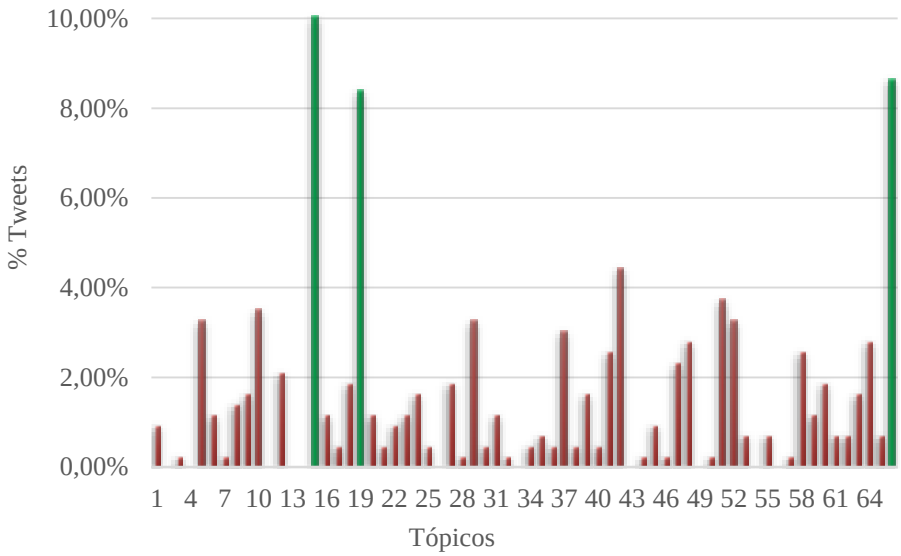


Figura 13: Porcentaje de tweets de enero por tópico

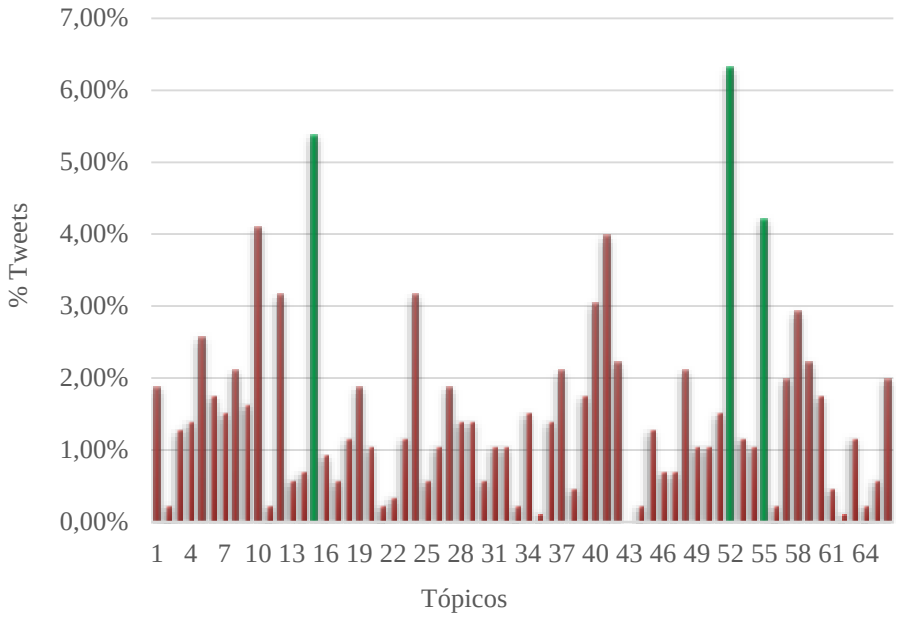


Figura 14: Porcentaje de tweets de febrero por tópico

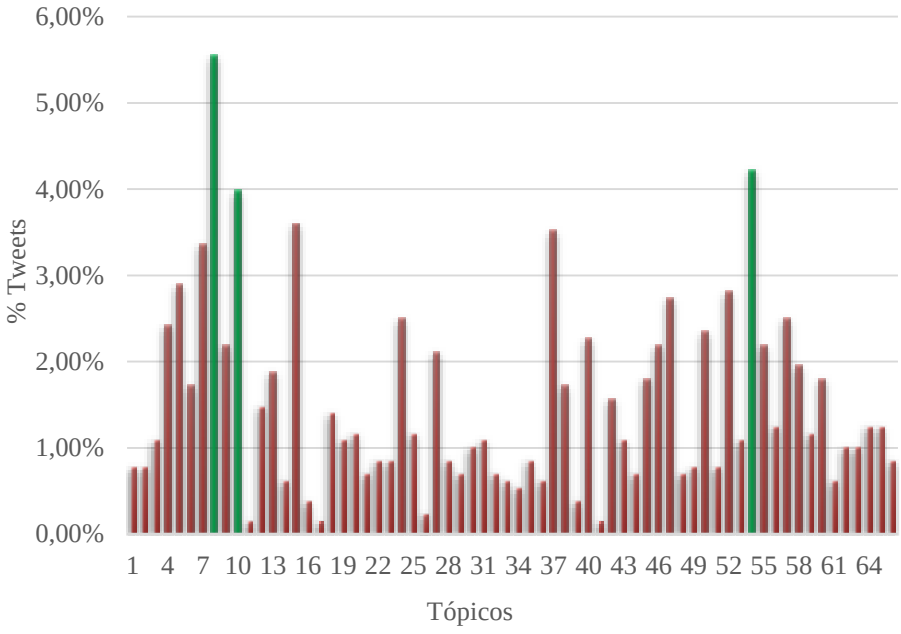


Figura 15: Porcentaje de tweets de marzo por tópico

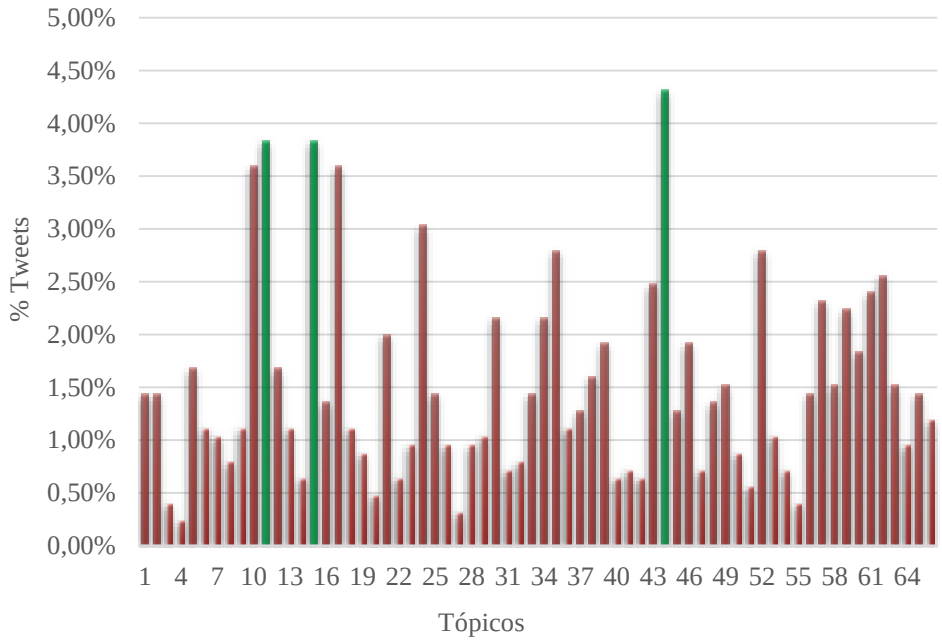


Figura 16: Porcentaje de tweets de abril por tópico

En conclusión, durante el primer cuatrimestre del año 2018, se denota que los tópicos: Formas para prevenir un resfriado-gripa (Tópico 10), Regulación de precios de medicamentos y cobertura total para enfermedad renal crónica (Tópico 15) y no informativo - Atención al cliente (Tópico 52), se ubican entre los 10 primeros tópicos de cada mes. Esto evidencia que la tendencia en enfermedades es la gripa y el interés de parte de las autoridades y entidades de salud pública a difundir estas alertas de cuidado y un tema de tendencia como lo era los precios de los medicamentos, así como el interés por dar solución a las inquietudes de los usuarios.

7.2.1.2 Análisis de impacto por Favoritos. En la Figura 17, se observa la cantidad de favoritos por tópicos en el periodo de enero a abril del 2018; sin embargo, se observa que en el tópico 15, al ser asignado a mayor cantidad de tweets , no es el más preferido como favorito por los usuarios presentando solamente un 2,56% de favoritos (508 Favoritos), esto se puede deber a que los temas relacionados (Regulación de precios de medicamentos y cobertura total para enfermedad renal

crónica) son de poco interés para los usuarios, pero de gran interés para las entidades que los twittean; por el contrario en el tópico 46 (Riesgos vacuna VPH), se observa que a pesar ser asignado al muestrear en el 1.03% de los documentos (60 tweets), presenta el porcentaje más alto de favoritos con el 4,53% (898 Favoritos).

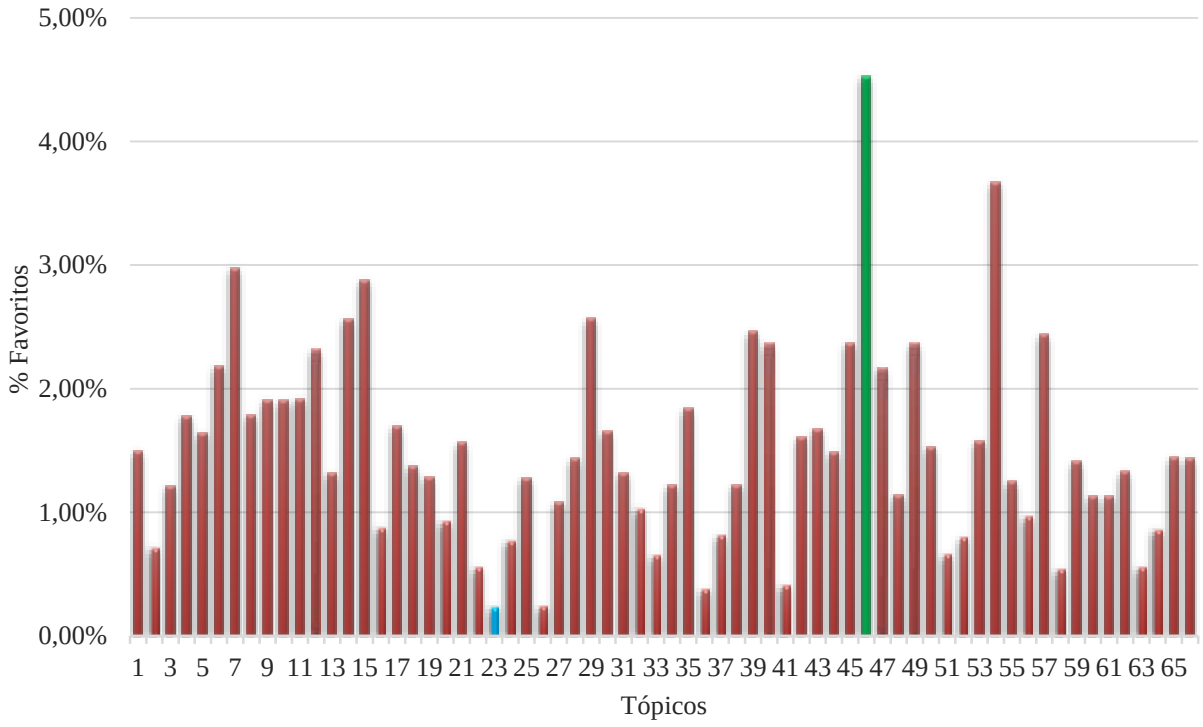


Figura 17: Porcentaje de favoritos por tópico

Adicionalmente, en la Figura 18, se muestra la frecuencia de los favoritos durante el cuatrimestre, encontrando 849 tweets que no representan ningún favorito y un tweet con 162 favoritos, el cual mencionaba: “La Fundación BBVA premia a Nubia Muñoz, precursora del desarrollo de la primera vacuna contra el cáncer de cérvix, vía @infosalus_com”.

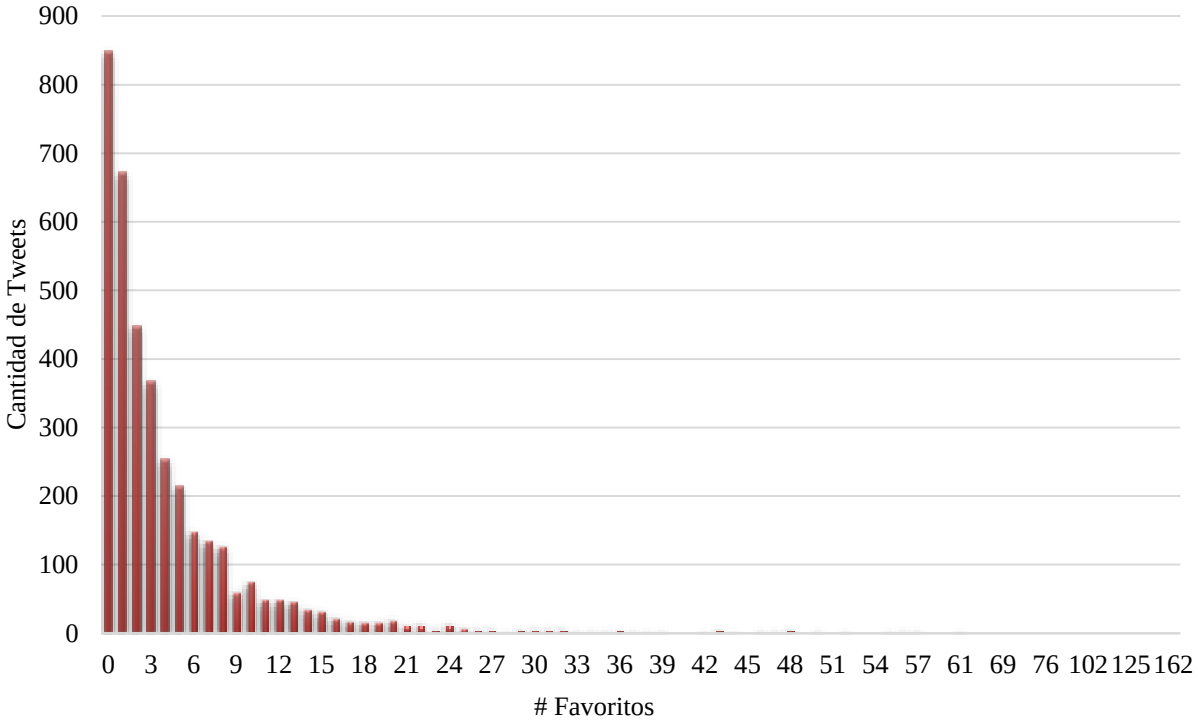


Figura 18: Cantidad de Tweet por favoritos

7.2.1.3 Análisis de impacto por Retweet. En la Figura 19 se muestra el porcentaje de Retweet por tópico, siendo el tópico 7 (Casos de sarampión prevención y control) el que presenta mayor número de Retweet con un 3% (936 Retweet) y el Tópico 23 el menor número de Retweets con el 0.11% (33 Retweets).

En la Figura 20, se observa que 1.017 tweets no fueron Retwitteados en el periodo de enero a abril del 2018, de los cuales se destaca que los temas de regulación de precios de medicamentos y cobertura total para enfermedad renal crónica (Tópico 15) y los sistemas y tratamientos de enfermedades transmisibles (Tópico 11), son los que presentan mayor cantidad de tweets sin ser Retwitteados, con el 60.65% (111 tweets) y 77,68% (108 tweets), respectivamente, con relación al total de tweets del tópico. Por otro lado, los tópicos: Lucha contra el cáncer y prevención del consumo de alcohol (Tópico 14), Cáncer de cuello uterino (Tópico 32) y Alimentación balanceada

y actividad física (Tópico 39), presentan todos sus tweets con al menos un Retweet.

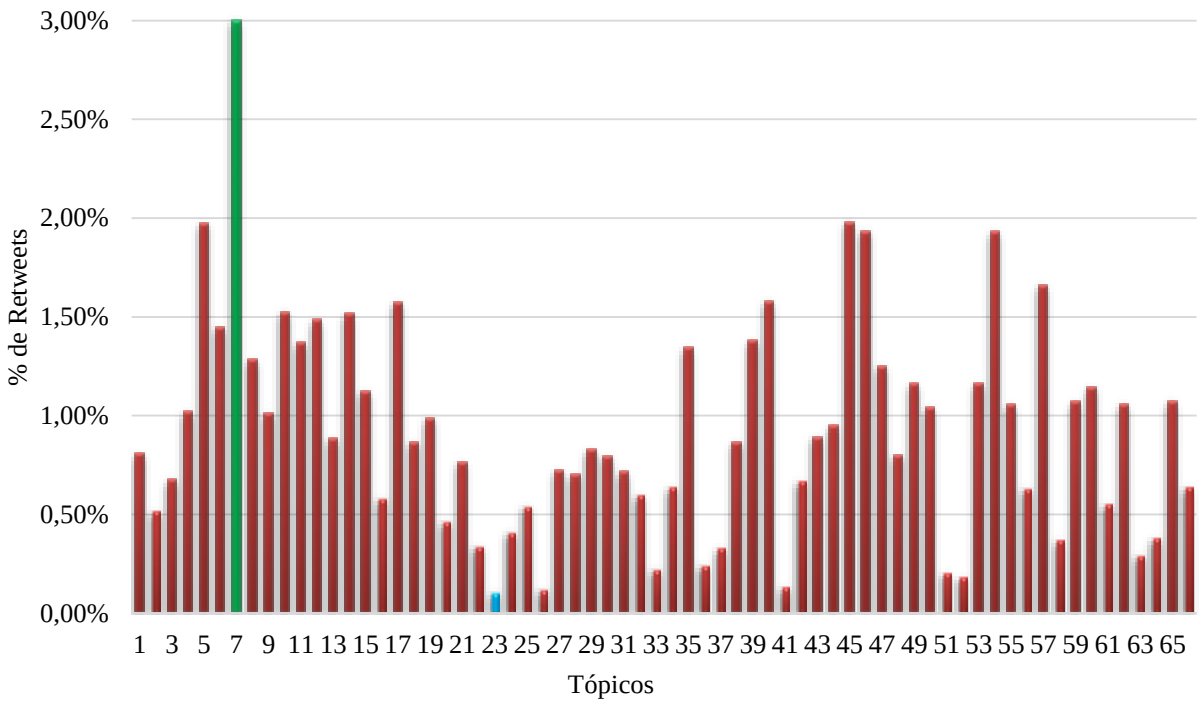


Figura 19: Porcentaje de retweets por tópicos

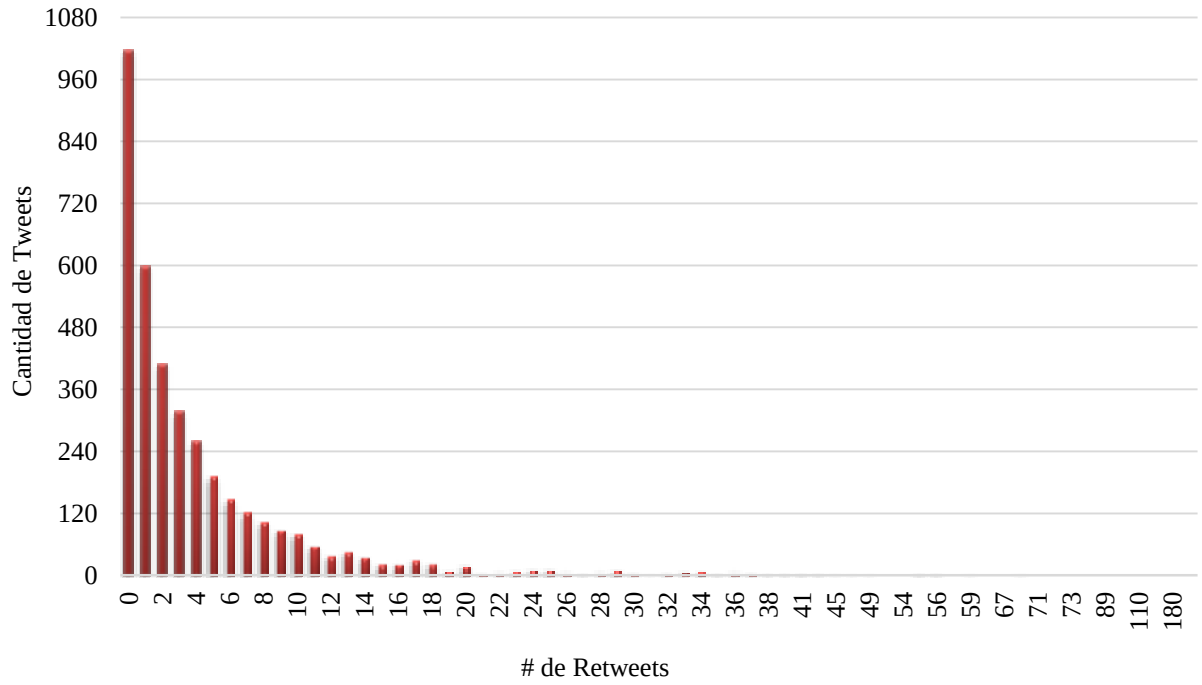
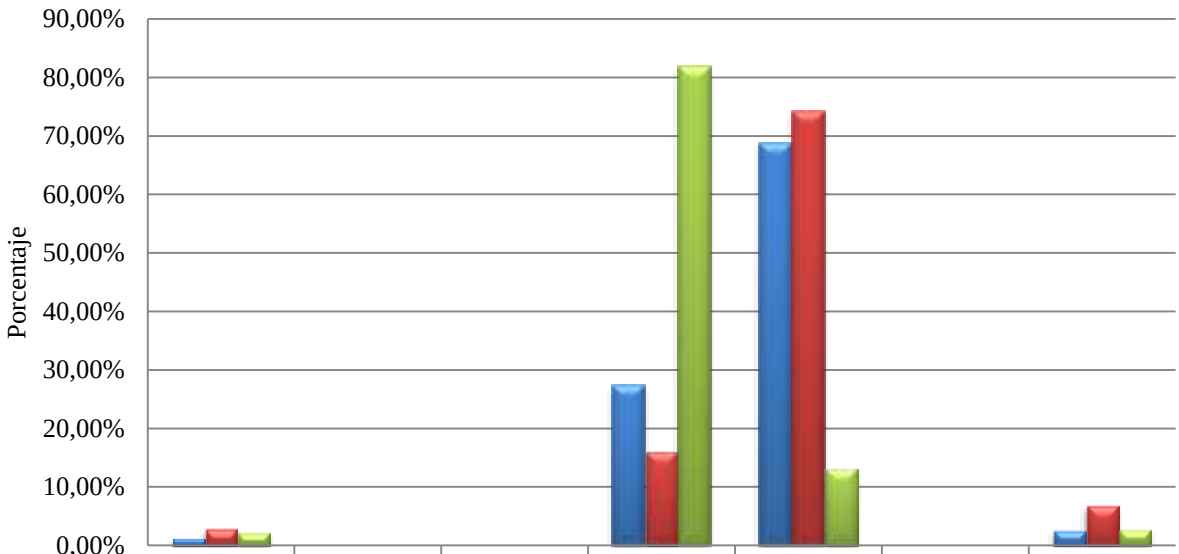


Figura 20: Cantidad de tweets por número de retweets

7.2.2 Tópicos Representativos por usuarios de Twitter.

7.2.2.1 Tópico 15. En la Figura 21, se observa por entidad los tweets, Retweets y favoritos, para el Tópico Regulación de precios de medicamentos y cobertura total para enfermedad renal crónica (tópico 15), teniendo en cuenta que este tópico se caracteriza por agrupar la mayoría de tweets. De este modo se puede denotar que el tópico es representativo en las publicaciones del INVIMA; sin embargo, presenta más favoritos y retweets por parte de las publicaciones que hace el MINSALUD lo cual se debe a que los tweets del INVIMA se relacionan más con dar información del compromiso que tienen de cuidar la vida, mientras el MINSALUD proporcionó información de precios, cobertura en la atención y medicamentos para la enfermedad renal crónica.



	dermatologico	INCancerologia	INSColombia	invimacolombia	MinSaludCol	OPSOMS_Col	Supersalud
Favoritos	1,23%	0,00%	0,00%	27,32%	68,83%	0,00%	2,63%
Retweet	2,86%	0,00%	0,00%	16,00%	74,29%	0,00%	6,86%
Tweets	2,19%	0,00%	0,00%	81,97%	13,11%	0,00%	2,73%

Figura 21: Porcentaje de favoritos/retweets/tweets del tópico 15 según la entidad

También muestra que a pesar de que el INVIMA realizó la mayor cantidad de Tweets en este

tópico 15; el MINSALUD obtuvo más retweets y favoritos en los pocos tweets que publicó; esto se puede deber a que el tema de precios y cobertura es más significativo para los usuarios, que el hecho de recalcar un compromiso de salud a nivel general.

También, cabe aclarar que de los 183 tweets del tópico 15, el 60.66% (111 Tweets) no fueron Retwitteados, siendo todos del INVIMA (Figura 22), los tweets que más fueron Retwitteados fueron del MINSALUD, como, por ejemplo:

- "Tenemos que resistir a la tentación de destruir sin haber construido": @agaviriau en "El futuro de la salud en Colombia". @ForosSemana
- «Si, guiados por muestras que no son representativas, percibimos que el sistema de salud es un fracaso sin atenuantes, tendremos la tentación de destruir sin haber construido»: @agaviriau ante autoridades y líderes de la salud del Meta.
- El 98% de colombianos tiene cobertura de salud. Hay garantías para quienes sufren enfermedad que exige tratamientos costosos #SaludParaTodos. #DíaMundialdelaSalud
- Colombia es ahora un referente en salud: regulamos precios de unos mil medicamentos y hay cobertura total para enfermedad renal crónica. #SaludParaTodos. #DíaMundialDeLaSalud

Es importante recalcar que, de los anteriores tweets, los dos primeros que más fueron Retwitteados no tienen relación directa con el tema principal del tópico respecto a la regulación de precios y medicamentos de la enfermedad renal.

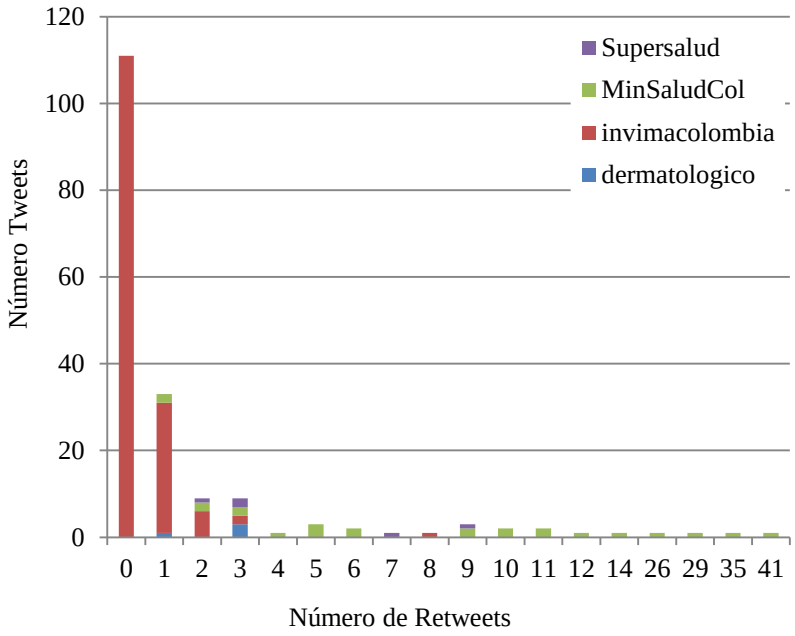


Figura 22: Porcentaje tweets por número de Retweets según la entidad para el tópico 15

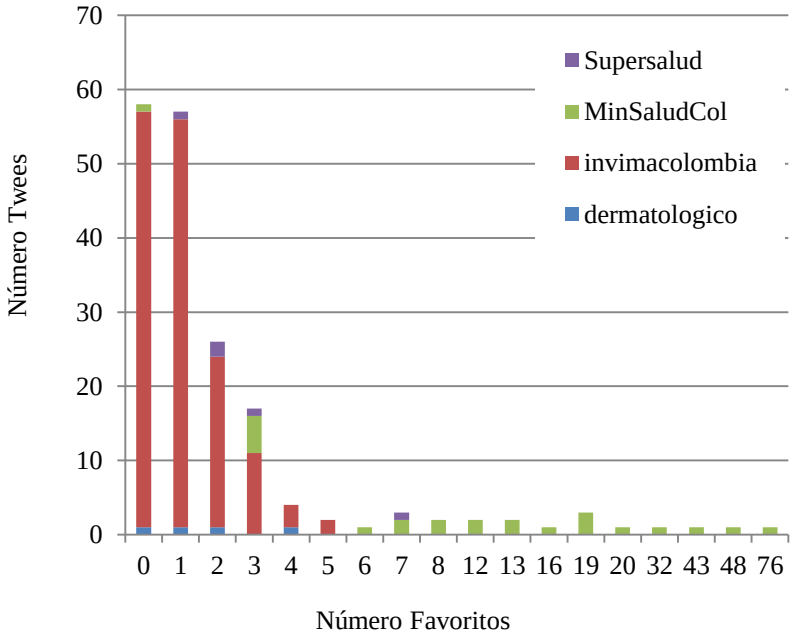


Figura 23: Porcentaje tweets por número de favoritos según la entidad para el tópico 15

De igual manera en la Figura 23 se muestra que los favoritos se concentraron en mayor medida en unos cuantos tweets del MINSALUD; coincidiendo con los tweets que más fueron Retwitteados.

Es importante destacar que el tópico 15, está compuesto de 14 tweets que fueron replicados durante los 4 meses, en la Figura 24, Figura 25, Figura 26, se observa respectivamente el número de tweets del tópico 15 con respecto al porcentaje de tweets, favoritos y Re tweets para cada uno de los meses. Así,

- Tweet 1: “Tenemos que resistir a la tentación de destruir sin haber construido”: @agaviriau en “El futuro de la salud en Colombia”. @ForosSemana
- Tweet 2: «Si, guiados por muestras que no son representativas, percibimos que el sistema de salud es un fracaso sin atenuantes, tendremos la tentación de destruir sin haber construido»: @agaviriau ante autoridades y líderes de la salud del Meta.
- Tweet 3: El 98% de colombianos tiene cobertura de salud. Hay garantías para quienes sufren enfermedad que exige tratamientos costosos #SaludParaTodos.
- Tweet 4: El vitíligo no tiene una causa determinante, sin embargo existen algunos factores de riesgo que se pueden asociar a su aparición #AmoyExaminoMiPiel @MinSaludCol @SectorSalud
- Tweet 5: Esta es la historia de vida de famosos como @winnieharlow y Asley Soto que padecen de vitíligo y crearon otro concepto de belleza.
- Tweet 6: #DermaDato: La fototerapia consiste en el tratamiento de diferentes enfermedades de la piel mediante la radiación ultravioleta y se utiliza en diferentes enfermedades dermatológicas como: psoriasis, dermatitis atópica, vitíligo, entre otras. #AmoYExaminoMiPiel

- Tweet 7: #Dermadato: La psoriasis ocupa el 2% dentro de las enfermedades dermatológicas. El Dr. Juan Raúl Castro nos explica los tipos de psoriasis que existen. #AmoyExaminoMiPiel @MinSaludCol
- Tweet 8: En #Invima protegemos lo esencial para cuidar tu vida
- Tweet 9: En #Invima protegemos lo esencial para que tu empresa crezca
- Tweet 10: Colombia es ahora un referente en salud: regulamos precios de unos mil medicamentos y hay cobertura total para enfermedad renal crónica. #SaludParaTodos.
- Tweet 11: Regulamos los precios de casi mil medicamentos y logramos cobertura total para la enfermedad renal crónica. #SaludParaTodos. #DíaMundialDeLaSalud
- Tweet 12: Las mujeres son más propensas a desarrollar enfermedades de carácter crónico. #GéneroySalud
- Tweet 13: #DíaMundialdelaSalud Hay garantías para quienes sufren enfermedad que exige tratamientos costosos #SaludParaTodos
- Tweet 14: #SancionesSNS @Supersalud sancionó a la Caja de Compensación Familiar de Cartagena y Bolívar con 30 SMLMV por no reportar de manera oportuna la información solicitada mediante Circular Única.

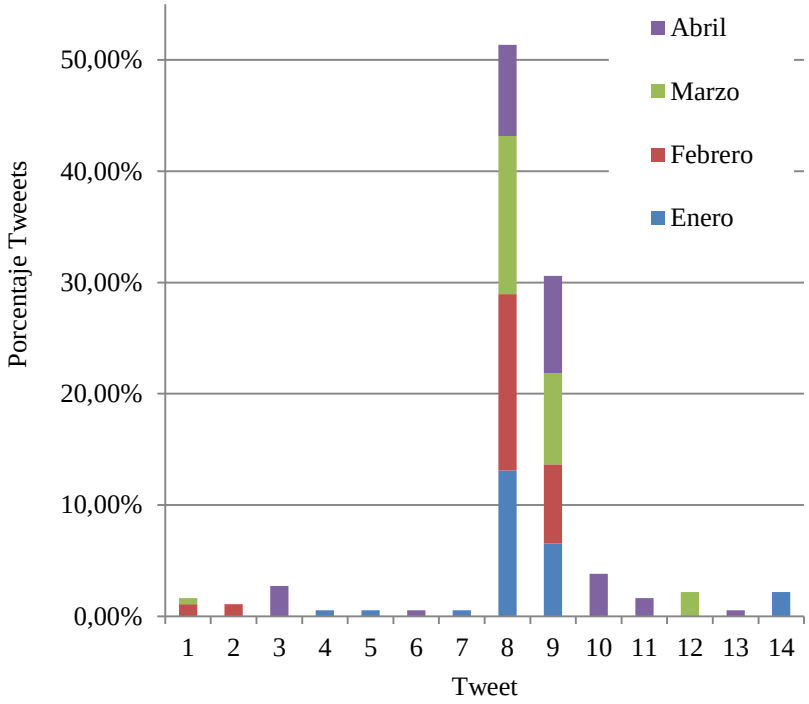


Figura 24: Número de tweets por mes según tweet tópico 15

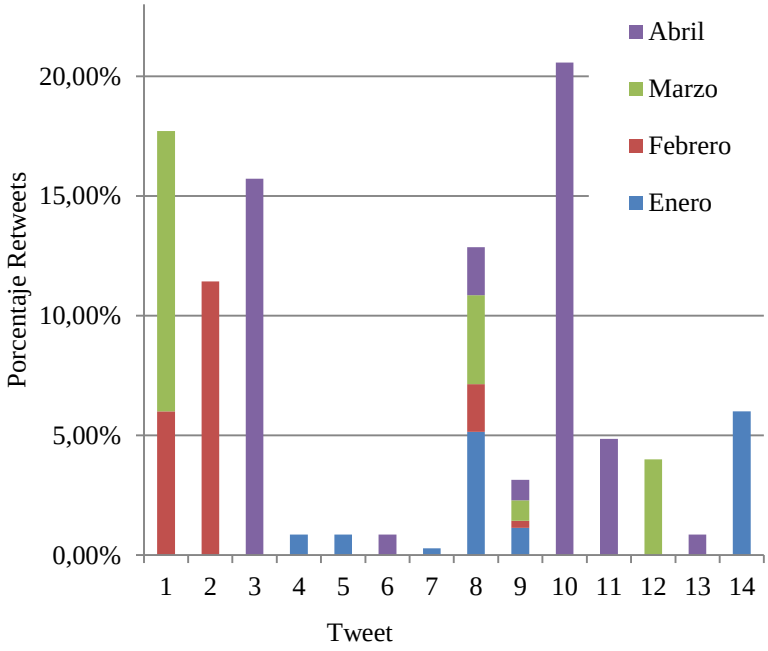


Figura 25: Número de Retweets por mes según tweet tópico 15

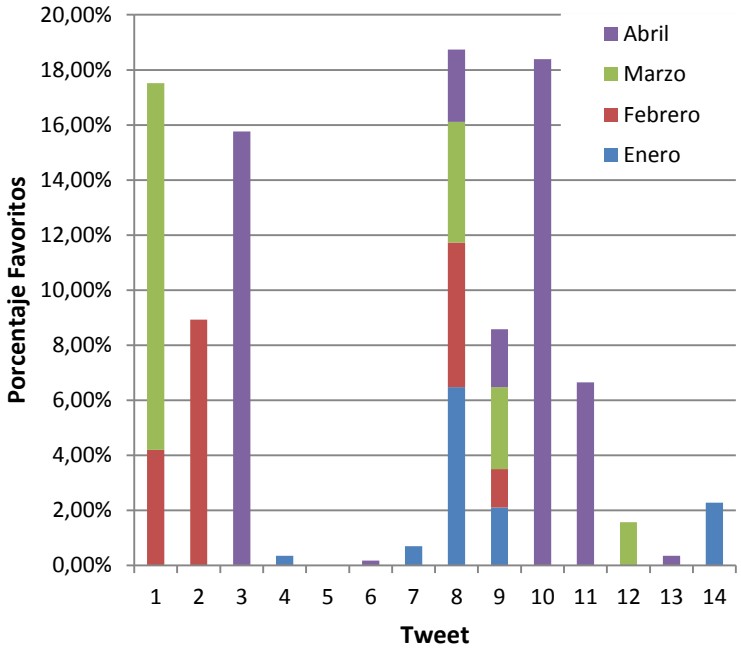
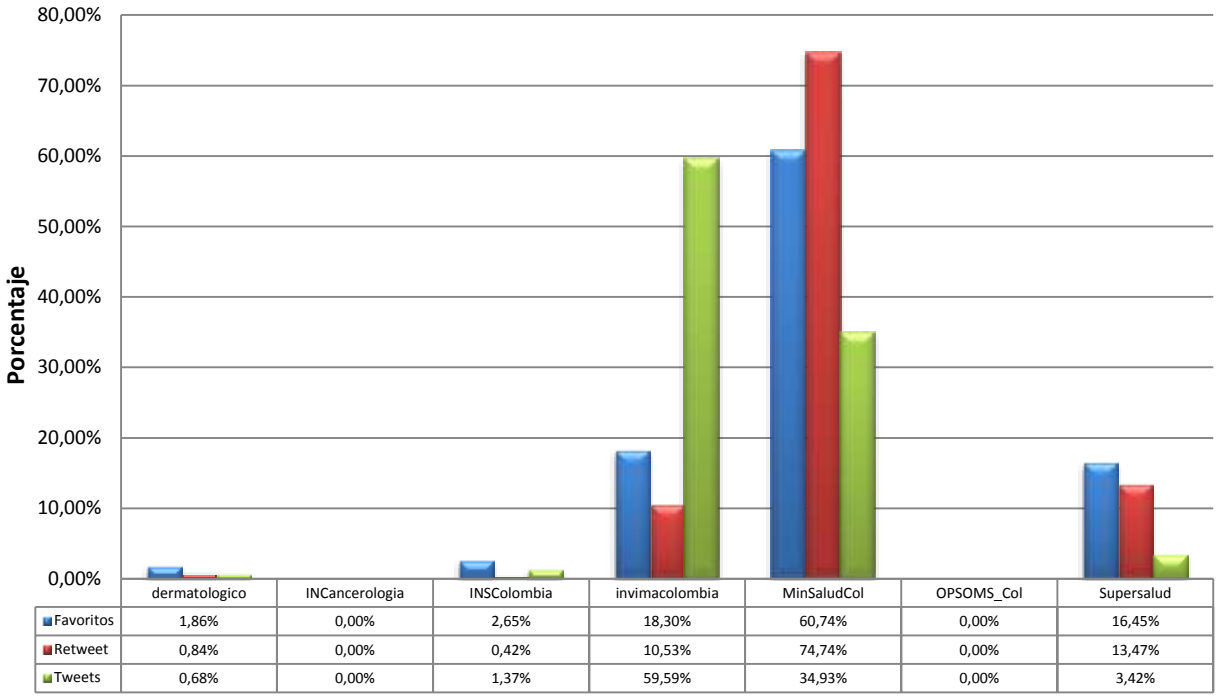


Figura 26: Número de favoritos por mes según el tweet tópico 15

Algunos tweets como “En #Invima protegemos lo esencial para cuidar tu vida” y “En #Invima protegemos lo esencial para que tu empresa crezca” (Ver figura 24), fueron twitteados de manera proporcional en los 4 meses; sin embargo, no son los tweets con más Retweets ni favoritos (Ver Figura 25 y Figura 26) del tópico; destacando que tweets con pocas publicaciones recibieron más interacción por parte de los usuarios, dentro de estos tweets se encuentran: “Tenemos que resistir a la tentación de destruir sin haber construido: @agaviriau en “El futuro de la salud en Colombia. @ForosSemana”, El 98% de colombianos tiene cobertura de salud. Hay garantías para quienes sufren enfermedad que exige tratamientos costosos #SaludParaTodos”, “«Si, guiados por muestras que no son representativas, percibimos que el sistema de salud es un fracaso sin atenuantes, tendremos la tentación de destruir sin haber construido»: @agaviriau ante autoridades y líderes de la salud del Meta”.

7.2.2.2 Tópico 10. Este tópico relacionado con las formas para prevenir un resfriado-gripa y

alertas sanitarias, tiene mayor presencia en los tweets del MINSALUD, INVIMA y Supersalud; sin embargo, los Retweets y favoritos se presentan en mayor proporción en MINSALUD y Supersalud, como se muestra en la Figura 27.



	dermatologico	INCancerologia	INSColombia	invimacolombia	MinSaludCol	OPSOMS_Col	Supersalud
Favoritos	1,86%	0,00%	2,65%	18,30%	60,74%	0,00%	16,45%
Retweet	0,84%	0,00%	0,42%	10,53%	74,74%	0,00%	13,47%
Tweets	0,68%	0,00%	1,37%	59,59%	34,93%	0,00%	3,42%

Figura 27: Porcentaje de favoritos/Retweets/tweets tópicos 10 según la entidad

Al Analizar los Retweets y favoritos se observa que de los 146 tweets que conforman el tópico, 59 tweets (40,41%) no son Retwitteados y 52 tweets (35,61%) no son marcados como favoritos (Ver Figura 28 y Figura 29); se resalta que la mayoría de tweets son publicados por el INVIMA, dentro de los cuales se encuentran: “#Visite 1. Alertas Sanitarias: 2. Canales de atención”, “Evite el contacto con personas que tengan gripa o tos. Prevenga #EnfermedadesRespiratorias.”

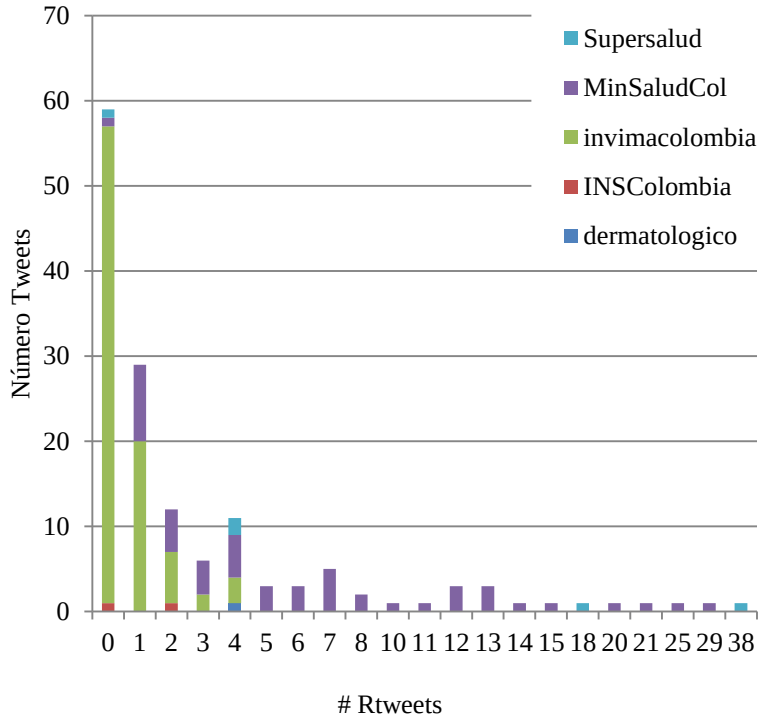


Figura 28. Número tweets por número de Retweets según la entidad para el tópico 10

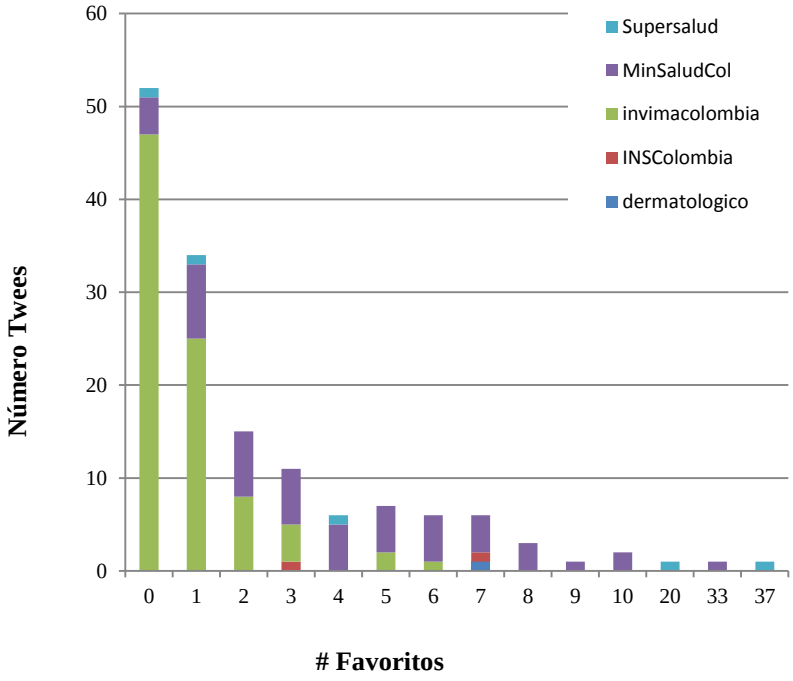


Figura 29: Número tweets por número de favoritos según la entidad para el tópico 10

Es importante destacar que el tópico 10, está compuesto de 10 tweets que son replicados durante los 4 meses. En la Figura 30, Figura 31 y Figura 32, se observa respectivamente el porcentaje de tweets, favoritos y Retweets por mes para cada uno de los tweets que conforman el tópico 10 .Los tweets que pertenecen a este tópico son:

- Tweet 1: La piel es el órgano más extenso y el más sensible del cuerpo. Amarla, cuidarla y protegerla debe ser vital para todos. #AmoyExaminoMiPiel
- Tweet 2: Conflicto armado, emergencias humanitarias, inmigración y desplazamiento influyen en la reproducción de los mosquitos que transmiten la #malaria, mencionó Chaparro @INSColombia
- Tweet 3: Participación en el curso de liderazgo en medidas sanitarias y fitosanitarias del 12 al 16 de marzo en Costa Rica por el @IICA de cooperación para la agricultura, U. para la Paz de las Naciones Unidas y el departamento de agricultura, presencia de 16 participantes de países de América
- Tweet 4: #Visite 1. Alertas Sanitarias 2. Canales de atención
- Tweet 5: Evite el contacto con personas que tengan gripa o tos. Prevenga #EnfermedadesRespiratorias.
- Tweet 6: Nuevos hospitales y puestos de salud de Caucasia, Vigía del Fuerte, Tierralta, Mocoa y Santander de Quilichao son un símbolo de paz para regiones “donde no había Estado porque había conflicto, y había conflicto porque no había Estado”: @agaviriau #SaludParaLaPaz
- Tweet 7: #AudicionSanaySegura en el curso de vida, unos siete millones de personas en

Colombia padecen problemas de audición.

- Tweet 8: Las brechas rural-urbanas y Caribe-centro no se han cerrado. El conflicto armado ha sido una de las causas: @agaviriau en #SeminarioIEEC en la @UninorteCO
- Tweet 9: #MentirasyVerdades Cuando tienes gripa cuando tienes resfriado
- Tweet 10: #MásCercaDeLosUsuarios continúa @Supersalud en las regiones en Neiva con la capacitación a víctimas del conflicto armado para la conformación de veedurías ciudadanas

Tweets como el número 4: "#Visite 1. Alertas Sanitarias 2. Canales de atención:" (Tweet 4), es el tweet que más replico y más twitteado en el tópico (Ver figura 30), y aunque si es marcado como favorito y Retwitteados; no es el tweet con más interacción de este tópico (Ver figura 31 y 32), como lo es el tweet 5: "Evite el contacto con personas que tengan gripa o tos. Prevenga #EnfermedadesRespiratorias. https://t.co/OFBDcXIvgZ"; que si es de gran interés entre los usuarios; aunque no tenga gran cantidad de tweets en el tópico.

De igual forma en el tópico hay tweets como "La piel es el órgano más extenso y el más sensible del cuerpo. Amarla, cuidarla y protegerla debe ser vital para todos. #AmoyExaminoMiPiel "(tweet 1), "Conflicto armado, emergencias humanitarias, inmigración y desplazamiento influyen en la reproducción de los mosquitos que transmiten la #malaria, mencionó Chaparro @INSColombia" (tweet 2), que no fueron tuvieron relevancias en términos de tweets, favoritos y Retweets.

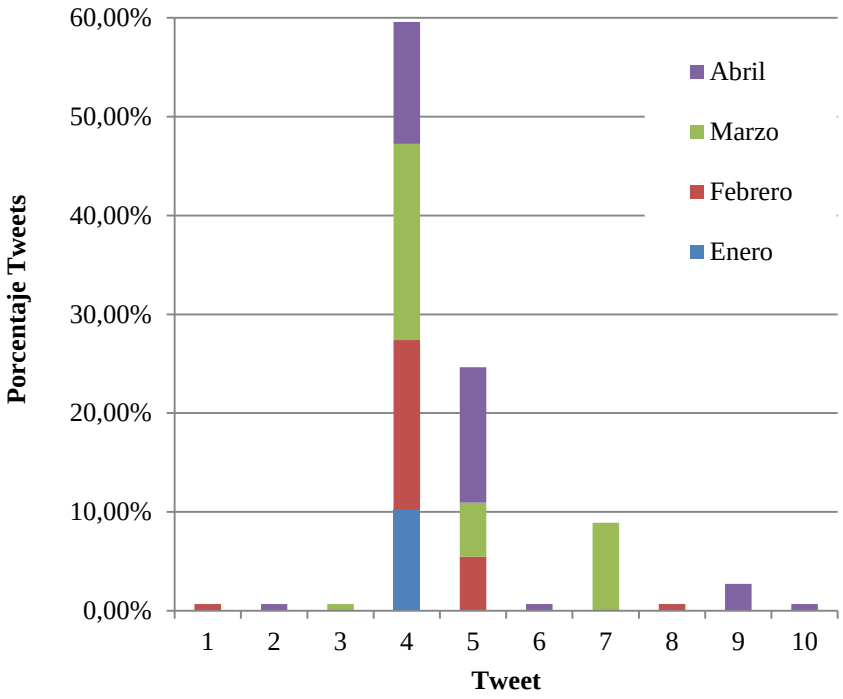


Figura 30: Número de tweets por mes según tweet tópico 10

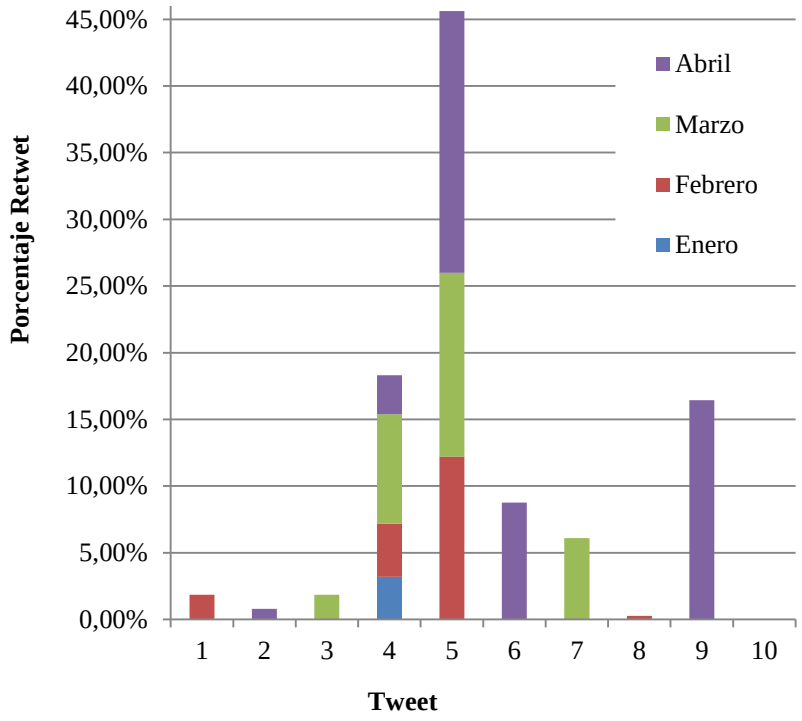


Figura 31: Número de Retweets por mes según tweet tópico 10

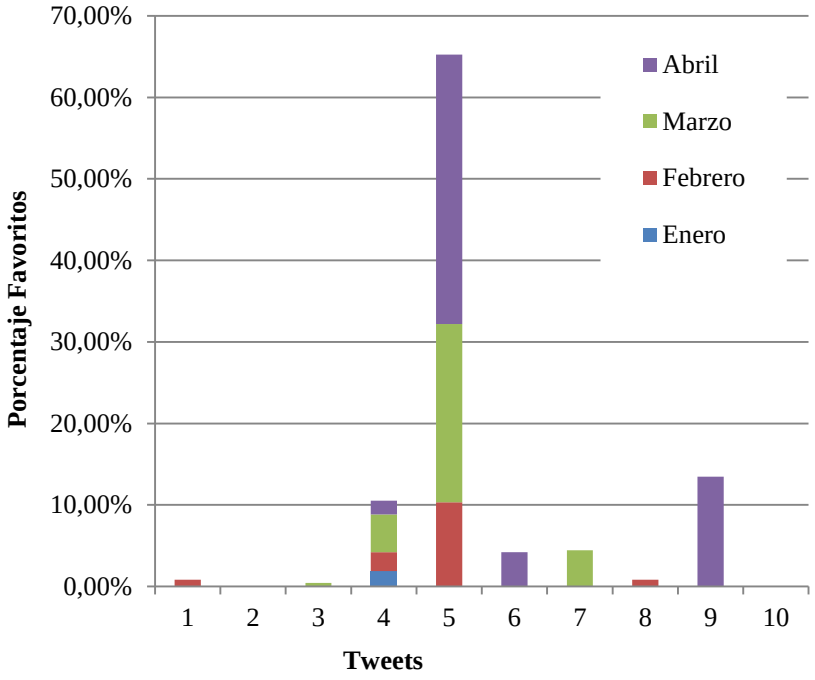


Figura 32: Número de favoritos por mes según el tweet tópico 10

8. Divulgación Académica

Se elabora un artículo y un repositorio en GitHub, con el objetivo de presentar los resultados y desarrollo del presente proyecto

En el artículo (Ver Apéndice B) se resumen los resultados obtenidos del presente proyecto.

El Repositorio de GitHub se encuentra disponible en https://github.com/martha0306/Autoridaades-Nacionales-de-Salud.git. En el cual se encuentra el código usado en las diferentes etapas del descubrimiento de conocimiento.

9. Conclusiones

Los documentos científicos incluidos dentro de la revisión de literatura permiten identificar técnicas de clasificación no supervisada para el análisis de contenido en redes sociales, siendo los más reconocidos k-medias, LDA, CLARA, PAM y DBSCAN, y dentro de estos se destaca LDA como técnica de modelado de tópicos. Así, estos hallazgos permiten ser una guía de métodos considerados en el problema objeto de estudio de este proyecto.

En el ajuste de los modelos de clasificación no supervisada aplicados a la base de datos objeto de estudio, se validan mediante criterios como el *loglikehood* y el tiempo de procesamiento empleado en la estimación de la mejor cantidad de temas óptimos, además, este proceso de validación se apoya con la aplicación de métricas que permiten comparar modelos basados en distribuciones probabilísticas como lo es la distancia de *Hellinger*.

Respecto al análisis de patrones y tendencias, se encuentra que en gran parte, las entidades analizadas publican temas relacionados con campañas de promoción de la salud y prevención de las enfermedades; sin embargo, supone un gran un esfuerzo el hecho de que se publican en diferentes fechas y horas del día, generando así que no todos los tweets lleguen a los usuarios de igual manera, ni se garantice que la mayor parte de los usuarios interactúen con la información publicada, como tampoco que los tweets tengan iguales interacciones, obteniendo una reducción en el impacto que genera el tweet.

Dentro de los tópicos relacionados con la promoción de la salud y prevención de la enfermedad, se encuentran los tópicos en los cuales existen tweets que son publicados en menor frecuencia durante el primer cuatrimestre del año; sin embargo, el contenido o la información proporcionada es de gran interés para los usuarios, como por ejemplo los tweets más predominantes en los tópicos

son los relacionados con gripas, tos, regulación de precios y mejoramiento del sistema de salud colombiano, y los tópicos con mayores favoritos y retweets fueron: promoción de medicina familiar (Tópico 54) , casos de sarampión prevención y control (Tópico 7) , Riesgos de vacuna VPH (Tópico 46).

Se encontró en gran medida que los usuarios usan las redes sociales de las entidades públicas, como medio para dar a conocer sus peticiones, quejas y reclamos, sobre los servicios de salud que les son prestados a ellos o a su núcleo familiar.

El uso de clasificadores no supervisados es una herramienta muy útil para que las autoridades de salud midan e identifiquen qué tan efectivas son sus campañas, si realmente llegan a los usuarios y generan el objetivo deseado

10. Recomendaciones

La realización de limpieza de los datos, teniendo en cuenta la gramática utilizada por las personas que publican en esta red social.

Se recomienda realizar un estudio de los estimadores de los modelos, dado que con un Alpha más pequeño se habrá logrado que hubiesen sido más el porcentaje de tweets en el que tópico tienen una probabilidad mayor al 50% al muestrear el tweet.

Dado que el modelo solo realiza análisis de contenido, se supone inconveniente la gramática usada por las entidades y que el modelo no clasifica de acuerdo a las interacciones, esto crea una diferencia en la identificación real de tópicos, dado que los clasifica de acuerdo a la cantidad de palabras por tópico.

Referencias Bibliográficas

Aggarwal, C. C., ChengXiang, Z., Crain, S. P., Zhou, K., Zha, S.-H., & Yang, H. (2012). Mining Text Data. In Charu C. Aggarwal & ChengCiang Zhai (Eds.), *Mining Text Data* (p. 519). https://doi.org/10.1007/978-1-4614-3223-4

Atkins, B. T. S. (1992). TOOLS FOR COMPUTER-AIDED CORPUS LEXICOGRAPHY: THE HECTOR PROJECT. *Acta Linguistica Hungarica*, *41*, 5–71. https://doi.org/10.2307/44308286

Benítez Andrades, J. A., & Valbuena López, J. A. (2011). *Motores de Búsqueda Web: Estado del arte en Probabilistic Latent Semantic Analysis y en Latent Dirichlet Allocation aplicado a problemas de acceso a la información en la Web.* Retrieved from https://www.jabenitez.com/personal/MASTER/MOTORES_DE_BUSQUEDA_WEB/TAREAS/MBW-TareaExtraFinal-JoseAlbertoBenitezAndrades-y-JuanAntonioValbuenaLopez.pdf

Blanco, E.-J., & Sanz, H. (2016). *Algoritmos de clustering y aprendizaje automático aplicados a Twitter*. Retrieved from https://upcommons.upc.edu/bitstream/handle/2117/82434/113257.pdf?sequence=1&isAllowed=y

Blei, D. M., & Lafferty, J. D. (2005a). *Correlated Topic Models*. Retrieved from www.jstor.org

Blei, D. M., & Lafferty, J. D. (2005b). *Correlated Topic Models*. Retrieved from www.jstor.org

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation Michael I. Jordan. *Journal of Machine Learning Research*, *3*, 993–1022. Retrieved from

http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Centro Dermatologico Federico Lleras Acosta. (n.d.). Misión y Visión. Retrieved April 15, 2018, from http://www.dermatologia.gov.co/la_entidad/misia_n_visia_n

Chae, B. (2015). Insights from hashtag #supplychain and Twitter Analytics: Considering Twitter and Twitter data for supply chain practice and research. *ELSEVIER*, *165*, 247–259. https://doi.org/10.1016/j.ijpe.2014.12.037

Chandía Sepúlveda, B. (2016). *APLICACIÓN Y EVALUACIÓN LDA PARA ASIGNACIÓN DE TÓPICOS EN DATOS DE TWITTER*. Retrieved from http://opac.pucv.cl/pucv_txt/txt-5000/UCD5100_01.pdf

Chona, S., Londoño, M., & Gross Beltrán, L. (2013). *ESTRATEGIAS DIGITALES DE MERCADEO APLICADAS A TRAVÉS DE LAS REDES SOCIALES*. Universidad Colegio mayor de nuestra señora del rosario, Bogota. Retrieved from http://repository.urosario.edu.co/bitstream/handle/10336/4493/1018420316-2013.pdf?sequence=1

Congreso de la república. (1993). Ley 100 de 1993. Retrieved August 3, 2018, from http://www.secretariasenado.gov.co/senado/basedoc/ley_0100_1993.html

Coronoda Matutti, M. A., Cárdenas Acosta, R., Bello Medina, K., & Carrasco Rodríguez, J. L. (2015). *DISEÑO E IMPLEMENTACIÓN DE UN SISTEMA INTELIGENTE DE INFERENCIA DE PROGRAMAS DE ESPECIALIZACIÓN EN INGENIERÍA USANDO TOPIC MODELING*. Lima. Retrieved from https://docplayer.es/33340210-Universidad-nacional-de-ingenieria-facultad-de-ingenieria-mecanica-instituto-de-investigacion.html

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996, March 15). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 37–37. https://doi.org/10.1609/AIMAG.V17I3.1230

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software*, *25*(5), 1–54. Retrieved from http://www.jstatsoft.org/v25/i05/

Finfgeld-Connett, D. (2015). Twitter and Health Science Research. *Western Journal of Nursing Research*, *37*(10), 1269–1283. https://doi.org/10.1177/0193945914565056

Garcia Diéguez, A. (2014). *Técnicas de clustering aplicadas alanálisis de trending topics en conjunto de tweets*. UNIVERSIDAD CARLOS III DE MADRID. Retrieved from https://e-archivo.uc3m.es/bitstream/handle/10016/22233/PFC_adrian_garcia_dieguez_2014.pdf

Gonzales, C., Barber, F., Burgarin Diz, A., Cañadas, J., Corcho, O., Delgado, M., … Vila, A. (2008). *INTELIGENCIA ARTIFICIAL: Métodos,técnicas y aplicaciones.* (J. L. Garrcia Jurado & B. Pecharromá, Eds.) (McGRAW-HI). España: España. Retrieved from https://kupdf.com/download/inteligencia-artificial-m-eacute-todos-t-eacute-cnicas-y-aplicaciones_586feab36454a7362f35c098_pdf

Grun, B., & Hornik, K. (2011). topicmodels: An R Package for Fitting Topic Models. *Journal of Statistical Software*, *40*(13), 30.

Grün, B., & Hornik, K. (2011). **topicmodels** : An *R* Package for Fitting Topic Models. *Journal of Statistical Software*, *40*(13). https://doi.org/10.18637/jss.v040.i13

Guzmán, E. L. (n.d.). *Métricas para la validación de Clustering*. Retrieved from http://www.disi.unal.edu.co/profesores/eleonguz/cursos/md/presentaciones/Sesion13_valida

cion_Clustering.pdf

Hernández, C. P. (2002). Explotación de los córpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento. *Estudios de Lingüística Del Español*, *18*(1139–8736), 1. Retrieved from https://ddd.uab.cat/record/53652

Hernández Orallo, J., Ramírez Quintanilla, M. J., & Ferri Ramírez, C. (2004). *Introducción a la Minería de Datos*. Retrieved from http://users.dsic.upv.es/~flip/LibroMD/

Hu, X., & Liu, H. (n.d.). TEXT ANALYTICS IN SOCIAL MEDIA. Retrieved from https://pdfs.semanticscholar.org/814a/ba8f1553266531138c29a9bc5830da1ad807.pdf

Imai, K. (2017). Quantitative Social Science: An Introduction. In *Quantitative Social Science* (p. 397). Retrieved from https://books.google.com.co/books?id=KLwODgAAQBAJ&pg=PA194&dq=Document-term+matrix&hl=es&sa=X&ved=0ahUKEwiO9OnR3pXeAhWCtlkKHepPAXYQ6AEIMTAB#v=onepage&q=Document-term matrix&f=false

Instituto Nacional de Cancerología. (2018). Mision Vision Valores Principios | Instituto Nacional de Cancerologia. Retrieved April 10, 2018, from http://www.cancer.gov.co/mision-vision-valores-principios

Kagashe, I., Yan, Z., & Suheryani, I. (2017). Enhancing seasonal influenza surveillance: Topic analysis of widely used medicinal drugs using twitter data. *Journal of Medical Internet Research*, *19*(9), e315. https://doi.org/10.2196/jmir.7393

Kalyanam, J., Katsuki, T., Lanckriet, G. R. G., & Mackey, T. K. (2017). Exploring trends of nonmedical use of prescription drugs and polydrug abuse in the Twittersphere using

unsupervised machine learning. *Addictive Behaviors*, *65*, 289–295. https://doi.org/10.1016/J.ADDBEH.2016.08.019

Kaufmann, morgan. (1990). *Readings in machine learning*. Morgan Kaufmann Publishers.

Kearney, M. W. (2018). rtweet: Collecting Twitter Data. Retrieved from https://cran.r-project.org/package=rtweet

León Boyajian, W., & Lamberti, P. W. (2011). *Una propuesta de distancia entre procesos cuánticos*. Retrieved from https://rdu.unc.edu.ar/bitstream/handle/11086/61/15867.pdf?sequence=1

Massey, P. M., Leader, A., Yom-Tov, E., Budenz, A., Fisher, K., & Klassen, A. C. (2016). Applying Multiple Data Collection Tools to Quantify Human Papillomavirus Vaccine Communication on Twitter. *Journal of Medical Internet Research*, *18*(12), e318. https://doi.org/10.2196/jmir.6670

Mesas Jávega, R. M. (2015). *ANALISIS DE TENDENCIAS Y MARCAS DEPORTIVAS ATRAVES DE TWITTER*. Universidad Autónoma de Madrid. Retrieved from https://repositorio.uam.es/bitstream/handle/10486/669057/Mesas_Javega_RusMaria_tfg.pdf?sequence=1

Ministerio de salud. (n.d.). ministerio de salud y proteccion social. Retrieved March 18, 2018, from https://www.minsalud.gov.co/Ministerio/Institucional/Paginas/mision-vision-principios.aspx

Ministerio de Tecnologías de la Información y las Comunicaciones. (2018). DATOS ABIERTOS GOBIERNO DIGITAL DE COLOMBIA.

Novillo Ortiz, D., & Hernández Pérez, A. (2015). *Acceso a información y uso de redes sociales en salud pública: un análisis de las autoridades nacionales de salud y de las causas principales de defunción en Latinoamérica"* Universidad Carlos III de Madrid. Retrieved from https://e-archivo.uc3m.es/bitstream/handle/10016/22158/Tesis_David_Novillo_Ortiz_2015.pdf

Odlum, M., & Yoon, S. (2015). What can we learn about the Ebola outbreak from tweets? *American Journal of Infection Control*, *43*(6), 563–571. https://doi.org/10.1016/J.AJIC.2015.02.023

Organización mundial de la salud. (2016). OMS | ¿Qué es la promoción de la salud? Retrieved January 22, 2018, from http://www.who.int/features/qa/health-promotion/es/

Organización Mundial de la Salud. (1998). Promoción de la salud. Glosario. ©*World Health Organization*, 36. Retrieved from https://www.mscbs.gob.es/profesionales/saludPublica/prevPromocion/docs/glosario.pdf

Organizacion mundial de la salud, O. (2005). *58ª ASAMBLEA MUNDIAL DE LA SALUD*. Ginebra. Retrieved from http://apps.who.int/gb/ebwha/pdf_files/WHA58-REC1/A58_2005_REC1-sp.pdf

Ospina, M. L. (2018). *Carta de trato digno al ciudadano*. Retrieved from https://www.ins.gov.co/AtencionAlCiudadano/Documents/CARTA TRATO DIGNO AL CIUDADANO 2018.pdf

Parks, R. G., Tabak, R. G., Allen, P., Baker, E. A., Stamatakis, K. A., Poehler, A. R., … Brownson, R. C. (2017). Enhancing evidence-based diabetes and chronic disease control among local health departments: a multi-phase dissemination study with a stepped-wedge cluster

randomized trial component. *Implementation Science*, *12*(1), 122. https://doi.org/10.1186/s13012-017-0650-4

R Core Team. (2018). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from https://www.r-project.org/

Romero Salamanca, G. (2018). 12 twitts en los doce años de twitter. Retrieved June 21, 2018, from http://www.eje21.com.co/2018/03/12-twitts-en-los-doce-anos-de-twitter/

Rus, V., Niraula, N., & Banjade, R. (2013). *Similarity Measures based on Latent Dirichlet Allocation*. Retrieved from http://deeptutor.memphis.edu/publications/2013/CICLing-2013.RusNiraulaBanjade.pdf

Sanabria Ruiz, V. A. (2017). Aplicación de técnicas de agrupamiento (clustering) para el análisis estadístico de tendencias en twitter basado en el lenguaje programación R. Retrieved from http://tangara.uis.edu.co/biblioweb/tesis/2017/168596.pdf

SCHIATTI SISÓ, L. C. (2017). *ESTUDIO COMPARATIVO DE DIFERENTES ALGORITMOS DE CLUSTERING PARA LA ESTIMACIÓN DE GRUPOS DE EVALUADOS QUE COMPARTEN DEBILIDADES CONCEPTUALES SIMILARES*. Universidad Tecnologica de Bolívar. Retrieved from http://biblioteca.unitecnologica.edu.co/notas/tesis/0069812.pdf

Silvestre Gómez, M. (2018). *Implementación de Asignación Jerárquica Latente de Dirichlet para Modelado de Temas*. Universidad de Sevilla. Retrieved from http://bibing.us.es/proyectos/abreproy/71106/fichero/TFM-1106-SILVESTRE.pdf

Supersalud. (2017). Espacios de Participación Ciudadana. Retrieved February 11, 2019, from https://www.supersalud.gov.co/es-co/delegadas/proteccion-al-usuario/direccion-de-

participacion-ciudadana

Tomeny, T. S., Vargo, C. J., & El-Toukhy, S. (2017). Geographic and demographic correlates of autism-related anti-vaccine beliefs on Twitter, 2009-15. *Social Science & Medicine*, *191*, 168–175. https://doi.org/10.1016/J.SOCSCIMED.2017.08.041

Torres Montero, A. V., & Cabrera Chaves, A. R. (2008). *LA COMUNICACIÓN EN LOS PROGRAMAS DE PROMOCIÓN Y PREVENCIÓN DE LA SALUD EN BOGOTÁ: SECRETARÍA DISTRITAL DE SALUD, EPS Y MEDIOS DE COMUNICACIÓN*. BOGOTA. Retrieved from http://www.javeriana.edu.co/biblos/tesis/comunicacion/tesis39.pdf

Trstenjak, B., Mikac, S., & Donko, D. (2014). KNN with TF-IDF based Framework for Text Categorization. *Procedia Engineering*, *69*, 1356–1364. https://doi.org/10.1016/J.PROENG.2014.03.129

Twitter INC. (2018a). Enterprise APIs — Twitter Developers. Retrieved July 29, 2018, from https://developer.twitter.com/en/docs/changelog/enterprise

Twitter INC. (2018b). Pricing — Twitter Developers. Retrieved July 29, 2018, from https://developer.twitter.com/en/pricing.html

Wang, S., Paul, M. J., & Dredze, M. (2015). Social media as a sensor of air quality and public response in China. *Journal of Medical Internet Research*, *17*(3), e22. https://doi.org/10.2196/jmir.3875

Witten, I. H. (Ian H. ., & Frank, E. (2000). *Data mining : practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann. Retrieved from https://www.researchgate.net/publication/30876208_Data_Mining_-

_Practical_Machine_Learning_Tools_and_Techniques_with_JAVA_Implementations