Análisis de la percepción de servicio en el sector de servicios públicos en la ciudad de Bogotá a través de técnicas de minería de texto aplicadas a Twitter

Carlos Eduardo Gómez Rivera

Juan Guillermo Jaramillo Yepes

Escuela Colombiana de Ingeniería Julio Garavito

Decanatura de Ingeniería de Sistemas

Maestría Gestión de Información

Bogotá D.C., Colombia

29 de noviembre de 2022

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO

Análisis de la percepción de servicio en el sector de servicios públicos en la ciudad de Bogotá a través de técnicas de minería de texto aplicadas a Twitter

Carlos Eduardo Gómez Rivera

Juan Guillermo Jaramillo Yepes

Trabajo de investigación para optar al título de

Magíster en Gestión de Información

Director

Ph.D Victoria Eugenia Ospina Becerra

Ph.D Dante Conti

Jurados

Daniela de la Rosa

Diana Patricia Rincón Velásquez

Escuela Colombiana de Ingeniería Julio Garavito

Decanatura de Ingeniería de Sistemas

Maestría en Gestión de Información

Bogotá D.C., Colombia

29 de noviembre de 2022

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO

Publicado en 2022 por la Escuela Colombiana de Ingeniería Julio Garavito. Avenida 13 No 205-59 Bogotá. Colombia
TEL: +57 – 1 668 36 00

Página de aceptación del jurado

El trabajo de grado de maestría titulado **"Análisis de la percepción de servicio en el sector de servicios públicos en la ciudad de Bogotá a través de técnicas de minería de texto aplicadas a Twitter"**, presentado por Carlos Eduardo Gómez Rivera y Juan Guillermo Jaramillo Yepes, cumple con los requisitos establecidos para optar al título de Magíster en Gestión de información.

_____

Victoria Eugenia Ospina Becerra
Director del Trabajo de Grado

_____

Dante Conti
Director del Trabajo de Grado

_____

Daniela de la Rosa
Jurado

_____

Diana Patricia Rincón Velásquez
Jurado

Bogotá, D.C., 29 de noviembre de 2022

Agradecimientos

Al ver el resultado de este trabajo de grado, solamente se me ocurre una palabra: Gracias. En primera medida a Dios por darme la oportunidad de recorrer este camino en compañía de grandes tutores: Dante Conti y Victoria Ospina, y un excelente profesional, compañero y amigo: Carlos Eduardo Gómez Rivera. Gracias a la Escuela Colombiana de ingeniería Julio Garavito, pues nuevamente juega un papel fundamental en una basta e importante etapa de mi vida. A mis padres: Diana María Yepes Medina y Guillermo Adolfo Jaramillo Gallego por su amor, paciencia, apoyo y palabras que han sido y seguirán siendo una gran guía y motivación. A mis abuelos: Francisco Jaramillo Ramírez (Q.E.P.D) y Oscar de Jesús Yepes Roldan (Q.E.P.D) dos caballeros y maestros que, con sus grandes enseñanzas inspiraron a más de una persona a ser mejores en su día a día. Finalmente, a mis amigos y familia que siempre estuvieron presentes aun en la distancia.

Quiero agradecer a Dios por la oportunidad brindada y por poner en mi camino a todas las personas que hicieron posible la culminación de este trabajo de grado; especialmente a nuestro director Dante Conti y Victoria Eugenia Ospina Becerra por guiar, compartir e instruir con todo su conocimiento y ayuda la culminación de este, además de agradecer a Juan Guillermo Jaramillo Yepes, por ser un excelente compañero que motivó y que gracias a su entrega el resultado de este trabajo es muy satisfactorio y a mi amada Escuela Colombiana de Ingeniería Julio Garavito, por ser la institución que me formó de manera profesional, integral y ética; y que me abrió nuevamente sus puertas para alcanzar este logro. Un agradecimiento especial a mis padres Stella Rivera y Juan Carlos Gomez; que gracias a sus palabras, motivación, comprensión y apoyo me dieron el ánimo y las ganas de continuar con este paso más en mi vida; a mi hermano Juan Sebastián Gomez Rivera por sus palabras y deseos de ser su ejemplo de vida y a mi novia Diana Carolina Arias, por ser mi apoyo profesional e incondicional y que con su ayuda fue el pilar para superar todos los obstáculos presentados. Finalmente, extender mis agradecimientos a mis compañeros, docentes y demás amigos que aportaron desde su perspectiva a lograr todo este resultado.

"*Los verdaderos viajes empiezan cuando se acaban los caminos*"

Jacques Lacan

Resumen

En Bogotá, las empresas prestadoras de los servicios públicos esenciales (Energía, Gas natural y Acueducto) son: Enel, Vanti y El Acueducto y Alcantarillado, las cuales cuentan con más de 1 millón de clientes cada una, pudiéndose catalogar como grandes empresas del sector; en donde el servicio al cliente se debe basar en una estrategia *customer centricity,* caracterizada por tener la VOC (voz del cliente) como principal insumo para orientar sus decisiones tácticas y operativas. Sin embargo, el top de las empresas con mayor cantidad de peticiones a nivel nacional según la (Super Intendencia de Servicios Publicos Domiciliarios, 2021) ubica en los primeros lugares a las tres empresas mencionadas anteriormente, lo que hace pensar que su percepción del cliente puede no ser la mejor. Con el fin de aportar a la solución de esta problemática, nace la presente investigación, teniendo como objetivo analizar los comentarios de los usuarios para identificar posibles áreas con oportunidades de mejora, mediante el aprovechamiento de Twitter; red social que, según los informes de gestión más recientes de las 3 compañías, ha sido adoptado por parte de los usuarios como canal de comunicación y por tanto un crecimiento relevante en el último año. Lo anterior, potencializado por las medidas de emergencia sanitaria decretadas por el gobierno, enfocadas en el fortalecimiento en canales digitales para todos los prestadores de servicios públicos, lo que se confirma en las cifras de la (Superintendencia de Servicios Públicos Domiciliarios, 2020) los contactos presenciales tuvieron un decrecimiento de cerca del 70%, aumentando la participación de canales virtuales.

Con el fin de encontrar áreas con mayores oportunidades de mejora y su comportamiento en el tiempo utilizando la voz del cliente como principal insumo, se propone realizar un análisis de sentimientos que permita conocer la percepción de los usuarios, además de aplicar técnicas como *Topic Modelling*, para plantear un marco de evaluación y generar interfaces de visualización de los datos que faciliten el análisis de esta información.

De esta manera, la presente investigación tiene como finalidad ofrecer un modelo de alta predictibilidad que plantee un punto de partida para este tipo de análisis, permitiendo conocer la situación actual de la percepción de servicio a través de los datos de las redes sociales para así lograr plantear métricas que aporten en la formulación de estrategias que garanticen una comunicación más eficiente con los usuarios, atendiendo a las necesidades transmitidas y transformar la percepción de manera positiva.

Abstract

In Bogotá, the companies that provide essential public services (Energy, Natural Gas and Aqueduct) are: Enel, Vanti and Acueducto y Alcantarillado, which have more than 1 million clients each, and can be classified as large companies in the sector, where customer service must be based on a customer centricity strategy, characterized by having the VOC (voice of the customer) as the main input to guide their tactical and operational decisions. However, according to the **(Super Intendencia de Servicios Publicos Domiciliarios, 2021)** the top of the national companies with the largest number of requests, places the three companies mentioned above in the first places, which suggests that their perception of the service could not be the best.

In order to contribute to the solution of this problem, this research was born, with the objective of analyzing user comments to identify possible areas with opportunities for improvement, by taking advantage of Twitter; social network that, according to the most recent management reports of the 3 companies, has been adopted by users as a communication channel and therefore has shown significant growth in the last year. All of the above, enhanced by the health emergency measures decreed by the Colombian government, focused on strengthening digital channels for all public service providers, which is confirmed in the figures of the **(Superintendencia de Servicios Públicos Domiciliarios, 2020)** : face-to-face contacts had a decrease of about 70%, increasing the participation of virtual channels.

For the purpose of finding areas with greater opportunities for improvement and their behavior over time using the voice of the customer as the main input, it is proposed to carry out a sentiment analysis that allows knowing the perception of users, in addition to applying techniques such as Topic Modeling, to propose a framework of evaluation and generate data visualization interfaces that facilitate the analysis of this information.

In this way, the present research aims to offer a highly predictable model that proposes a starting point for this type of analysis, allowing to know the current situation of the perception of service through data from social networks to propose metrics that contribute to the formulation of strategies that guarantee more efficient communication with users, attending to the needs transmitted and transforming perception in a positive way.

Índice General

Índice de Figuras

Índice de Tablas

Índice de Ecuaciones

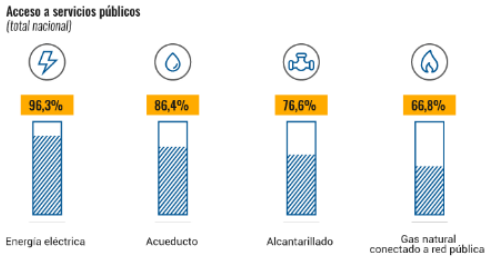
Anexos

1. Introducción

Para cualquier país la prestación de los servicios públicos es vital para apoyar y mejorar la calidad de vida de sus ciudadanos, ya que permiten la satisfacción de las necesidades básicas, como por ejemplo el acceso a agua potable. Según el artículo 430 del Código Sustantivo del trabajo, lo servicios públicos pueden definirse precisamente, como un conjunto de actividades organizadas y centradas en satisfacer las de necesidades generales de la población.

De acuerdo con el (DANE, 2018) y los resultados obtenidos en la aplicación del censo nacional de población y vivienda, se obtuvieron las siguientes cifras respecto a la cobertura de los servicios públicos domiciliarios básicos para los colombianos:

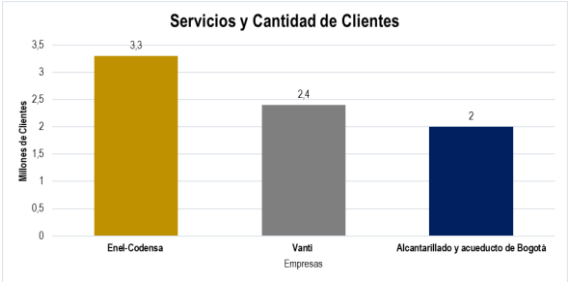
Figura 1. Acceso a servicios públicos domiciliarios en el territorio colombiano



Fuente: (DANE, 2018)

Teniendo en cuenta lo anterior es importante mencionar que, en Colombia más precisamente en Bogotá, la prestación de estos tres servicios esenciales está a cargo de las empresas que se observan en la Figura **2**, donde aproximadamente el 91.67% de la población cuenta con la cobertura y suministro de estos servicios:

Figura 2. Cantidad de clientes titulares que figuran en cada empresa de servicios públicos de Bogotá

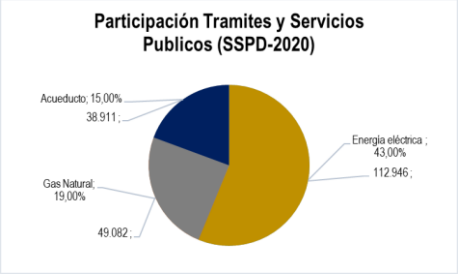


Fuente: (Enel - Codensa, 2021), (Vanti, 2020) y (Empresa de Acueducto y Alcantarillado de Bogotá, 2021). Autoría Propia.

Con estas cifras, se identifican estas compañías, como grandes empresas del sector que requieren de una gran infraestructura técnica, no solo para cumplir con su misión sino también para atender de una manera adecuada y oportuna a tal número de usuarios, es por ello que la estrategia de cada una debe contemplar el "servicio al cliente", en donde la atención que se le brinde al usuario debe estar basada en lo que él piensa, opina y siente, es decir una estrategia *customer centricity*.

Por otro lado, este modelo debe estar enmarcado en acciones reactivas que permitan ante ponerse a situaciones que disgusten a los usuarios, capitalizando los momentos de verdad en experiencias gratas que generen sentimientos de empatía y afinidad con el cliente. Sin embargo, según (Super Intendencia de Servicios Publicos Domiciliarios, 2021) en su comunicado número 01, publicado el 5 de enero del 2021, se presentan las siguientes cifras:

Figura 3. Trámites y servicios atendidos por la Superintendencia de Servicios públicos Domiciliarios en el año 2021 (SSPD)



Fuente: (Super Intendencia de Servicios Publicos Domiciliarios, 2021). Autoría Propia.

Esto lleva a pensar que las empresas que mayor cantidad de solicitudes (entendidas como: quejas, reclamos, derechos de petición, etc.) radicadas ante la Super Intendencia de Servicios Públicos, son las prestadoras de los servicios básicos en la ciudad de Bogotá - Acueducto y Alcantarillado, Enel, Vanti - lo cual es confirmado en el mismo comunicado de la (Super Intendencia de Servicios Publicos Domiciliarios, 2021) a partir del cual se adaptó la Tabla 1, en donde estas cuatro compañías se ubican en los primeros lugares del Top 10, de las empresas con mayor número de trámites y solicitudes de atención presentadas ante la SSPD.

Tabla 1. Top 10 Empresas con mayor cantidad de trámites y solicitudes nacionales presentadas ante la SSPD en 2020

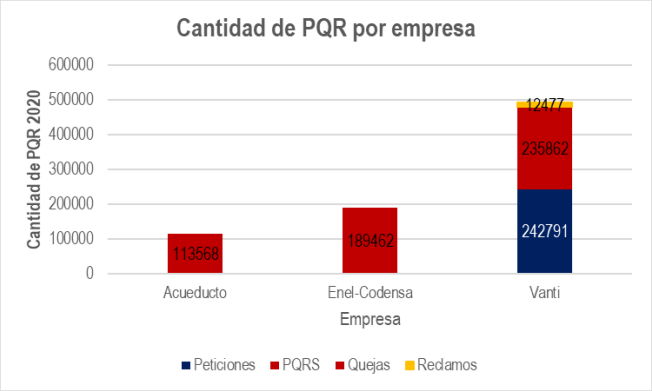
Empresa / prestador	Número de trámites y solicitudes de atención personalizada	%
Enel	44.250	17,02
Electrificadora del Caribe (ene-sep-2020)	39.017	15,01
Vanti	31.740	12,21
Empresa de Acueducto y Alcantarillado de Bogotá	15.716	6,04

Fuente: Autoría Propia.

Esta información, permite preguntarse si realmente la estructura del servicio al cliente está orientada y alineada con el sentir de los consumidores y usuarios, ya que se hace evidente que dejando de lado a la compañía "Electrificadora del Caribe", el 35% de los trámites a nivel nacional se asocian con las 3 empresas en cuestión.

Lo anterior, es secundado por las cifras publicadas por las empresas prestadoras de servicios públicos de la siguiente manera:

Figura 4. Cantidad de PQR atendidos por empresa de servicios públicos domiciliarios durante el año 2020



Fuente: (Empresa de Acueducto y alcantarillado de Bogotá, 2021, pág. 42) (Enel, 2020, pág. 44) y (Vanti, 2020, pág. 44). Autoría Propia.

Se hace entonces evidente que la percepción del servicio público puede no ser la mejor, aun cuando a más 10 mil usuarios que reportaron se les dejo de atender (Super Intendencia de Servicios Publicos Domiciliarios, 2021) y que es imperativo conocer lo que los clientes piensan y opinan a través de la voz del cliente (VOC), para identificar oportunidades de mejora y generar acciones que estén acordes a las expectativas y deseos de los usuarios, pues atender sus solicitudes a tiempo y con la calidad necesaria, influye en la prestación de un servicio público básico adecuado.

Otro dato relevante publicado por la (Super Intendencia de Servicios Publicos Domiciliarios, 2021) en el comunicado 01 de 2021, se relaciona con el cambio generado por la pandemia en las modalidades de atención al cliente: “Por su parte, las solicitudes de atención personalizada sumaron 11.762, un 72% menos que las 41.919 recibidas en 2019; reducción explicable por las condiciones de aislamiento preventivo que vivió el país a partir del pasado mes de marzo.”

Lo que se hizo también necesario dado el decreto 385 del 12 de marzo de 2020, en donde se declaró la emergencia sanitara por parte del Ministerio de Salud y Protección Social, lo que género que la misma (Superintendencia de Servicios Públicos Domiciliarios, 2020), solicitara a los prestadores de servicios públicos el fortalecimiento, desarrollo, creación y divulgación amplia sobre los canales de atención no presenciales (virtuales y digitales) para garantizar los protocolos de bioseguridad.

Las cifras anteriores, dan indicios sobre un cambio de paradigma y es que, en los canales de atención y los modelos de servicio al cliente, se presentó una transformación radical en sus estructuras, viéndose en la necesidad de entrar en el mundo digital de una manera acelerada aumentando la participación de los canales virtuales en términos de utilización por parte de los usuarios. Dentro de los canales virtuales de las empresas de servicios públicos se destaca Twitter, que como se expone en la Tabla 2, las interacciones a través de este medio se incrementaron a lo largo del 2020.

Tabla 2. Hitos Canales Digitales empresas de servicios públicos de Bogotá año 2020

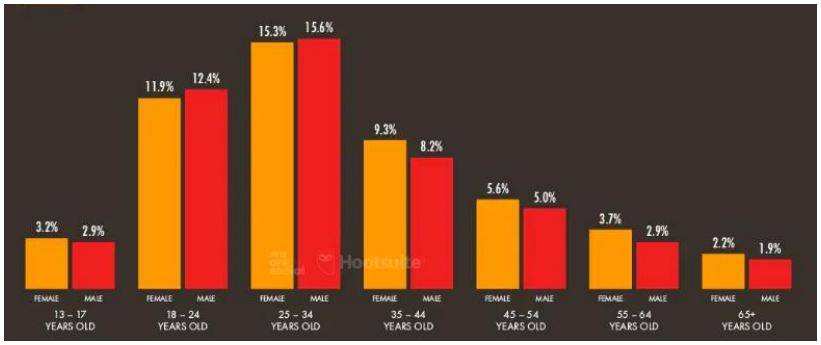
Empresa	Hito
Enel – Codensa	Según el informe de sostenibilidad presentado por (Enel, 2020) el canal con más crecimiento fue Twitter; teniendo 1.447 publicaciones, 43.799 seguidores y 121.274 comentarios, los cuales crecieron 147% más que lo reportado en el cierre de 2019.
Vanti	En el informe de Sostenibilidad presentado por (Vanti, 2020) se evidencia un aumento en los canales digitales cerrando el año con 53.994 solicitudes. El canal Twitter llego a 9.044 seguidores.
Empresa de Acueducto de Bogotá	*Sin información reportada en el informe de sostenibilidad.

Fuente: Autoría Propia.

Esta red social, ha tenido una especial acogida durante la última década en cuanto al uso que se le da por parte de los consumidores y clientes de productos y servicios, ya que se basa en la premisa de permitir a los usuarios compartir sus pensamientos, opiniones y sentimientos mediante un máximo de 280 caracteres (*microbloging*), abriendo paso a la integración de la vida diaria de las personas con la tecnología y cambiando hábitos, decisiones, procesos de compra y expectativas de las personas que la usan.

Como lo indica el estudio denominado *Digital 2021 Global Overview Report*, (WeAreSocial, 2021) sobre el uso del internet y las redes sociales a nivel mundial; hace una corta, pero completa reseña sobre el panorama de Colombia en términos del uso del internet y comportamientos de consumo de contenido digital a través de redes sociales con corte a enero de 2021. Este estudio concluye que alrededor del 68% de la población del país tiene acceso a internet, y de este porcentaje el 77% aproximadamente utiliza por lo menos alguna red social de manera activa. En Colombia, se ha evidenciado que la distribución de usuarios de redes sociales se da de la siguiente manera:

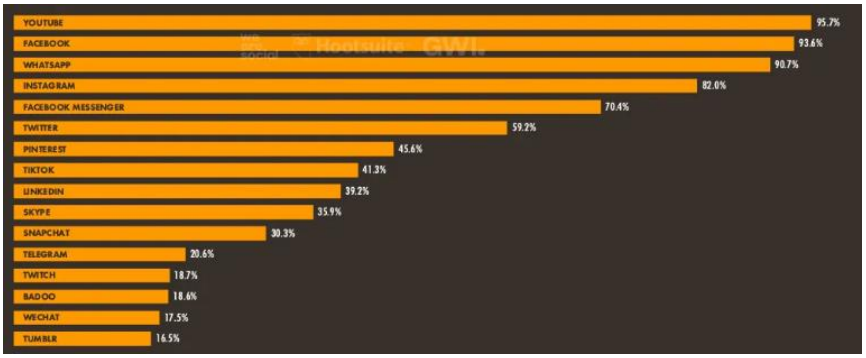
Figura 5. Distribución de usuarios de redes sociales en Colombia para el año 2021, discriminado por rango de edad y género



Fuente: (WeAreSocial, 2021)

Por otro lado, los usuarios que utilizan estas plataformas de manera recurrente tienen en promedio 10 perfiles en diferentes plataformas o redes sociales y estos equivalen al 90% de la población encuestada y en términos de usabilidad, se destacan las siguientes redes sociales con su participación en Colombia:

Figura 6. Porcentaje de usuarios de redes sociales con edad entre 16 y 64 años que usaron cada una en enero de 2021



Fuente: (WeAreSocial, 2021).

Actualmente en Colombia, existen en promedio alrededor de 3.35 millones de usuarios activos en la plataforma y se generan aproximadamente 5.2 millones de tuits al día (tomado a partir de la participación de usuarios en Twitter y cantidad de tuits por segundo generados a nivel mundial estimados en 6000[1]), por lo que se puede pensar que la utilización de las redes sociales, en especial Twitter, se transforma en una necesidad más que en una ventaja competitiva para adaptarse a un entorno variable y dinámico, en especial cuando la información sobre lo que los usuarios expresan se encuentra disponible de manera gratuita y fácilmente accesible.

A este punto, se hace evidente que entender la percepción de los clientes de las empresas de servicio público es de vital importancia aún más cuando: se ubican entre las que más peticiones tienen ante el ente regulador SSPD, la percepción sobre el costo de estos aumentó durante el año 2020 debido al teletrabajo (Bogotá Como vamos, 2021), además en este mismo informe se registró que en el último año se presentó una satisfacción por debajo del 70%.

Para llegar a tal fin, se propone realizar una evaluación y análisis de la percepción del servicio al cliente (considerando todos los comentarios realizados exclusivamente por clientes en una ventana de tiempo definida) involucrando la red social Twitter como canal de atención a los usuarios (debido a su crecimiento en el último año como previamente se expuso), utilizando técnicas de minería de texto para limpiar, estructurar y estandarizar los datos derivados de los mensajes y así aplicar análisis estadísticos y de sentimientos que permitan obtener información, indicadores y visualizaciones para las empresa del Acueducto, Enel y Vanti; que en conjunto consolidaran un estudio robusto del uso actual de esta red social como canal de atención empresarial, resaltando el aprovechamiento de esta fuente de información en la implementación de soluciones tecnológicas para la toma de decisiones, identificación de mejoras e implementación de estrategias que puedan ser de utilidad para acercar a los clientes a la organización y generar un impacto positivo producto de lo evidenciado.

2. Objetivos

2.1 Objetivo General

Analizar la percepción del servicio en las empresas de servicios públicos de la ciudad de Bogotá: Enel, Acueducto y Vanti; para identificar potenciales áreas con oportunidades de mejora, mediante la aplicación de técnicas de minería de texto y evolución temporal a los Tuits de los usuarios.

[1] Tomado de https://www.internetlivestats.com/twitter-statistics/

2.2 Objetivos específicos

- Diseñar un modelo de procesamiento y clasificación de texto a partir de los datos extraídos de las páginas oficiales de cada entidad para tener un *dataset* de referencia.
- Evaluar mediante técnicas de análisis de sentimientos las emociones de los usuarios para cuantificar y cualificar la percepción del servicio prestado por cada empresa.
- Definir un sistema de evaluación que permitan establecer métricas para valorar la percepción de los servicios prestados por las compañías, a partir de la utilización de los datos obtenidos mediante el modelo de procesamiento y clasificación de texto.
- Presentar un modelo de visualización de datos a partir de los indicadores evaluados para identificar *insights*, correlaciones, comparaciones y tendencias relacionados al estado actual de las organizaciones.

3. Marco teórico o estado del arte

Los grandes cambios basados en la disponibilidad de la información han repercutido en la vida cotidiana de las personas e industrias, la voz de cliente se ha convertido en uno de los insumos más valiosos, que permite tomar en consideración la percepción y sentimiento de los usuarios; causado por la proliferación de las plataformas basadas en opinión (como por ejemplo redes sociales) que han generado que los comentarios y reseñas sean un fenómeno emergente y cada vez más consolidado, en donde las decisiones de compra del consumidor y la expectativa del servicio pos venta, se han visto influenciadas por la cantidad de información que se puede conseguir fácilmente (Zhiwei & Park, 2015).

Durante la última década, se ha presentado un especial interés en estudiar diversas maneras sobre cómo puede explotarse dicha información de las redes sociales, en especial Twitter; ya que la transformación digital ha llevado a las empresas a convertir estos medios virtuales en canales de atención al cliente, creando espacios en donde el contenido generado (micro reseñas o micro blogs) nace a partir de experiencias, pensamientos y sentimientos relacionados a los individuos (Salaberry, 2020). Twitter es una empresa que ofrece la posibilidad de acceder a sus datos de manera gratuita: “Para compartir información en Twitter de la forma más amplia posible, también les proporcionamos a las empresas, los desarrolladores y los usuarios acceso programático a los datos de Twitter mediante nuestras API (interfaces de programación de aplicaciones)” (Twitter, S.F).

Uno de los principales métodos que ha sido ampliamente estudiado para aprovechar estos datos es la minería de texto o *Data mining*, la cual es un conjunto de técnicas que han sido aplicadas en los últimos años en diversos temas, especialmente, en el análisis de texto lo que incluye reseñas online de redes sociales. (Kuo,

Riantama, & Chen, 2020), dentro de la minería de texto, se recalca el análisis de sentimientos (*SA*) u *opinion mining*, las cuales se basan en extraer la polaridad (positiva, negativa o neutra) a partir de un texto dado.

En la literatura revisada se encuentran una gran variedad de estudios, algunos de los cuales se centran en generar un *frame work* de referencia para futuros trabajos; este tipo de investigaciones se considera de alta relevancia, debido a que permiten tener una base no solo para proponer nuevas ideas para analizar los comentarios, sino también realizar búsquedas más exhaustivas que permitan tener un entendimiento más profundo del tema. Un ejemplo de lo anterior se evidencia en el artículo "*The Impact of Sentiment Analysis on Social Media to Assess Customer Satisfaction: Case of Rwanda*" (Ngaboyamahina & Sun, 2019), en donde se presenta un modelo que contiene 4 etapas básicas: Recolección de datos, Gestión de datos, Extracción y clasificación, visualización de resultados, que permiten recoger los datos de redes sociales para aplicar análisis de sentimientos y contar con ciertos elementos de visualización sobre la percepción de los servicios en el sector público y privado de Ruanda.

En Colombia, se han realizado algunas investigaciones que utilizan Twitter como fuente de datos para realizar análisis de ciertos sectores, por ejemplo (Ávila Rodríguez, 2020) propone un modelo basado en redes neuronales y aprendizaje automático que permite entender y personalizar las tarifas asociadas a los seguros de vida. En este estudio, se recalca la utilización de una gran variedad de herramientas empleadas en las distintas etapas de la metodología de la investigación, por ejemplo: *Tweet Archiver*, que permite realizar una conexión entre las hojas de Excel de Google Drive con Twitter, *OpenRefine*, software que realiza la limpieza de los datos de Twitter sin necesidad de utilizar códigos de programación y *Biome-textSe* mediante el cual se carga de la data ya procesada y permite realizar los clúster de los Tuits basados en: características particulares del negocio de seguros, utilizando las métricas asociadas a la matriz de confusión (precisión, *recall* y F1). Por último, es importante mencionar que la visualización está también presente en este estudio, al comparar el desempeño del modelo en términos de cada categoría de clasificación.

Este tipo de investigaciones permite sentar una línea base para ahondar en los avances que se han desarrollado en la utilización de la minería de texto y el análisis de sentimientos, enfocados en aprovechar la voz del cliente (VOC) expresada en medios digitales. Como se mencionó previamente, este análisis puede ser aplicado en las diferentes etapas del proceso de compra del consumidor; desde el sector del turismo en posventa, donde (Songpan, 2017) realiza un análisis de la concordancia entre lo escrito en una reseña de hotel con la calificación brindada por los cliente para así generar y comparar un par de algoritmos basados en árboles de decisión y el clasificador *naive bayes*, permitiendo predecir y explicar la valoración del cliente a partir de lo comentado, obteniendo una precisión del 94.37% de su modelo propuesto.

Por otro lado, (Bello-Orgaz, Menéndez, Okazaki, & Camacho, 2014), aplican una serie de combinaciones entre técnicas de clasificación de textos y *clustering* de datos, cuantificándolas mediante una serie de métricas que

permiten evaluar el rendimiento de cada combinación para ser aplicada a una muestra de 100 Tuits sobre las opiniones de los clientes acerca de IKEA; los resultados muestran un mejor rendimiento para clasificar textos con *C4.5 Tree,* clusterización con *Dirichlet Process Algorithm* medido con la métrica *Tanimoto Coefficient Distance*, mostrando una mejor personalización al combinarlos respecto a la aplicación de técnicas simples.

En el sector de las comidas rápidas, (Kuo, Riantama, & Chen, 2020) concluyen que los principales atributos a los que los clientes hacen alusión en las reseñas son: la calidad de la comida, servicio al cliente, la agilidad y tiempos de espera; donde utilizaron el método *TD-IDF* para saber cuáles son las palabras que más se repiten, seguido por el algoritmo *LASSO,* para categorizar y clusterizar *top words*, adicionalmente de realizar una comparación cronológica pre y pos-pandemia.

Los autores (Kuffo, Vaca, Izquierdo, & Bustamante, 2018) cómo lo indican en su artículo, realizan una investigación sobre la experiencia posventa a partir de los comentarios generados en redes sociales como Facebook, Twitter y *FourSquare*, extrayendo una muestra robusta sobre 90 empresas rankeadas en *ForeSee CX Index Ranking* y *Forrester CX Index Ranking.* Con esto, pretenden aplicar diferentes técnicas de clasificación de textos utilizando diferentes combinaciones de unigramas, bigramas, trigramas y tetragramas para así encontrar cual tiene un mejor rendimiento con los datos extraídos para cada red social y observar cómo cada uno de estos se ajusta de mejor manera al cambiar el tamaño de la muestra analizada. Entre sus resultados, resaltan el modelo *Random Forrest* al alcanzar una precisión máxima para los datos extraídos de Facebook con una muestra de 14.000 cometarios y SVM al lograr una precisión máxima para los datos obtenidos de Twitter con una muestra 4.000 comentarios; además de indicar con cuales N-Gramas tienen un mejor rendimiento de acuerdo con la red social estudiada.

La percepción del cliente es otro factor común que se ha identificado en los trabajos previos, y es que este concepto es la base para incorporar la VOC en las decisiones ya sea de una organización o de una figura pública; como por ejemplo, (Ba & Lee , 2012) en su artículo evalúan la percepción de los Tuits de 13 figuras públicas: *"Barack Obama, Donald J. Trump, Bill Gates, Oprah Winfrey, Larry King, Lady Gaga, Ashton Kutcher, Britney Spears, Dalai Lama, TechCrunch, Mashable, CNN Breaking News, BBC Breaking News"* escogidas a partir de la cantidad de seguidores que tienen, donde el objetivo es utilizar el análisis de sentimientos para identificar la influencia positiva o negativa de los Tuits de estas figuras, su variación en el tiempo y correlación con situaciones del mundo "real". Se utiliza una API para obtener todos los Tuits en cierto periodo de tiempo, se realiza una limpieza y se aplica el método *LIWC* (un método basado en un diccionario para calcular la polaridad de cada palabra y por ende da cada tweet), este estudio es especialmente importante ya que realiza análisis de correlación, propone series de tiempo de la evolución de la percepción de cada usuario y una evaluación de esta a través de una posible métrica: *Positive – Negative Influence*.

De acuerdo con la investigación realizada por (Niño Martinez, Vaca, Rios, & Rey, 2020), utilizan la minería de texto a través de Twitter para determinar el impacto de las campañas realizadas y publicadas que mencionan la cuenta oficial del Ejército Nacional de Colombia. Dentro de los resultados, se utilizan los *hashtags* como una ayuda para identificar los temas específicos de conversación de los usuarios respecto a las publicaciones realizadas y cómo evoluciona su aceptabilidad a través del tiempo, cuantificando el número de favoritos de los comentarios que llevan las etiquetas. Finalmente, implementan diferentes visualizaciones para comprender en términos de relevancia cuales son los hashtags más usados además de presentar el análisis de sentimientos respectivo mediante la aplicación de métodos definidos en lenguaje Python y el diccionario VADER para polarización de palabras en español.

En el sector de salud pública se presenta también otro ejemplo de la evaluación de la percepción de los clientes, en este caso, los ciudadanos relacionados con las medidas de prevención adoptadas por el gobierno con respecto a la pandemia, específicamente el aislamiento social obligatorio. Al igual que en los estudios anteriores, se descargan todos los tweet (con los *hashtag* "cuarentenaobligatoria"), posterior a través de diversas librerías del software R studio (R Core Team, 2021) se realiza la limpieza y *tokenización* de los mismos, obteniendo cuales son los términos que más se repiten y a partir de los cuales se generan nubes de palabras, después se obtiene la polaridad de cada tweet, el resultado final indica más del 49% de los usuarios presentaban un sentimiento negativo y solo cerca del 10% se expresaban de manera positiva (para el restante no logra identificarse el sentimiento) (Salaberry, 2020).

Otro de los sectores que demuestra la transversalidad de los estudios relacionados, corresponde a la percepción y evaluación de las nuevas tecnologías, el autor (Cortez Reyes, 2018) propone aplicar técnicas de *data mining* para que a partir de un conjunto de Tuits, se pueda extraer información que permita conocer los principales temas que se está hablando en el sector de la inteligencia artificial y así generar conceptos generales, que por medio del uso de diferentes tipos de representaciones gráficas presenten una interfaz de los *trending topics.* Se recalca que en este estudio se utiliza también *R studio* (R Core Team, 2021) y que igualmente se genera una *term-document matrix (TDM)*, permitiendo la generación de las gráficas de concepto.

El estudio de la percepción del cliente puede llevar a realizar un análisis externo de las compañías de una forma más detallada al involucrar los servicios prestados o ciertas características de los productos, llegando a ser una buena herramienta para generar *Benchmarks* basados en minería de texto y análisis de sentimientos, en donde se parte de definir ciertas características particulares para cada investigación y que son la base para realizar la comparación, encontrando oportunidades de mejora y de mercado, sin tener que realizar esfuerzos costosos en métodos tradicionales como encuestas de satisfacción (Ogudo & Dahj Muwawa Jean, 2019).

Un ejemplo de lo anterior se evidencia en la investigación de los autores (Chamlertwat, Bhattarakosol, Rungkasiri, & Haruechaiyasak, 2018), en la cual se desarrolla un nuevo método para analizar la opinión de los

clientes con respecto a algunas marcas de celular, se realiza una comparación sobre las principales características (por ejemplo, la pantalla, el sistema operativo) y finalmente se generan gráficas para visualizar fácilmente los hallazgos. Este *framework* llamado *MSAS –Micro-blog Sentiment Analysis System-* se compone de los siguientes módulos principales: Preparación: En esta etapa se configura la API para descargar los Tuits, se busca el diccionario que se utilizará para realizar el análisis de polaridad. Por otro lado, se determina cuales características de las marcas se van a analizar y se define el algoritmo de clasificación relacionado a que Tuits son opiniones y cuáles no, a través de las siguientes etapas: *Message Collecting Module (MCM)* (recolección de Tuits mediante la API), *Opinion Filtering Model (OFM)* (aplicación del modelo para saber que comentarios son opiniones), *Polarity Detecting Moduel (PDM)* (definición de la polaridad o sentimiento de cada tweet), *Feature Classification Module (FCM)* (se atribuye a cada marca los sentimientos de sus Tuits evaluados) y *VisuaLization Module (SVM)* (se proponen vistas como graficas de radar, de líneas y de barras para realizar la comparación de la polaridad encontrada relacionada a las marcas y los atributos de sus productos).

En otra línea de estudio, (Ali, Ahmad, Ur Rehman, & Kamal, 2018) realizan un análisis de competitividad de 3 marcas de gaseosas a través del uso de Tuits; donde la investigación se lleva a cabo bajo un modelo de evaluación de contenido publicitario de cada una de ellas, para identificar los sentimientos de los consumidores y así conocer los puntos fuertes de cada marca y las oportunidades de mejora. Los resultados obtenidos sugieren desarrollar una serie de políticas de riesgos potenciales que afecten la imagen y percepción de marca, además de implementar o mejorar estrategias comerciales.

Los autores (Ranjan, Sood, & Verma, 2019) realizan una investigación sobre el sector de telecomunicaciones de la India, donde el objetivo es analizar el impacto que tendrá sobre los seguidores de las páginas oficiales de cada una, la incursión de una nueva empresa y cómo el análisis de sentimiento ayuda a entender la dinámica del mercado con este nuevo oferente. Dentro de su análisis, propone un nuevo modelo de predicción de crecimiento de seguidores en las redes sociales de las empresas del sector, a partir del modelo de correlación de Pearson que se alimenta por medio de la cuantificación obtenida del análisis de sentimientos al aplicar el método TF-IDF, con un diccionario personalizado en lenguaje hindú para dar más precisión al modelo. Algunos de los resultados demuestran una alta predictibilidad de su modelo de crecimiento de usuarios mes a mes con respecto a las proyecciones realizadas con datos reales de la industria; además de resaltar la importancia de personalizar el diccionario utilizado para el análisis de sentimientos de acuerdo con la lengua nativa y modismos usados en la región como factor determinante para llegar al resultado.

En el sector de transportes, la minería de texto ha tomado relevancia para identificar las oportunidades de mejora y nuevas necesidades comunicadas por los usuarios, lo cual ha sido foco de investigación de (Sari, Wierfi, & Setyanto, 2019) al realizar un análisis de sentimientos para identificar la polaridad de estos sobre dos empresas de transporte privado de la India; a través de plataformas Online, usando el clasificador de *Nayve Bayes*

dado a que es un método tradicional con un alto nivel de precisión. La extracción de datos se realiza a través de la API de Twitter y su cuantificación utiliza el método TF-IDF a través el software RapidMiner. Por otro lado, sobresale el uso del método *10-Cross Fold Validation,* para elegir aleatoriamente los datos de prueba de los datos de entrenamiento, además de medir el rendimiento del clasificador. Dentro de los resultados encontrados, se evidencia que el método *Nayve Bayes* tiene un gran nivel de exactitud de acuerdo con los datos analizados (73.24%) y remarca la importancia de mantener un diccionario enriquecido con palabras en lenguaje nativo, además de incluir modismos para mejorar la precisión de los métodos.

Este tipo de estudios puede también aplicarse a empresas de un sector en particular, con el fin de conocer que oportunidades de mejora que hay al interior de cada compañía, llevado el insumo generado por los análisis de la VOC a las esferas de la dirección de áreas como operaciones y servicio al cliente, un ejemplo de lo anterior se presenta en la investigación propuesta por (Zhan, Han, Tse, Helmi Ali, & Hu, 2021), en la cual se analizan tres grandes compañías del sector de *retail* del Reino Unido, con el objetivo de proponer un modelo que permita a las empresas realizar la integración entre los análisis basados en redes sociales y las estrategias de la compañía, para generar nuevos planteamientos, programas y acciones basadas en *customer centricity*. En este estudio, al igual que en los anteriores, se propone un *framework* o modelo de implementación, con las siguientes fases: la primera se relaciona con la recolección de los datos desde diferentes fuentes externas e internas (HTML, APIS o manual), utilizando en esta investigación las API en *R studio* (R Core Team, 2021). Posterior a esto se aplica la fase 2, en donde se realiza el preprocesamiento, creando 3 *data set* (uno por cada compañía) y se eliminan los Tuits repetidos; además de excluir comunicaciones en otros idiomas y los signos de puntuación gramaticales. La fase tres, está estrechamente relacionadas con la aplicación del algoritmo de clasificación *LDA y* el análisis de sentimientos, involucrándose también la visualización. Este estudio demuestra que es posible realizar un análisis para enfocar las decisiones estratégicas y tácticas de los directivos de áreas del sector.

Otro estudio aplicado a niveles de servicio ha sido expuesto por (Ogudo & Dahj Muwawa Jean, 2019) quienes realizan una investigación para determinar el nivel de detracción potencial y de promoción en tres operadores de redes móviles en Sudáfrica, con el objetivo de crear un vínculo bidireccional entre clientes o seguidores y los operadores. Los tuits analizados son extraídos a través de la API de Twitter y procesados a través del software R Studio (R Core Team, 2021), aplicando diferentes librerías destinadas para este procesamiento. Se ejecuta un EDA, para encontrar patrones importantes en la información y correlaciones entre los diferentes parámetros de las comunicaciones (se emplean línea de tiempo para determinar la relación entre seguidores y Tuits), además de extraer los datos de georreferenciación para establecer clústeres de acuerdo con la información geográfica de quien pública en la red social. Se utiliza LDA para predecir y clasificar la polaridad de los datos. Parte de los resultados mostrados, utilizan el algoritmo de Porter para realizar el proceso de Derivación (Eliminación de prefijos, sufijos y afijos) y así formar la visualización de nube de palabras más utilizadas, además de realizar un

análisis de vocablos que expresan sentimientos negativos precedidos por la palabra “no”. Otras de las visualizaciones tratan de obtener el rendimiento del NLP a través de los datos analizados, teniendo un 96.52% de aciertos con este modelo. Finalmente, realizan una red neuronal a partir de pares de palabras que permitirán identificar el nivel de relación entre los usuarios y su relación con los temas encontrados. Dentro de las conclusiones del trabajo, se identifica la obtención del nivel de satisfacción de los clientes para con los operadores móviles, además de hallar más información a parte de calidad de servicio. Se muestra como el análisis de texto puede ser una respuesta en tiempo inmediato a problemas relacionados con la prestación del servicio y esto como puede afectar el NPS medido.

A pesar de los avances que se han mostrado en los potenciales usos que ha tenido la minería de texto y el análisis de sentimientos sobre las redes sociales, la percepción del cliente sigue siendo un problema para las compañías, en especial para las prestadoras de servicios públicos, ya que áreas como servicio al cliente pueden presentar grandes oportunidades de mejora. Por otro lado, dentro de la literatura revisada, no se tienen evidencias ni antecedentes de que se utilizara técnicas de minería de texto aplicado a los canales de atención al cliente como método de medición de la percepción del servicio o su utilización en términos de mejora de procesos y procedimientos internos, e incluso siendo esto utilizado en inteligencia de negocios y ventajas competitivas. En Colombia, más precisamente en Bogotá, no se han realizado estudios a tal nivel de profundidad, y a pesar de que existen algunas encuestas (Bogotá Como vamos, 2021), no se hacen de forma recurrente ni tienen información empresarial que permitan a las áreas estratégicas de las organizaciones tomar decisiones y ajustar sus procesos de acuerdo con los hallazgos.

4. Metodología

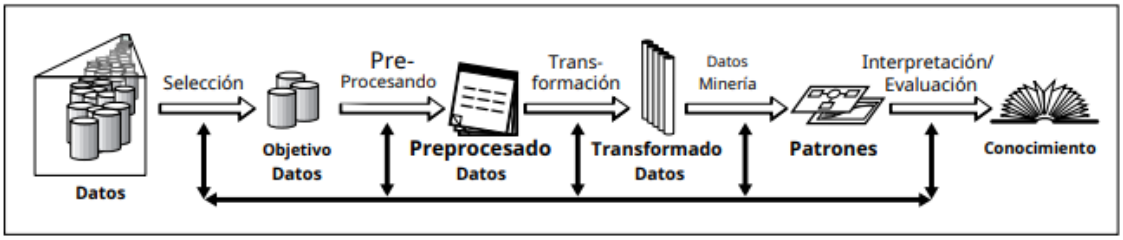
Para cumplir con el objetivo de la investigación y realizar un análisis profundo que permita conocer la percepción del servicio al cliente en las tres empresas mencionadas – Enel, Vanti, Acueducto y Alcantarillado de Bogota – se propone una solución basada en las siguientes etapas genéricas: Extracción y clasificación de datos, procesamiento y visualización de resultados. Para esto se propone el uso de la metodología *Knowledge Discovery Databases (*KDD) propuesto por (Fayyad, Piatetsky-Shapiro, & Smyth, 1996), la cual comprende las fases anteriormente mencionadas e incluye tres más que permiten obtener un mayor conocimiento de los datos.

El modelo KDD se define como proceso no trivial (es decir; busca estructuras, modelos, patrones o parámetros) de identificar patrones válidos, novedosos, potencialmente útiles y comprensibles a partir de datos. El objetivo de este proceso es proporcionar una descripción general acerca de la variedad de actividades y como se interrelacionan para obtener el resultado.

La solución propuesta se basa en la utilización del software R studio (R Core Team, 2021), esto a razón de su versatilidad en cuanto a librerías para aplicar minería de texto, análisis de sentimientos, modelos de *clusterización* y visualizaciones disponibles. Según la literatura encontrada, es un programa que es bastante útil en este tipo de soluciones.

En cuanto a la investigación se propone un enfoque aplicado, dado que se utiliza como base lo revisado en el marco teórico y el método KDD para proponer una solución que integre un análisis mixto, entre lo cuantitativo (de la cantidad de tuits y usuarios catalogados en base a un valor número como positivos, negativos o neutros) y lo cualitativo, al asignarle los diferentes tipos de sentimientos a los usuarios de cada empresa, y a partir de lo anterior poder tener un modelo de comparación. Además de incluir una segunda etapa, en donde se propone realizar análisis temporales que incluyen el *topic modeling* y una evaluación del análisis de sentimiento en el tiempo (variables tanto cuantitativas como cualitativas).

Figura 7. Proceso de descubrimiento de conocimiento método KDD



Fuente: (Fayyad, Piatetsky-Shapiro, & Smyth, 1996)**.**

4.1 Fases modelo KDD

4.1.1 Comprender el dominio de aplicación:

a. **Conocimientos previos relevantes:** Para este paso, se realiza el estado del arte (3. Marco teórico o estado del arte) con el fin de realizar una revisión a trabajos previos sobre el uso de Twitter como fuente de datos y potenciales aplicaciones en diferentes sectores y a partir de los resultados evidenciados en las investigaciones junto con los objetivos del presente trabajo, definir el aporte de esta investigación.
b. **Objetivos de la aplicación:** Se define el objetivo principal para el usuario final, seleccionando las fuentes de datos e identificando si el resultado de la aplicación de la metodología sobre los datos puede ser usados para clusterizar, visualizar, explotar y explorar diferentes puntos de vista.

4.1.2 Extraer la base de datos objetivo:

Consta de realizar la solicitud de la API a través de la página oficial de Twitter (https://developer.twitter.com/en/docs/twitter-api). Una vez se tiene acceso, se instalan las librerías "twitteR", "dplyr" y "tidyr" disponibles en R studio (R Core Team, 2021), con el fin de realizar la extracción de todos los Tuits en la ventana de tiempo que contengan el *hastags* y mención de las 5 cuentas oficiales de las empresas del sector de servicios públicos – (@codensaservicio, @enelcolombia, @enelclientesco, @AcueductoBogota, @grupovanti).

4.1.3 Preparar los datos:

Se aplican una serie de pasos para eliminar nombres de usuarios, espacios, URL's, hashtags, acentos de las palabras que no sean necesarios; para obtener una estructura de datos estándar que tendrán las condiciones necesarias para la clusterización y análisis de sentimientos pretendidos. En esta fase se hará el uso de librerías como "tm" (para aspectos de forma de los textos).

Una vez se tiene el corpus o conjunto de Tuits con la calidad requerida, se busca crear los tokens para realizar el análisis de sentimientos posterior. Este paso es de vital importancia, ya que también permitirá conocer cuáles son los términos que más se repiten. Se puede llevar a cabo mediante técnicas especializadas como el algoritmo TF-IDF.

4.1.4 Minería de datos:

A partir de los resultados obtenidos en la Fase 3, se elabora un modelo que permita medir la percepción del servicio al cliente. Es importante recalcar que en esta etapa se tiene en cuenta el análisis de sentimientos y los valores cuantitativos encontrados a partir de los modelos seleccionados, incluyendo el algoritmo LDA.

Para el análisis de sentimientos, se utilizará la técnica de cuantificación de términos utilizada por (Ba & Lee , 2012). En términos de *topic modeling*, se realizará una evaluación de diferentes técnicas de clusterización para determinar la cantidad de grupos a trabajar con el corpus; y así realizar por cada uno de ellos el análisis correspondiente.

4.1.5 Interpretación:

Con los resultados de las Fase 4 se definen las métricas e indicadores de medición de percepción de servicio al cliente propuesto, se realiza un análisis del estado actual de la percepción del servicio para las empresas de y resolver preguntas clave para el negocio como: ¿Qué áreas de mejora tienen en común las empresas servicios públicos de Bogotá? Además, se proponen una serie de visualizaciones que permitan

simplificar el entendimiento de resultados y así mismo ofrecer diferentes alternativas a las organizaciones para comprender la interrelación de variables extraídas de las redes sociales comentarios (basados en el análisis de sentimientos)?, ¿Cuáles son los indicadores de percepción de servicio que están mejor o peor calificado por cada compañía?, ¿Cuáles son las *keywords* más comunes? y el cambio en el tiempo de *trending topics*, sentimientos, tópicos entre otros.

4.1.6 Utilizar el conocimiento descubierto:

En este apartado y dada la naturaleza de la investigación, se realizará la estructuración de un escrito de divulgación científica que sea base para conocer la situación actual analizada y además funcione como motivación para realizar futuras investigaciones sobre este campo.

5. Resultados y contribución

Con el fin de detallar de una manera ordenada y concisa los resultados obtenidos, se propone una explicación de la aplicación de cada fase de la metodología implementada (KDD), lo cual se detalla a continuación:

5.1 Comprender el dominio de aplicación

5.1.1 Conocimientos previos relevantes

A partir de la revisión teórica que se realizó relacionadas con técnicas de minería de texto y análisis de sentimientos aplicados a reseñas en diferentes medios, se encontró que a nivel general el tema se ha implementado en distintos sectores, pasando por el de *retail* hasta el de servicios y turismo, se encontró también que existen algoritmos y técnicas que permiten determinar la percepción del cliente, utilizándola como insumo para mejorar los procesos de las compañías. Se recalca que, existe una cantidad considerable de investigaciones que pueden ser tomadas como *frameworks* de referencia y que permiten tener una idea general sobre cómo aplicar una solución de minería de texto. A pesar de lo anterior, en Colombia no se evidencian soluciones de este tipo empleadas en los procesos de servicio al cliente, lo que lleva a considerar que la presenta investigación puede sentar una base para aportar a la temática en cuestión. En el anexo 1, se puede visualizar un resumen, mostrando una descripción de los principales aportes relacionadas con en esta temática, clasificados por autor, sector de aplicación, temática general y su fuente.

5.1.2 Objetivos de la aplicación

A partir de lo anterior y teniendo en cuenta la problemática actual de las empresas públicas de servicio, en donde se presenta una alta cantidad de quejas por partes de los clientes y un incremento en la volumetría de los canales digitales, se define el objetivo general como: Analizar la percepción del servicio al cliente en las empresas de servicios públicos de la ciudad de Bogotá: Enel, Acueducto y Vanti; para identificar potenciales áreas con oportunidades de mejora, mediante la aplicación de técnicas de minería de texto y evolución temporal a los Tuits de los usuarios. (en la sección 2. Objetivos, se describen los objetivos específicos de la investigación).

5.2 Extraer la base de datos objetivo

En este paso se detalla todo el proceso relacionado con la solicitud de credenciales y elección de librerías para realizar el proceso de descarga de datos a través de R (R Core Team, 2021). En primera instancia, se debe considerar tener algunas cuentas de Twitter activas con el fin de registrarlas en la aplicación que otorga las API (*Twitter Developer*). El proceso de solicitud de los *tokens* y *keys* para acceder a la información, debe

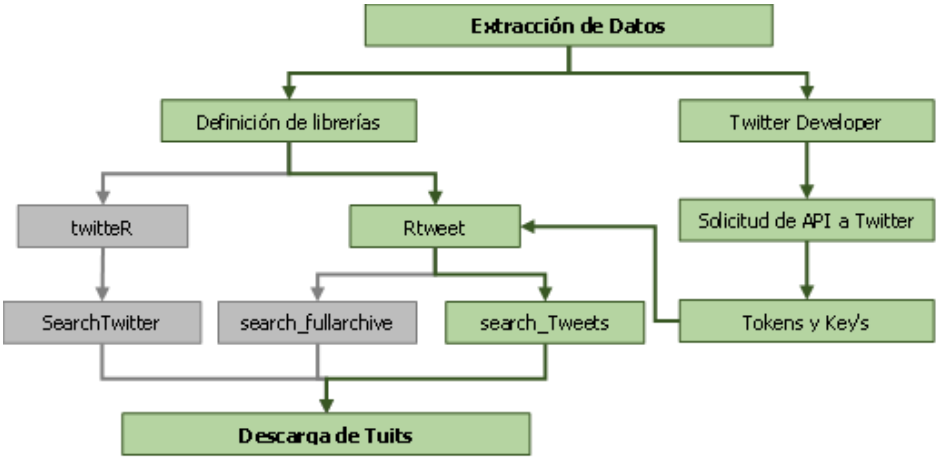
tener en cuenta la ejecución de varias sesiones donde se justifica el uso de los datos, la finalidad del estudio, el tipo de acceso que se debe tener a los datos y las cuentas de las cuales se extraerá la información. En el caso del trabajo de investigación, se solicitó acceso con dos cuentas mediante API's Académicas y con otras dos cuentas, acceso a las API's gratuitas para realizar la extracción de una muestra de información y así construir el código en R (R Core Team, 2021). Las cuentas de Twitter a usar para la investigación son las siguientes:

Tabla 3. Cuentas de Twitter objeto de la investigación

@ codensaservicio	@EnelColombia	@EnelClientesCo
@ AcueductoBogota	@grupovanti	

Fuente: Autoría Propia.

Figura 8. Proceso de selección de librerías para descarga de tuits



Fuente: Autoría Propia.

Una vez obtenidas las credenciales, a partir de un análisis de la literatura relacionada, se definieron las librerías a utilizar para llevar a cabo la tarea de extracción de información: *twitteR* y *Rtweet*. En términos de usabilidad, se usaron ambas librerías para comparar su funcionalidad, encontrando las siguientes características:

Tabla 4. Comparación de librerías de R para generación de token de conexión a Twitter

Variable	twitteR	Rtweet
Tipo de conexión	Directa, sin confirmación en la Twitter Developer	Con confirmación de conexión desde Twitter Developer
Nombre de la API	Se debe definir	No se debe definir

Variable	twitteR	Rtweet
Clave de la API	Se debe definir	Se debe definir
Código de la clave de la API	Se debe definir	Se debe definir
Token de la API	Se debe definir	Se debe definir
Código del token de la API	Se debe definir	Se debe definir

Fuente: Autoría Propia.

Por otro lado, y a partir de las librerías anteriormente mostradas; se compara el uso de tres funciones con el fin de realizar la extracción de datos de cada una de las cuentas, donde se definieron algunos atributos importantes y se logró identificar las siguientes características:

Tabla 5. Comparativo entre funciones de R y sus atributos, para descarga de Tuits

Atributo \ Librería	twitteR – SearchTwitter	Rtweet – Search_fullarchive	Rtweet – search_Tuits
Número de variables	34	-	90
Control de Tuits	Si	-	Si
Control de Retuits	Si	-	Si
Fecha y hora de creación tuit	Si	-	Si
Cadenas de tuits	No	-	Si
URL's de perfil	Si	-	Si
Información básica de usuario	Si	-	Si
URL's de tuits	Si	-	Si
URL's de cadena de tuits	No	-	Si
Fuente de mensaje	Si	-	Si
Información detallada de usuario	No	-	Si
URL's de imagen de perfil	No	-	Si
Descripción del perfil de usuario	No	-	Si
Descarga de tuits por fecha	No	Si	No
Locaciones	No	Si	Si

Fuente: Autoría Propia.

- **search_Tuits:** Esta función permite descargar un objeto tipo *Dataframe* con toda la información detallada de usuarios, fechas, estado de los mensajes e información general y especifica de las cuentas de los usuarios que interactúan con las entidades objeto de estudio.
- **SearchTwitter:** Esta función permite descargar un objeto tipo lista con la información general de los usuarios y el mensaje principal que interactúa con las entidades objeto de estudio.
- **Search_fullarchive:** Esta función puede ser usada si se utiliza una versión paga (premium) de la API, donde permite consultar los tuits de las entidades por rango de fechas.

De acuerdo con las características más importantes de estas diferentes funciones que se usaron para la descarga de tuits a través de R (R Core Team, 2021), se optó por el uso de la función **search_Tuits;** que a pesar de que la descarga se hace en una ventana de tiempo de máximo 8 días atrás a la fecha de consulta, permite tener una mayor cantidad de variables que al ser usadas en el estudio logrando obtener información de mensajes completos, usuarios, interacciones y dispositivos usados para la comunicación. Por otro lado, no se considera el uso de la función **Search_fullarchive** dada la complejidad que se tiene para pagar la versión premium de la API y desconociendo la cantidad de variables e información que contengan cada una de ellas.

La información consultada con la función definida; se transfiere a un *Dataframe* para que de esta manera pueda ser descargado en formato Excel y así estos datos puedan ser depurados y seleccionados en una estructura que permita manipularlos a lo largo del proceso de minería de texto.

5.3 Preparar los datos

Antes de detallar lo realizado en esta etapa, es importante mencionar que, debido al poco reconocimiento del lenguaje latino en R (R Core Team, 2021), se creó un nivel de limpieza adicional al que inicialmente se planteaba, enfocado netamente en la eliminación de palabras sin valor del lenguaje español latino a partir de una muestra de tuits de una de las empresas evaluadas (lo cual se explica más adelante).

5.3.1 Pre Procesador

Mediante el desarrollo de una herramienta programada, se elaboró un procesador de datos tipo ETL como un paso previo a usar las funciones de R (R Core Team, 2021) para la limpieza de los tuits, justificándose en las siguientes razones:

1. Organizar y estructurar la información de interés para la investigación para poder manipular los datos a través de R (R Core Team, 2021).
2. Seleccionar los mensajes de los usuarios, a través de la eliminación de textos de cuentas corporativas, gubernamentales y distritales que con su contenido pueden desviar los resultados que se pretenden obtener con la investigación.

Este preprocesador funciona en 5 pasos:

3. **Cargue:** La herramienta permite seleccionar el archivo descargado de R (R Core Team, 2021) desde cualquier locación del equipo para que sea cargado y transformado.

4. **Selección de variables:** El preprocesador tiene identificadas las 23 variables de interés que se debe extraer del archivo descargado de R (R Core Team, 2021). De esta manera configura una nueva estructura de archivo más compacta.
5. **Consolidación de hilos de usuario:** En esta etapa, se identifican los tuits que tienen "hilos" para considerar la información del mensaje y así el mismo quede completo con la finalidad de obtener toda la trazabilidad del texto.
6. **Eliminación de cuentas:** En este paso, la herramienta contiene con un listado de cuentas gubernamentales, corporativas y distritales con la finalidad de que estas sean eliminadas y así únicamente dejar los tuits de los usuarios.
7. **Consolidación de datos cargados:** El preprocesador mantiene la información trabajada de cada cargue y la exporta una vez se defina que la base ya está completa para ser trabajada en R. (R Core Team, 2021).

Con la selección y eliminación de tuits que se acaba de plantear, se pierde en promedio para cada entidad estudiada el 24.06% de los tuits iniciales. Una vez obtenidas las bases finales para cada empresa, estas son cargadas en R (R Core Team, 2021) para iniciar el proceso de limpieza de los textos.

5.3.2 Limpieza

La limpieza de los tuits consta de eliminar términos, caracteres normales y especiales sin valor, adicionalmente, retirar URL, y menciones que se tienen en los datos descargados. Este subproceso también consta de estandarizar los comentarios que se bajan, con el fin de poder aplicar técnicas de conteo y visualización básica. Inicialmente solo se contemplaban dos pasos, pero para tener resultados más precisos se desarrolla un paso adicional (el cual se explica en secciones posteriores).

5.3.2.1 Primer nivel de limpieza

Una vez se tienen los tuits consolidados se eliminan todas las direcciones mencionadas por los usuarios, así como, hashtag, cuentas, espacios en blanco y doble espacio. Finalmente, se transforman los tuits en mayúsculas, con el fin de facilitar la eliminación de los *stopwords* en siguiente nivel. Para mayor detalle, dentro del código, los comandos utilizados pueden encontrarse en la sección "LIMPIEZA – Primer nivel de limpieza".

5.3.2.2 Segundo nivel de limpieza

Como se mencionó previamente, esta etapa no estaba considerada inicialmente, sin embargo, al realizar las primeras graficas de frecuencia de términos y nubes de palabra, se evidencia que la cantidad de palabras sin valor aún era bastante alta, esto debido al poco reconocimiento de términos del software. Para mitigar este impacto, se realizó una descarga de una muestra de 2000 tuits de la empresa Enel, una vez hecha la *tokenizacion* y la transformación en formato *Tidy*, se exporto un listado de los términos con su respectiva frecuencia de aparición, la cual se evaluó termino por termino para así consolidar el listado de palabras sin valor.

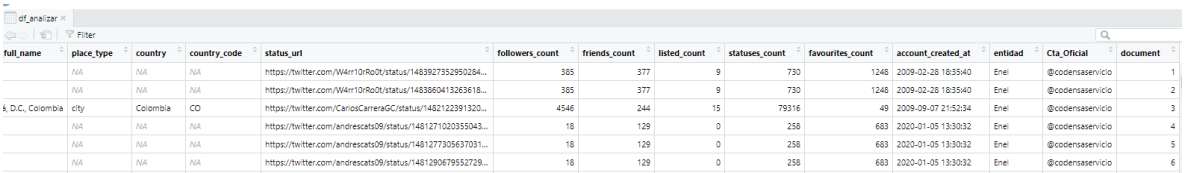
Esta lista, es el principal insumo del segundo nivel de limpieza, lo primero que se realizo fue su cargue en R (R Core Team, 2021), como se puede apreciar en el apartado del código "LIMPIEZA – Segundo nivel de limpieza <- terminos sin valor", posterior a lo anterior, se transforman en mayúsculas todos los vocablos y a través de la función "*removeWords"*, se eliminaron dichas palabras de los datos que se están analizando.

Se considera importante mencionar que, este registro puede ser complementado para un mayor nivel de precisión del análisis y que puede ser adaptado dependiendo del lenguaje y contexto en el cual se esté aplicando la solución propuesta, para realizar esto, únicamente es necesario modificar el archivo que se carga, mencionando en una única columna del archivo de Excel los vocablos a eliminar.

5.3.2.3 Auto numerador

Al igual que el apartado anterior, esta asignación no se contemplaba inicialmente ya que al generar el corpus (lo que se puede encontrar en la siguiente etapa), la única columna que era tenida en cuenta contenía el tuit del usuario, por lo cual, más adelante no se podrían realizar cruces importantes para los análisis posteriores entre los *Dataframe* generados, por ejemplo, no se podría identificar la cantidad de tuits por usuario. Para mitigar esto, se asigna un valor autonumérico al tuit mediante la función "*mutate"* con el argumento "*row_number()"*. Para el ejercicio realizado, la nueva columna que contenía esta información recibió el nombre de "*document*", como se puede apreciar en la Figura 9. Dataframe a analizar con el autonumerador. Lo anterior puede detallarse, en el código "LIMPIEZA – autonumerador"

Figura 9. *Dataframe* a analizar con el autonumerador



df_analizar

Filter

full_name	place_type	country	country_code	status_url	followers_count	friends_count	listed_count	statuses_count	favourites_count	account_created_at	entidad	Cta_Oficial	document
	NA	NA	NA	https://twitter.com/W4rr10rRo0t/status/1483927352950284...	385	377	9	730	1248	2009-02-28 18:35:40	Enel	@codensaservicio	1
	NA	NA	NA	https://twitter.com/W4rr10rRo0t/status/1483860413263618...	385	377	9	730	1248	2009-02-28 18:35:40	Enel	@codensaservicio	2
á, D.C., Colombia	city	Colombia	CO	https://twitter.com/CarlosCarreraGC/status/1482122391320...	4546	244	15	79316	49	2009-09-07 21:52:34	Enel	@codensaservicio	3
	NA	NA	NA	https://twitter.com/andrescats09/status/1481271020355043...	18	129	0	258	683	2020-01-05 13:30:32	Enel	@codensaservicio	4
	NA	NA	NA	https://twitter.com/andrescats09/status/1481277305637031...	18	129	0	258	683	2020-01-05 13:30:32	Enel	@codensaservicio	5
	NA	NA	NA	https://twitter.com/andrescats09/status/1481290679552729...	18	129	0	258	683	2020-01-05 13:30:32	Enel	@codensaservicio	6

Fuente: Autoría Propia.

5.3.2.4 Corpus

Una de las fases básicas que se consideró desde la planeación de la investigación, correspondió a la formación del corpus, estaba claro desde el principio que la utilización de esta técnica permitía generar los *tokens* además de complementar la limpieza mencionada anteriormente. Para empezar y como se puede apreciar en el apartado del código "CORPUS – Formación de corpus", a través de la función "*Corpus"* se transformaron los comentarios de los clientes en un el formato requerido para poder aplicar técnicas de minería de texto. Una vez realizado lo anterior, a través de la función "*tm_map"* se eliminaron: números, puntuación, se transformaron en minúsculas los tuits, y se retiraron las *stopwords* definidas en el sistema, con el argumento que permitió reconocer el lenguaje en español y términos distintos a "no" y "si". Todo lo anterior, se puede decir que constituye el tercer y último nivel de limpieza.

5.3.2.5 DTM – Document Term Matrix

Para posibilitar el conteo de términos de los tuits de los usuarios, fue necesario transformar los datos del formato del corpus a una matriz de documentos y términos (*Document Term Matrix*), en donde cada una de las columnas correspondía a una palabra, y cada columna al número de documento. La matriz esta codificada en un código binario, en el que los 0 indica que ese término no se encuentra en el documento de esa fila, mientras que un 1 significaba lo contrario; es decir, el término si se encuentra en ese documento.

La función que permitió realizar esto, se conoce como "*DocumentTermMatrix"* tomando como argumento principal el corpus generado y teniendo algunos controles importantes para no eliminar palabras valiosas, específicamente el relacionado al *stemming* y al de la *longitud* de las palabras, este último, requirió de un tiempo de consulta mayor dado que inicialmente, se estaban eliminando términos importantes.

Figura 10. DTM para Corpus Enel

corte	cuenta	dice	entiendo	envía	garantía	hará	lectura	mes	necesito	primera	promedio	sino	soluciona
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	2	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	2	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0

Showing 1 to 14 of 3,177 entries, 6297 total columns

Fuente: Autoría Propia.

Según lo esperado, esta matriz tenía un tamaño considerable como lo muestra la la Figura 10. DTM para Corpus Enel, donde se puede evidenciar que se tenían 3177 filas por 6297 columnas, a pesar de que se pudo exportar en otros formatos fuera de R (R Core Team, 2021), fue necesario convertirla en formato *Tidy*, para poder realizar los análisis posteriores de una manera más rápida y eficiente. Lo anteriormente descrito, se encuentra en el apartado "DOCUMENT TERM MATRIX – DTM" del código, allí se evidencian todos los controles inactivos que se utilizaron para no afectar el conteo de palabras.

5.3.2.6 Formato *Tidy*

Como se mencionó anteriormente, para un manejo optimo y relativamente más sencillo de la *DTM*, se debió emplear la transformación a formato *Tidy* de los datos, esta conversión, consistió en totalizar la cantidad de veces que aparecía cada uno de los términos en la totalidad de documentos. En otras palabras, contar cuantas veces aparecía cada palabra en cada tuit.

La función que permitió realizar lo anterior se conoce como *tidy,* tomando como principal argumento la *DTM* y generando una tabla como la que se muestra en la siguiente ilustración:

Figura 11. Conteo de términos en formato Tidy para corpus de Enel

	document	term	count
36	3	solicitar	1
37	4	año	1
38	4	comunicado	1
39	4	contador	1
40	4	corte	1
41	4	cuenta	1
42	4	dice	1
43	4	entiendo	1
44	4	envía	1
45	4	garantía	1
46	4	hará	1
47	4	lectura	1
48	4	mes	2
49	4	necesito	1
50	4	no	2

Showing 36 to 50 of 28,343 entries, 3 total columns

Fuente: Autoría Propia.

En el código, "*FORMATO TIDY"* es la sección en donde se implementó esta función, a partir de este punto, se pudieron empezar a generar las visualizaciones genéricas, además fue el punto de partida para poder aplicar los distintos diccionarios que más adelante se describen.

5.3.3 Diccionarios

Durante el preprocesamiento de datos y teniendo en cuenta el análisis de sentimientos que se realizó, se seleccionaron tres diccionarios con la finalidad de cuantificar y cualificar las palabras del corpus. Estos diccionarios son BING, NRC y Affin.

De acuerdo con la literatura revisada, los antecedentes de los trabajos realizados; R (R Core Team, 2021) trae predeterminado la configuración de estos tres diccionarios, y según los resultados mencionados por otros autores tiene un mejor resultado un diccionario de lengua nativa con todas sus variaciones semánticas y mediante un modelo de asignación de palabras hacia el DTM, la probabilidad de puntuación y clasificación de términos aumenta considerablemente respecto al uso de los diccionarios que posee las librerías de R (R Core Team, 2021).

Con lo anterior, se extrajeron los diccionarios que utiliza R[2] (R Core Team, 2021), los cuales contienen un conjunto de palabras sin modismos ni regionalismos y disponible en diferentes idiomas. Con el diccionario Affin, se pretende obtener el *score polarity* que es la cuantificación de una palabra entre una escala de -1 y 1 (siendo los valores negativos de la escala, palabras negativas; 0 palabras neutras y los valores positivos de la escala, palabras positivas). El diccionario Bing permite obtener un calificativo para cada palabra entre las categorías positivas, negativas y neutras; y finalmente el diccionario NRC (*National Research Council Canada*) clasifica cada palabra en una o varias categorías que están predefinidas para este diccionario (ira, anticipación, asco, miedo, alegría, tristeza, sorpresa, confianza); usando notación binaria para cada palabra (1 asignado al sentimiento, 0 no asignado al sentimiento).

Para efectos de la presente investigación, se extrajeron de cada uno de los diccionarios descargados, la base de palabras correspondientes en idioma español y se realizó un trabajo de exploración manual para eliminar palabras repetidas dando manejo a cada diccionario de la siguiente manera respecto a cada atributo:

- **Diccionario Bing:** Se unificaron los conceptos dando prioridad a los calificativos extremos (POSITIVO y NEGATIVO). En caso de que el concepto no sea concluyente con sus valores extremos, se utiliza la ayuda del diccionario Affin para que con sus puntajes las palabras sean correctamente asignadas.
- **Diccionario Affin:** Se promediaron los puntajes de cada palabra.

[2] Los diccionarios fueros descargados de la página web http://saifmohammad.com/WebPages/lexicons.html, el cual es un blog diseñado por Saif Mohammad; quien es el creador de los diccionarios para las librerías de R que se usan en análisis de sentimientos.

- **Diccionario NRC:** Se unificaron las asignaciones de las palabras repetidas en una sola fila, con el objetivo de que las categorías de sentimientos que se consideren apliquen a cualquier contexto en el que sea usado este diccionario.

Finalmente, para obtener una mayor puntuación y abarcar una mayor proporción de los corpus; se ingresan manualmente las posibles variaciones de cada palabra (asignándole el valor de su palabra raíz o verbo), incluyendo terminaciones como se indica en la tabla 5; además de considerar las acentuaciones y puntuaciones de cada palabra para brindar un contexto correcto al ejercicio.

Tabla 6. Terminaciones de palabras usadas en variaciones semánticas

Terminación de la palabra	Terminación de la palabra	Terminación de la palabra
A	Er	en
E	Ir	in
I	S	m
O	R	n
U	L	d
Ar	An	z

Fuente: Autoría Propia.

Una vez finalizadas las aplicaciones de cada variación de las palabras de cada diccionario, se procede a cargar cada uno de estos en un *Dataframe* en R (R Core Team, 2021) para su uso mediante un modelo de asignación entre el DTM y cada uno de ellos.

5.3.3.1 Diccionario – Asignación al DTM y *Dataframe* final

Para obtener el *Dataframe* final que permitió realizar toda la minería de texto y análisis de sentimientos, se utilización tres tipos de diccionarios para cuantificar y cualificar los tuis. Es importante mencionar que a partir de una muestra aleatoria de 1928 tuits de la empresa Enel, se realiza un análisis del rendimiento del porcentaje de asignación de cada método (lo que se presenta en la Tabla 7, teniendo como criterio el de selección el método que permitiera un porcentaje de asignación más alto), el primero al emplear los diccionarios predefinidos en sistema y el segundo al estructurarlos y mejorarlos.

- **Diccionarios predefinidos:** Se utiliza la librería *syuzhet*: usando la función *get_nrc_sentiment* para el diccionario NRC, la función *get_sentiment (method = "bing")* para el diccionario Bing y la función *get_sentiment (method = "Affin")* para el diccionario Affin.

- **Diccionarios estructurados:** Se utiliza la función *merge* y los *Dataframe* de los diccionarios previamente cargados, que mediante una asignación teniendo como campo clave las palabras, se pueda consolidar un *Dataframe* definitivo.

Tabla 7. Resultados obtenidos al comparar método de R y método de asignación de palabras

Diccionario	Librería syuzhet		Merge (Asignación)	
	Tuits afectados	%	Tuits afectados	%
Bing	948	49.17 %	1367	70.90 %
Affin	728	37.75 %	1226	63.59 %
NRC	845	43.82 %	732	37.97 %

Fuente: Autoría Propia.

Con estos resultados obtenidos y dado la mejora obtenida con la personalización de los diccionarios; se define utilizar mediante la asignación al DTM el diccionario Bing y Affin, y utilizar el diccionario NRC predefinido en R (R Core Team, 2021) para formar el corpus final. Los hallazgos que a continuación se describen aplican para ambos métodos previamente explicados:

- **Problemas de ortografía:** El análisis de los textos de los tuits revela que muchas de las palabras no contienen algún valor o asignación, esto se debe a que el usuario ha tuiteado de manera incorrecta en términos de ortografía cometiendo uno o más falencias en una misma comunicación. A estos términos no se les puede asignar un valor, dado que la coincidencia de la palabra debe ser exacta a como está en los diccionarios.
- **Letras repetidas consecutivamente, en una palabra:** Algunas de las palabras contenían más de una letra repetida, que por lo general está posicionada al final de cada palabra lo que genera que el diccionario no las detecte ni les asigne algún valor o calificativo. (por ejemplo: “ayudaaaa”).
- **Abreviaciones:** Dentro de los vocablos que no obtuvieron valor, se identificaron abreviaciones a términos que dado a que no se encuentran en ningún diccionario no se puede asignar valores (por ejemplo: “cll” abreviación de “calle”).
- **Palabras unidas:** Algunas palabras de los tuits carecían de un espacio entre ellas que las separara para que el diccionario las lograra tomar, lo que incrementa el porcentaje de términos no reconocidos (por ejemplo: “nocontestan”).
- **Nombres de personas, barrios, ciudades, departamentos, locaciones, entre otros:** La Investigación demostró que los usuarios reportan su ubicación (utilizando el nombre de los municipios, calles, barrios etc) estos términos no se contemplan dentro de los diccionarios dado que son palabras propiamente del dialecto colombiano.

5.4 Indicadores a partir de diccionarios

5.4.1 Indicadores Diccionario BING

Dentro de la revisión literaria que se realizó con respecto al tema de la aplicación de análisis de sentimientos y minería de texto, se considera de alta relevancia la propuesta ejecutada por los autores (Ba & Lee , 2012), como se describió previamente, el objetivo era determinar la percepción de ciertas figuras públicas (en su mayoría políticos) y realizar algunos análisis estadísticos, para encontrar correlaciones con los sucesos en el mundo "real" (el detalle de esto se encuentra en la sección: 3. Marco teórico o estado del arte), en este orden de ideas, se recalca que los autores plantean algunos indicadores que permiten realizar una comparación entre la percepción de las figuras públicas analizadas, como lo afirman:

"We define the sentiment score and *polarity* of each tweet and user in order to detect the positive or negative audience of a popular user. We then define the positive-negative ratio of a popular user. We define the sentiment score for each popular user, i, audience user, j, and tweet, k, as the ratio of the positive word count versus negative word count. If a sentiment score is greater than 1, we can classify the *polarity* as positive; otherwise, it is negative. The positive-negative ratio (PN ratio) for a popular user is defined as the number of positive tweets or users from the audience divided by the number of negative tweets or users from the audience" (Ba & Lee , 2012).

Estos fueron acogidos, adaptados e implementados en la presente investigación, partiendo de la asignación que se realizó con el diccionario BING; a continuación, se mencionan los principales:

- **TPS (*Tweet Polarity*):** Esta variable asigna la etiqueta tuit positivo, tuit neutro o tuit negativo de acuerdo con resultado de la diferencia entre el número de palabras positivas y negativas de cada tuit.

Ecuación 1. Cálculo del tuit *Polarity* mediante la asignación por palabra del diccionario BING

$$Etiqueta_{Tuit} = No.palabras^{+} - No.palabras^{-} \begin{cases} Si\ Etiqueta_{Tuit} > 0, Tuit\ Positivo \\ Si\ Etiqueta_{Tuit} = 0, Tuit\ Neutro \\ Si\ Etiqueta_{Tuit} < 0, Tuit\ Negativo \end{cases}$$

Fuente: Autoría Propia.

- **UP (*User Polarity*):** Esta variable asigna la etiqueta usuario positivo, usuario neutro y usuario negativo de acuerdo con los resultados obtenidos de TP, donde la diferencia entre tuits positivos y tuits negativos dan la asignación al usuario.

Ecuación 2. Cálculo del User *Polarity* por tuit para cada empresa

$$Para\ cada\ entidad\ (i), Para\ cada\ usuario\ (j),$$

$$User\ Polarity_{i,j} = No.Tuits\ positivos_{i,j} - No.Tuits\ negativos_{i,j}\ ;$$

$$\begin{cases} Si\ User\ Polarity_{i,j} > 0, Usuario\ Positivo \\ Si\ User\ Polarity_{i,j} = 0, Usuario\ Neutro \\ Si\ User\ Polarity_{i,j} < 0, Usuario\ Negativo \end{cases}$$

Fuente: Autoría Propia.

- **Ratio PN Tuit:** Relación entre la cantidad de tuits positivos sobre la cantidad de tuits negativos.

Ecuación 3. Cálculo del PN *Tuit Value* por empresa

$$PN\ Tuit\ Value_{Empresa} = \frac{No.Tuits\ Positivos}{No.Tuits\ Negativos}$$

Fuente: (Ba & Lee , 2012).

- **Ratio PN User:** Relación entre la cantidad de usuarios catalogados como positivos sobre la cantidad catalogados como negativos.

Ecuación 4. Cálculo del PN *User Value* por empresa

$$PN\ User\ Value_{Empresa} = \frac{No.Usuarios\ Positivos}{No.Usuarios\ Negativos}$$

Fuente: (Ba & Lee , 2012).

5.4.1.1 Conteo General Palabras - BING

Inicialmente, se planteó determinar la cantidad total de palabras positivas, negativas y neutras que se encontraron en cada conjunto de tuits de las empresas. Para realizar esto, se utilizaron las funciones: "select", "count", "agrupación", "prop.table" y "round". Una vez se extrajeron del *Dataframe* principal las columnas a evaluar (documento, la entidad y la calificación Bing de cada termino (pudiendo ser únicamente: Positivo, Negativo o Neutro)), se realizó un conteo general sobre este último campo, a partir del cual se construyó una tabla genérica, en donde se muestra el porcentaje y cantidad de cada grupo de palabras encontrado.

5.4.1.2 Bing aplicado a tuits

Uno de los indicadores importantes a cuantificar es la polaridad de los tuits y el ratio entre los tuits positivos y negativos; esto con el objetivo de identificar para cada una de las empresas la composición de

los mensajes bajo el etiquetado que maneja este diccionario y establecer la proporción de tuits positivos, neutrales, negativos y los no clasificados.

- **Tuit *Polarity* Score (TPS):** Dentro del código, la sección identificada como TPS, corresponde a la implementación en el código del indicador TPS (*Tweet Polarity*) (Ba & Lee , 2012). Para esto, se extrae del *Dataframe* principal las siguientes variables: *document, term*, entidad y Bing (clasificación asignada a cada palabra), realizado el conteo por cada uno de los tuits para determinar la diferencia de las palabras positivas y negativas y basado en el "*Tweet Polarity Score*". Las funciones utilizadas en este párrafo corresponden a: "count", "spread", "subset", "if else".
- **Ratio PN Tuits:** Con todo lo anteriormente descrito, se pudo encontrar la proporción entre los tuits identificados como positivos sobre los tuits identificados como negativos. Esta razón se realizó haciendo un conteo de los tuits con polaridad positiva y negativa; lo que permitió identificar un único valor por corpus.

5.4.1.3 Bing aplicado a Usuarios

Es importante encontrar la relación entre los usuarios identificados con una polaridad positiva y negativa con el objetivo de aportar una idea respecto a los usuarios que se encuentran satisfechos y contentos o si por el contrario, presentan comunicaciones negativas por Twitter.

- **UP (User *Polarity*):** En este apartado es necesario contar con el nombre del usuario *(screen_name*), además del *document, term*, entidad y Bing; para realizar el conteo para cada usuario en cada tuit de los términos positivos, negativos y neutros. Luego se eliminan los términos que no puntuaron o no se encontraron en el diccionario y finalmente para cada cuenta de cliente, se calcula la razón entre los tuits positivos y negativos (*User sentiment Score*), siendo esto la base para encontrar la polaridad del usuario (UP).
- **Ratio PN Usuarios:** Para encontrar la relación entre usuarios asociados a un sentimiento positivo y negativo, uno de los posibles indicadores que se considera relevante en la percepción de servicio, se realizó un conteo que determinaba la totalidad general de los usuarios que habían escrito un tuit considerado como positivos o negativo, y a partir de esta encontrar el valor del ratio entre estas dos variables.

5.4.2 Indicadores Diccionario AFINN

Como se mencionó en la sección 5.3.3, este diccionario permite asociar a cada palabra un número que varía entre -1 y 1 dependiendo si se considera un término positivo o negativo, permitiendo que través del *score polarity*, se pueda a cuantificación del tuit basado en las palabras que este contiene.

Es importante considerar esta cuantificación, dado que permite llevar los resultados obtenidos del tuit a la escala NPS que cuantifica el nivel de satisfacción del cliente con los servicios prestados, variando las opciones de respuesta de 1 a 10 agrupadas de la siguiente manera:

- Los puntajes obtenidos de 1 a 6 se consideran detractores.
- Los puntajes obtenidos entre 7 y 8 se consideran neutros o neutrales.
- Los puntajes obtenidos entre 9 y 10 se consideran promotores.

5.4.2.1 Asignación Valor NPS basado en AFINN

Para realizar una homologación con el proceso anterior, se asignó el valor cuantitativo del diccionario Afinn a cada una de las palabras que compone cada tuit. Después se multiplica la frecuencia de cada palabra en cada tuit (n) con su respectivo puntaje y finalmente se procedió a promediar los valores encontrados de cada termino, incluyendo los repetidos y los negativos.

Ecuación 5. Cálculo para cuantificar el puntaje de cada tuit aplicando el diccionario Affin

$$Para\ cada\ tuit\ (i), para\ cada\ palabra\ (j);$$

$$Puntaje_i = Prom\big(Suma(Puntaje\ Affin_j * n_j)\big)$$

Fuente: Autoría Propia.

5.4.2.2 Homologación de Escala NPS

Realizado lo anterior, se prepararon los resultados para ajustarlos a la escala y así determinar a la proporción de usuarios que pertenecen a cada categoría. El ajuste de los resultados a la escala se realiza de la siguiente manera:

- Se elimina el 2.5% de los valores más altos y el 2.5% de los valores más bajos, trabajando así con el 95% de los datos. Esto permite eliminar los valores atípicos para cada base de datos y reducir la variabilidad de estos.

- Se calculan los quintiles para cada conjunto de datos; esto posibilita dividir de manera homogénea los resultados y así etiquetarlos por quintil (pares de 2 en 2 tomando el valor mínimo como 1 y el valor máximo como 10).
- Después de determinar a cuál quintil corresponde cada resultado, se realiza el etiquetado de los datos para posteriormente determinar su cantidad y representación porcentual sobre los mismos.

Con el anterior procedimiento realizado, se puede calcular a partir de los datos restantes los indicadores NPS por tuit y usuarios como se explica a continuación:

- **Cálculo del NPS por Tuit:** Para poder contar con el NPS basado en la cantidad de tuits por cada empresa, se extrajeron las variables *document*, calificación y entidad de cada *Dataframe* por entidad y posteriormente se realizó la asignación de la etiqueta del sentimiento que representa el tuit como se indica en la **Ecuación 6** y así determinar si es un tuit detractor, neutral o promotor:

Ecuación 6. Etiquetado de cada tuit con diccionario Affin

$$\text{A partir de la } \textbf{Ecuación 5}$$

$$Puntaje_i \begin{cases} Si\ Puntaje_i > 0, Promotor \\ Si\ Puntaje_i = 0, Neutro \\ Si\ Puntaje_i < 0, Detractor \end{cases}$$

Fuente: Autoría Propia.

Con lo anterior, se obtuvo el NPS promedio general de tuit para cada empresa; en el cual se agrupó mediante un conteo las etiquetas de los tuits y se determina por medio de la diferencia entre los tuits promotores y detractores; el puntaje final sobre la totalidad de tuits etiquetados.

Ecuación 7. Cálculo del NPS promedio por tuit para cada entidad

$$Para\ cada\ entidad\ (i);$$

$$NPS\ Prom\ por\ tuit_i = \frac{No.Tuits\ Promotores_i - No.Tuits\ Detractores_i}{No.Total\ Tuits_i}$$

Fuente: Autoría Propia.

- **Cálculo del NPS por Usuario:** Para tener la cantidad de usuarios por etiqueta, se extraen las mismas variables de la sección incluyendo el *screen_name.* Posterior se realizó el conteo de la cantidad de usuarios en cada etiqueta (detractor, neutral y promotor) y finalmente se calcula la

diferencia entre la cantidad de tuits promotores y detractores por usuario para establecer si el usuario es detractor, neutral o promotor por empresa:

Ecuación 8. Cálculo para etiquetar a cada usuario por entidad

$$Para\ cada\ tuit\ (i), para\ cada\ usuario\ (j)$$

$$Diferencia_j = Prom\left(Puntaje\ Affin_j\right) \begin{cases} Si\ Diferencia_i > 0, Promotor \\ Si\ Diferencia_i = 0, Neutro \\ Si\ Diferencia_i < 0, Detractor \end{cases}$$

Fuente: Autoría Propia.

Con lo anterior, se obtuvo el NPS promedio general de usuario para cada empresa; en el cual se agrupó mediante un conteo las etiquetas los usuarios y se determina por medio de la diferencia entre los usuarios promotores y detractores; el puntaje final sobre la totalidad de los usuarios etiquetados:

Ecuación 9. Cálculo del NPS promedio por Usuario para cada entidad

$$Para\ cada\ entidad\ (i);$$

$$NPS\ Prom\ por\ Usuario_i = \frac{No.Usuarios\ Promotores_i - No.Usuarios\ Detractores_i}{No.Usuarios\ Tuits_i}$$

Fuente: Autoría Propia.

5.4.3 Indicadores Diccionario NRC

De acuerdo con lo mencionado en la sección 5.3.3, este diccionario cuenta con 8 sentimientos definidos donde cada palabra se categoriza en uno de estos. Es importante considerar este diccionario dentro del análisis de sentimientos, ya que permite identificar la cantidad de palabras por categoría y, además, conocer la composición del corpus por sentimiento y con esto determinar cuáles son aquellos con los que los usuarios se sienten identificados.

5.5 Análisis de sentimientos y visualizaciones generales

5.5.1 Visualizaciones generales

Con el fin de poder generar algunas conclusiones y análisis respecto a los corpus obtenidos, se realizaron algunas visualizaciones básicas, las cuales se explican a continuación. Las gráficas de frecuencias y las nubes de palabra se plantean para cada uno de los corpus de los tuits de las empresas y es importante resaltar que, esta es la primera fase del análisis la cual está orientada a identificar ideas generales de lo que

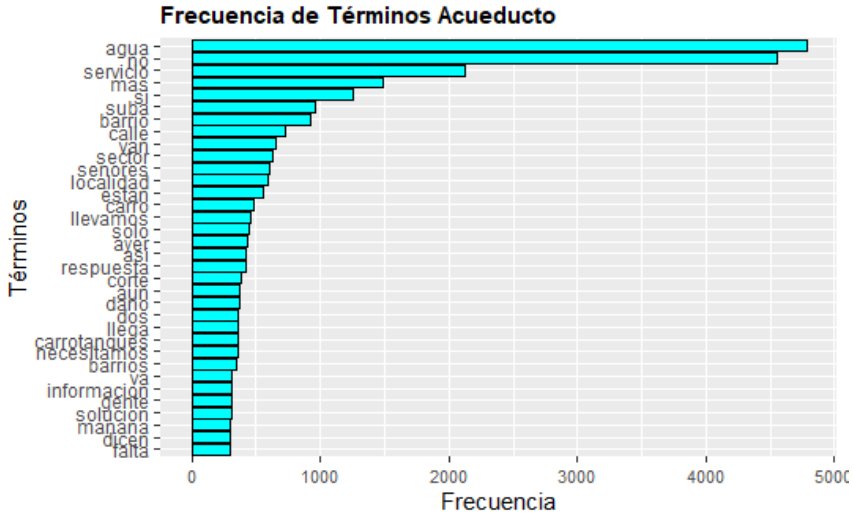
los clientes están tuiteando. Así mismo, para todos los puntos que se detallan a continuación es necesario que los datos estén en formato *tidy*¸ y que previamente se hayan realizado los niveles de limpieza requeridos.

5.5.2 Graficas de frecuencias de palabras

Para realizar esta visualización, se emplearon las funciones: "*count", "filter", "mutate" y "ggplot".* Se realizó el conteo de los términos con mayor frecuencia para así filtrarlos y garantizar dejar los términos que aparecen en promedio 3 veces por día, se reordenaron de mayor a menor y se generó un gráfico de barras horizontal. Este proceso, se puede encontrar en el apartado del código: "Grafica Frecuencias".

5.5.2.1 Acueducto

Figura 12. Grafica de Frecuencias Corpus Acueducto (Frecuencia mayor a 300)



Fuente: Autoría Propia.

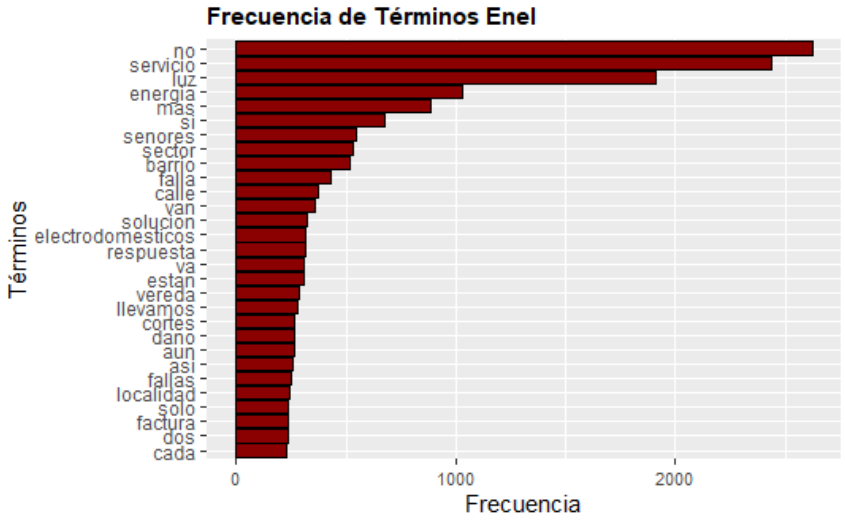
De la Figura 12 se puede resaltar la amplia cantidad y variedad de términos, fundamentando en que el acueducto es la empresa con mayor cantidad de tuits que tiene en su corpus. Esto último, explica también por qué además de tener un número superior de términos que se repiten, cuenta con un número más alto de repeticiones de palabras a nivel general (5000). Se consideraron que los términos más relevantes son los siguientes: "Suba, Barrio, localidad, sector, carro, llevamos, respuesta, corte, solución. Información y solución".

Al igual que para las otras dos empresas, las palabras de mayor frecuencia: "agua" y "servicio" permiten inferir que los mensajes van dirigidos a la compañía adecuada. Para este corpus, un hallazgo importante a considerar es que aparece la localidad de "suba" por lo cual la gran mayoría de mensajes giran

en torno a los usuarios ubicados en esta zona (se confirma dado que palabras como: “calle” “barrio” y “sector” permiten determinar que los cortes en el servicio afectan zonas amplias y que los clientes envían su ubicación y quejas de manera masiva por este canal), y así términos como: “carro”, “llevamos”, “corte”, “carrotanque” y “necesitamos” indican que los usuarios solicitan a través del canal una solución para los cortes de servicio de manera ágil. Por último, las palabras “información”, “falta” y “respuesta” permiten identificar que las personas requieren que este canal adquiera más relevancia para informar sobre novedades en el servicio y que puede haber una posible falta de respuesta hacia los usuarios.

5.5.2.2 Enel

Figura 13. Grafica de Frecuencias Corpus Enel (Frecuencia mayor a 230)



Fuente: Autoría Propia.

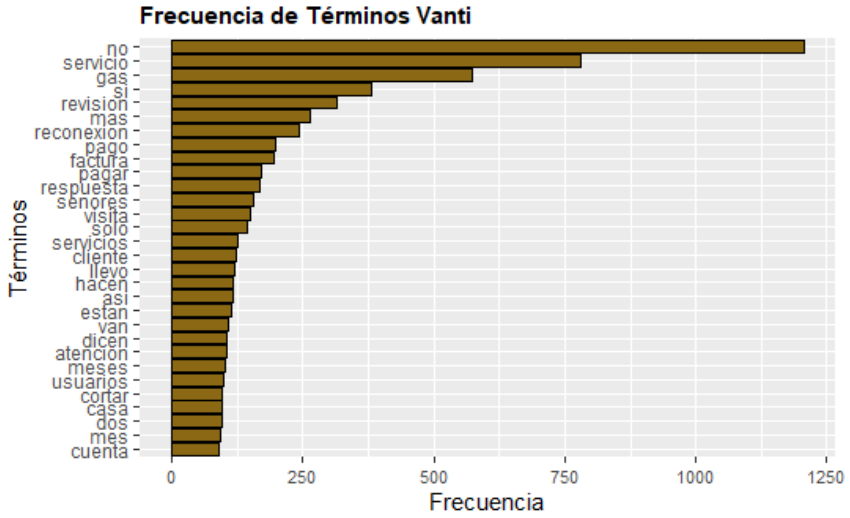
En el caso de Enel, los resultados obtenidos se muestran en la Figura 13 donde se puede evidenciar que la máxima frecuencia de repetición asociada a un término es de 2650, siendo el segundo corpus que mayor cantidad de tuits presenta. De las palabras que aparecen en el gráfico se consideran de mayor relevancia: “no, servicio, sector, barrio, falla, calle, van, solución, electrodomésticos, respuesta, llevamos, vereda, cortes, daño, localidad, factura”.

Términos como “servicio” y “luz” permiten inferir que los tuits se dirigen directamente a la empresa prestadora del servicio debido a su alta frecuencia de repetición, seguido de “sector”, “calle” y “barrio” lo que indica que los clientes usan este canal de comunicación como medio para reportar sus incidentes relacionados con la continuidad del servicio y que, además; los “cortes” y “fallas” afectan zonas bastante amplias (explicando también la aparición, aunque en menor frecuencia de palabras como “localidad”). Otras

palabras como “solución” y “respuesta” en este contexto para el corpus analizado, cobran una relevancia significativa ya que se puede inferir que los clientes no tienen respuesta de sus incidentes y que por lo tanto no tienen una solución o respuesta de este, llevando a que se hagan contactos reiterativos al mismo usuario, lo que se traduce en la aparición de términos como: “llevamos” y “aun. El resto de términos de la Figura 13, también permiten inferir los temas afectan a los usuarios como por ejemplo “fallas” y “cortes”; que se asocian a cortes en el servicio, “electrodoméstico” y “daño” que se traduce en la intermitencia que afecta este tipo de bienes de los clientes, por lo que se observa que gran cantidad de los reclamos se relacionan con este tipo de incidente y el término “factura” que puede llegar a relacionarse con un ámbito económico que afecta directamente al cliente. Por último, el término “vereda” permite concluir que los clientes en centros rurales hacen uso de este canal y no únicamente los ubicados en Bogotá.

5.5.2.3 Vanti

Figura 14. Grafica de Frecuencias Corpus Vanti (Frecuencia mayor a 90)



Fuente: Autoría Propia.

Como se observa en la Figura 14, dado que esta compañía es la que menor cantidad de tuits tiene, su cantidad de palabras con alta frecuencia, así como el número máximo de repeticiones (1250) es la menor de los tres conjuntos de tuits analizados.

Sin embargo, se encuentran algunas palabras de alta relevancia como “Servicio, revisión, factura, pagar, reconexión, respuesta, visita, llevo, casa, nunca, cortar, meses, atención y tiempo.” Al igual que para los corpus anteriores, las palabras con mayor frecuencia permiten identificar que el usuario se dirige a la compañía prestadora del servicio correspondiente.

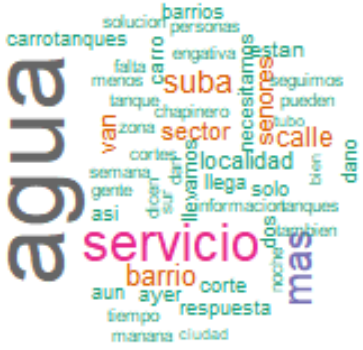
Se infiere que la aparición de una alta frecuencia para las palabras como “reconexión”, “visita”, “casa”, indica que el proceso en general de asistir al predio del cliente para realizar algún tipo de tramite está afectando lo que se comenta vía Twitter. Se puede observar que aparecen términos vinculados a “factura”, “pago” y “pagar”; esto quiere decir que los clientes están contactando por este medio para reportar algún tipo de novedad con el ámbito económico que los afecta directamente. Finalmente, aparecen palabras como “respuesta”, “cliente”, “atención”, “usuarios” y “cuenta” por lo cual se puede inferir que el proceso de atención al cliente sufre algún tipo de novedad que influye en los mensajes que el cliente emite a través de la red social.

5.5.3 Nubes de palabras

Con el fin de poder complementar los resultados obtenidos de los gráficos de frecuencia, se realizaron las siguientes nubes de palabras las cuales permiten realizar un análisis a mayor profundidad e identificar fácilmente nuevos términos que por su poca frecuencia no se evidenciaron anteriormente. Para generar esto, se puede consultar el código en el apartado “*Wordcloud”,* en donde la función principal que se utilizó se conoce como: “*wordcloud”*. Finalmente, para lograr interpretar de una manera amena y ágil los resultados, se considera un máximo de cincuenta términos para cada visualización obtenida por cada entidad.

5.5.3.1 Acueducto

Figura 15. Nube de palabras Corpus Acueducto (Top 50)



Fuente: Autoría Propia.

La nube de palabras permite complementar lo anteriormente mencionado, en especial el tema de las localidades y sitios en específico desde donde los clientes escriben, por ejemplo, aparece junto con **“**suba” nuevas localidades como: “Engativá”, y “chapinero”, además de que los términos como **“**barrios”, “calle**”**, “ciudad” y **“**sur**”** recalcan la alta frecuencia con la que el cliente reporta su ubicación cuando existe

alguna novedad. Se evidencian términos relacionados con la razón de las posibles fallas más frecuentes, por ejemplo, **"**tubo", "daño", "falta", "cortes" por lo cual puede ser un posible indicio de los procesos que esta empresa debe analizar. Vocablos como: "solución", "seguimos", "dos", "noches", "aun" , "ayer" y "mañana" puede indicar que el tiempo de respuesta es bastante elevado.

5.5.3.2 Enel

Figura 16. Nube de palabras Corpus Enel (Top 50)

Fuente: Autoría Propia.

Para el corpus de Enel, aparecen términos como **"**pésimo" palabra que demuestra una gran inconformidad por parte de los usuarios, se hace evidente y se confirma que los clientes utilizan el canal para reportar las "fallas", esto debido a la aparición de los siguientes vocablos: "suba", "zona" "casa". Se recalca la aparición de la palabra **"**caso" y "numero**"** lo que indica que los usuarios reportan por allí su incidencia (o que previamente ya se les ha emitido por otro medio) y llegan a Twitter para solicitar seguimiento (cliente reincidente), esto último se confirma con la presentación en el gráfico de términos como: "seguimos", "nuevamente", "aun", "cada" y "problema". Por otro lado, la palabra "publico" se identifica con fallas en las luminarias o alumbrado público. Así mismo las palabras: "problema", "ayuda" y "solución" puede ser sinónimo de que el cliente está requiriendo de solución o respuesta a su inconveniente que muchas veces puede ser urgente y de mucha relevancia para el mismo y que pueden existir oportunidades de mejora en el porcentaje de solicitudes o inquietudes que se responden (esto se explica por la presentación de términos como: "nadie" y "responde"). Como en las otras empresas, aparecen términos como: "factura", "cuenta" y "pagar", lo que indica que el cliente tiene algún tipo de novedad en el ámbito económico.

5.5.3.3 Vanti

Figura 17. Nube de palabras Corpus Vanti (Top 50)

Fuente: Autoría Propia.

Al aumentar la cantidad de términos que se pueden visualizar aparecen nuevas palabras como **"**ladrones" y "peor", lo que indica malestar a nivel general de los usuarios, en este sentido y como posible causa de ello, se evidencia la aparición de palabras como: "reconexión", "cortar", "nunca" y "solución**,** lo que se traduce en la no respuesta y cortes sin fundamentos que sufren los clientes. Como complemento a lo anterior aparece la palabra "periódica" lo que puede indicar que los usuarios también tienen novedad con las solicitudes de revisiones periódicas y obligatorias que se deben hacer cada cierto tiempo. A nivel general, se puede intuir que el principal tema de los tuits se asocia a las visitas técnicas que se realiza por parte de la empresa, en donde pueden dejar "esperando" al cliente y no se da "respuesta" por ningún otro canal (esto último se relaciona con la palabra "línea", "datos" y "solución").

5.5.4 Análisis de sentimientos a partir de diccionarios

5.5.4.1 Diccionario BING

Al aplicar el diccionario BING a los corpus de cada una de las empresas, se puede visualizar en la Tabla 8 la cantidad de palabras que fueron asignadas de acuerdo con las etiquetas positivas, neutras y negativas y la cantidad de palabras que no lograron ser asignadas al no encontrarse en el diccionario. Dentro de los resultados obtenidos, se resalta la cantidad de vocablos de cada corpus que no lograron ser asignados en promedio el 26.84%, esto permite inferir que esta cantidad de términos están relacionados a modismos y regionalismos propios del idioma colombiano, por lo cual el diccionario requiere un enriquecimiento más robusto para mejorar el etiquetado de datos y precisar los resultados; por otro lado, la etiqueta de palabras neutras supera por el doble a las etiquetas positivas y negativas, lo que permite

determinar que la percepción de servicio a nivel general para las tres empresas analizadas es negativa. Otro aspecto para tener en cuenta es que, de acuerdo con el etiquetado de las palabras positivas y negativas; se evidencia que las primeras son superadas por las últimas casi en una relación de 2 a 1 para las 3 empresas, por lo que se puede inferir que los mensajes que comunican los usuarios a través de la red social perciben palabras en tono negativo y que el canal de comunicación a nivel general se asocia a sentimientos negativos.

Tabla 8. Resultado BING por palabras para cada empresa

Entidad / BING	POSITIVAS		NEGATIVAS		NEUTRAS		NA	
	No. Palabras	% Participación	No. Palabras	% Participación	No. Palabras	% Participación	No. Palabras	% Participación
Acueducto	14018	50,99%	22563	45,80%	49434	50,86%	38784	57,75%
Enel	8939	32,51%	17610	35,75%	35605	36,63%	19522	29,07%
Vanti	4535	16,50%	9086	18,45%	12164	12,51%	8856	13,19%

Fuente: Autoría Propia.

Para analizar las métricas establecidas por (Ba & Lee , 2012), se utiliza el etiquetado de las palabras obtenidas con el diccionario BING y se se obtienen los resultados de la Tabla 9 y el indicador *PN Tuit Value* para cada empresa. (Si el tuit no tiene palabras positivas, neutras o negativas; se etiqueta automáticamente como **Tuit no Etiquetado**).

Tabla 9. Resultado de etiquetado por Tuits para cada empresa e indicador PN tuit Value

Entidad \ BING	Tuits Positivos	Tuits Neutros	Tuits Negativos	Tuits No Etiquetados	PN *Tuit Value*
Acueducto	2823	2951	6357	348	0.444
Enel	1514	1818	4887	139	0.3100
Vanti	449	463	1824	37	0.246

Fuente: Autoría Propia.

Analizando los resultados y los indicadores, el Acueducto tiene el mejor indicador sobre las 3 empresas; ya que por cada 100 tuits negativos la empresa recibe 44 tuits positivos, lo que indica que los usuarios de las empresas están divididos casi equitativamente, ocasionando que pueda pensarse que la percepción no esta tan polarizada al lado negativo. En el caso de Enel, se obtiene que por cada 100 tuits negativos la empresa existe 31 positivos, lo que muestra que la percepción del servicio es muy negativa. Finalmente, para Vanti se obtiene que por cada 100 tuits negativos se reciben 25 tuits positivos, lo que demuestra que además de ser la empresa con el menor puntaje; es la compañía que tiene un valor critico en cuanto a la cuantificación de la percepción de servicio. Un aspecto importante para resaltar es la cantidad

de comunicaciones que no tienen etiqueta, esto se debe a que las palabras utilizadas por los usuarios en el mensaje no fueron categorizadas por el diccionario, en promedio es el 1,93% del total de tuits por empresa. Al considerar estas palabras dentro del enriquecimiento del diccionario BING, podrían presentarse cambios en el indicador obtenido.

Otra de las métricas analizadas es el PN *User Value*; el cual utiliza el etiquetado de los tuits de la Tabla 9 para agruparlos por usuario y así determinar la cantidad de tuits positivos, neutros y negativos. Es Importante mencionar que dentro de los resultados obtenidos que se pueden observar en la Tabla 10, se identifican algunos usuarios que tuitean tanto mensajes positivos como negativos, dada la naturaleza del indicador no se tendrán en cuenta dentro del cálculo para no alterar el resultado; de igual manera se considera la misma premisa para el grupo de usuarios que se identificaron con la etiqueta neutros.

Tabla 10. Resultados calculo PN *User Value* para cada empresa e indicador

Entidad \ BING	User Positivos	User Neutros	User Negativos	User Positivos – Negativos	User No Etiquetados	PN *User Value*
Acueducto	1018	841	2558	826	150	0.398
Enel	557	520	1976	509	52	0.282
Vanti	218	160	837	126	14	0.260

Fuente: Autoría Propia.

En cuanto los usuarios que no tienen asignado alguna etiqueta, en promedio representan el 1,75% de los clientes de cada empresa, al observar el resultado se puede decir que es consecuente con el grupo de tuits de la Tabla 9, ya que la identificación de usuarios depende del resultado de asignación de esta; por lo cual se puede establecer una relación directa de dos mensajes en promedio por usuarios que pueden cumplir esta característica.

Dentro de los resultados a destacar se encuentra la cantidad de usuarios que transmiten mensajes con sentimientos negativos, ya que es muy superior en número respecto a la cantidad de clientes con tuits positivos y neutros juntos. En el caso del Acueducto, es la empresa que contiene el indicador más alto, y este hace referencia a que por cada 10 usuarios que tuitean a la empresa de forma negativa, 4 lo hacen de manera positiva. En el caso de Enel, se tiene un indicador de usuario bajo, permitiendo afirmar que la percepción del servicio es negativa; dado que por cada 10 usuarios que presentan tuits con sentimientos negativos, se tienen 3 clientes que se asocian a comunicaciones positivas. Finalmente, el indicador más bajo de usuario se obtiene en Vanti; donde por cada 10 usuarios que tuitean de manera negativa, 2 usuarios lo hacen de manera positiva, permitiendo secundar lo anteriormente mencionado, en cuanto al valor critico de la cuantificación del servicio.

5.5.4.2 Diccionario Affin

Mediante el diccionario Affin, se pueden calcular dos indicadores derivados de la homologación con la escala del NPS promedio por tuit y usuario; que se obtienen a partir de la medición de la intensidad de las palabras que usan los usuarios en cada mensaje para definir las etiquetas sobre tuits (Esto se explica al detalle en la sección 5.4.2).

En la Tabla 11 se pueden observar los resultados; para el acueducto, se puede observar que los indicadores del NPS tanto para el tuit como para los usuarios son negativos y a nivel general es la empresa que tiene un desempeño inferior de las 3 compañías analizadas mediante este diccionario, esto se debe a que los tuits detractores son mayores a los promotores en una proporción de 2 a 1. Al analizar el indicador a nivel de usuarios, se presenta un resultado negativo dado que el etiquetado de los clientes depende del etiquetado de tuits y por ende, al tener mayor cantidad de mensajes detractores, el total de usuarios en esta categoría también aumenta-

Tabla 11. Resultados diccionario Affin aplicado al Acueducto y cálculo del NPS por tuit y usuario

Etiqueta NPS	Número de Tuits	Número de Usuarios	NPS TUIT	NPS USUARIO
DETRACTOR	6611	2897	-0.242	-0.268
NEUTRO	2276	1044		
PROMOTOR	3592	1452		
Total	12479	5393		

Fuente: Autoría Propia.

En el caso de Enel; se puede observar en la Tabla 12 que los indicadores del NPS tanto de tuit cómo de usuario son negativos (aunque con un mejor resultado respecto al Acueducto), donde la proporción entre tuits promotores y detractores es de 0.59, este ratio a pesar de ser aceptable indica que los tuits en su mayoría son negativos, por lo que se afirma que los mensajes están asociados a este sentimiento. Para el NPS por tuit, se puede ver que los mensajes tienden a transmitir aspectos negativos, de igual manera el NPS por usuario, permite evidenciar que la diferencia entre promotores y detractores es negativa, mostrando que la proporción respecto al total es del -23,4%; siendo la mejor puntuación obtenida para las tres empresas.

Tabla 12. Resultados diccionario Affin aplicado a Enel y cálculo del NPS por tuit y usuario

Etiqueta NPS	Número de Tuits	Número de Usuarios	NPS TUIT	NPS USUARIO
DETRACTOR	4329	1883	-0.213	-0.234
NEUTRO	1484	694		
PROMOTOR	2545	1037		
Total	8358	3614		

Fuente: Autoría Propia.

Finalmente, para Vanti se pueden observar los resultados en la Tabla 13; donde se resalta que el NPS por tuit es negativo dado que la cantidad de tuits detractores es superior a la cantidad de tuits promotores en una proporción de 1.5:1 y esto permite inferir que los mensajes comunican aspectos negativos de la empresa del total de tuits analizados, donde se espera que por cada 10 tuits dirigidos a la empresa sean 5 los que transmitan un mensaje detractor. Analizando el NPS por usuario, se tiene el mejor resultado entre las 3 empresas, y esto se traduce que a pesar de tener un NPS por tuit intermedio; esta empresa tiene una menor cantidad de usuarios que transmiten mensajes con sentimientos negativos.

Tabla 13. Resultados diccionario Affin aplicado a Vanti y cálculo del NPS por tuit y usuario

Etiqueta NPS	Número de Usuarios	Número de Tuits	NPS TUIT	NPS USUARIO
DETRACTOR	1453	710	-0.223	-0.217
NEUTRO	470	237		
PROMOTOR	850	408		
Total	2773	1355		

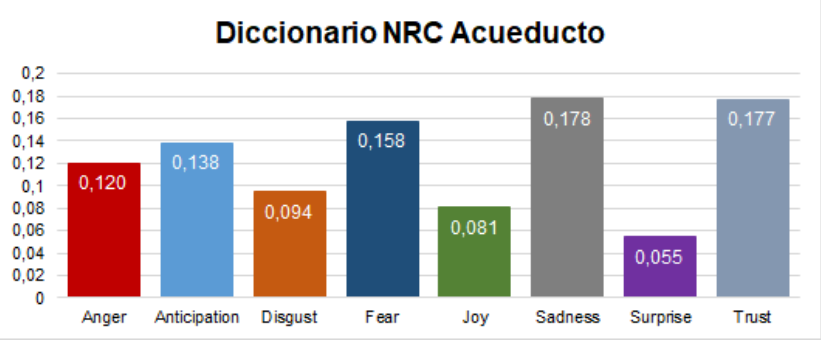
Fuente: Autoría Propia.

5.5.4.3 Diccionario NRC

Para obtener el resultado de la aplicación del diccionario NRC, se implementó el método *get_NRC_sentiment* que ofrece la librería *"syuzhet"* de R para cuantificar y asignar cada palabra de cada corpus a los 8 sentimientos que maneja este. Una vez todas las palabras fueron cualificadas, se suman los resultados y se totalizan para extraer la participación por sentimiento para cada empresa.

Para el acueducto, se obtiene la distribución del corpus a través de los sentimientos NRC como se muestra en la Figura 18; donde el 50% del corpus se clasifican en los sentimientos *Sadness, Trust* y *Fear*. Este alto porcentaje, indica que la mitad de las palabras que usan los usuarios para comunicarse con la empresa, transmiten sentimientos negativos y de malestar general. Un aspecto importante para resaltar es que los sentimientos positivos solo ocupan el 13% del corpus analizado, con lo cual se puede evidenciar la falta de empatía y apropiación que presentan los usuarios.

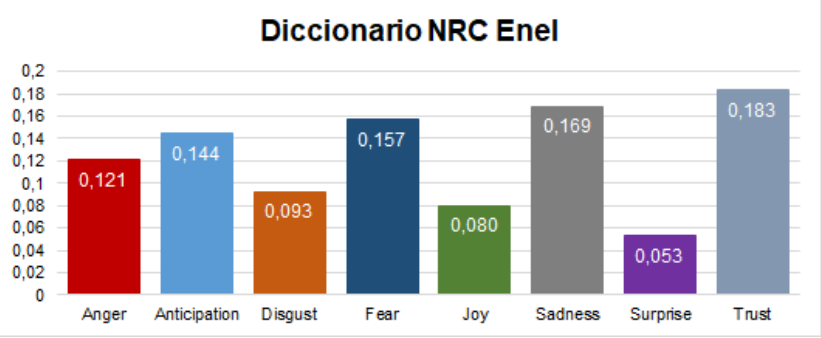
Figura 18. Proporción de sentimientos NRC asignados para los tuits del acueducto.



Fuente: Autoría Propia.

En el caso de Enel, en la Figura 19 se obtiene el resultado de la distribución del corpus en los 8 sentimientos. Al igual que en corpus del Acueducto, el 51% de las palabras están categorizadas en los sentimientos *Trust, Sadness* y *Fear* respectivamente; indicando que las palabras usadas por los usuarios en los mensajes dirigidos a la empresa transmiten sentimientos negativos y en general que la percepción del servicio no es muy buena. Detallando los sentimientos positivos, estos se encuentran en las dos últimas posiciones ocupando el 13% del total del corpus; con lo cual se infiere que los usuarios no tienen sentimientos buenos hacia el servicio prestado por la empresa y la imagen a nivel general.

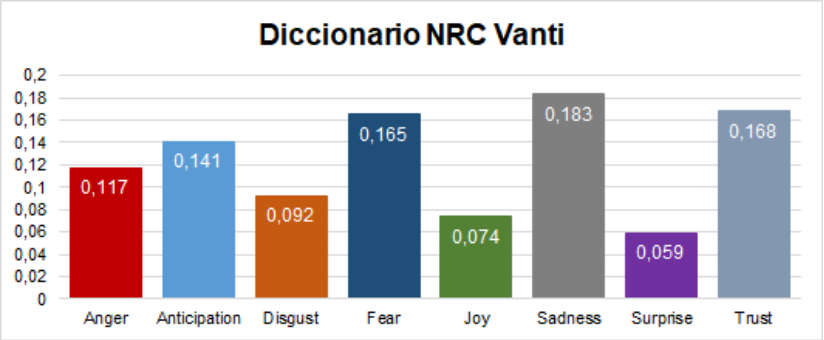
Figura 19. Proporción de sentimientos NRC asignados para los tuits de Enel.



Fuente: Autoría Propia.

Finalmente, en la Figura 20 se puede observar el resultado de la aplicación del diccionario NRC al corpus de Vanti; donde de igual manera los sentimientos que ocupan el 50% del corpus son *Trust, Sadness* y *Fear* respectivamente. Al igual que los corpus del Acueducto y Enel, los usuarios transmiten sentimientos negativos por medio de los mensajes dirigidos a la empresa y demuestra el malestar e inconformidad que tienen estos con los servicios que brinda la empresa. Detallando los sentimientos positivos, estos ocupan las últimas dos posiciones dentro de la participación con un 13%; lo que indica que los usuarios tienen sentimientos negativos con la empresa y sienten malestar con los servicios prestados y la imagen de esta.

Figura 20. Proporción de sentimientos NRC asignados para los tuits de Vanti.



Fuente: Autoría Propia.

Finalmente, a nivel general, se recalca la participación de la categoria *Anticipation*, cuyo porcentaje de asignación presenta un valor muy cercano al top 3 de los sentimientos de cada compañía; por lo cual puede pensarse que existen terminos compartidos en los tuits de los usuarios para las Enel, Vanti y El Acueducto que se utilizan para lograr obtener una respuesta a su mensaje de manera rapida y en la menor cantidad de interacciones posibles. Puede proponerse que este grupo esta estrechamente relacionado con peticiones de una urgencia alta que requieran una respuesta oportuna.

5.6 Comportamiento de tuits y evolución de los sentimientos a través del tiempo

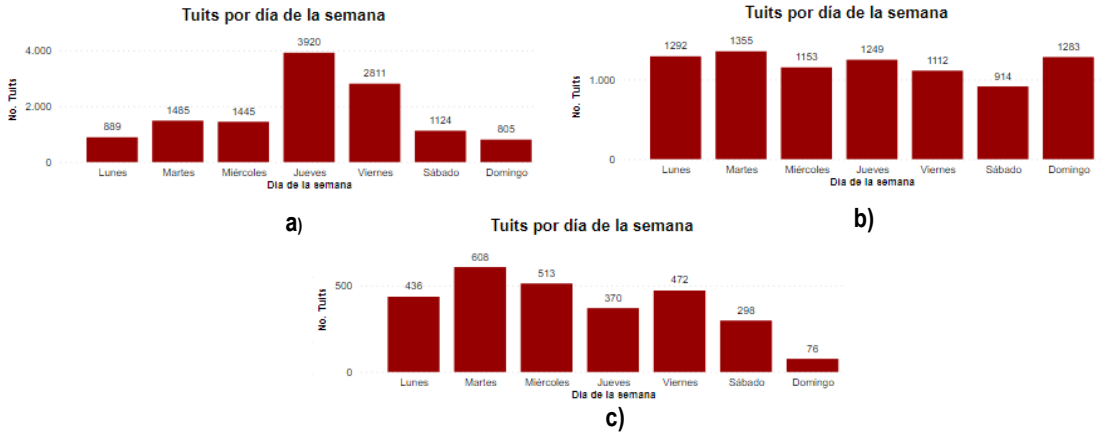
Analizar la evolución de los sentimientos a través del uso de cada uno de los diccionarios permite definir patrones, comportamientos, variaciones y hallazgos sobre los usuarios de Twitter. Para esto, se han realizado una serie de visualizaciones por entidad que contienen la información más relevante para la investigación durante el periodo analizado.

5.6.1 Líneas de tiempo

A continuación, se analizarán los datos recolectados a través de una serie temporal sobre todos los tuits descargados para cada entidad, por medio de lo cual se obtiene información sobre las tendencias y anomalías que se presentaron.

Para los datos recopilados se realizó un análisis para determinar qué día de la semana por empresa se presentaba alto tráfico de tuits como se puede visualizar en la Figura 21. En el caso del Acueducto, la tendencia que se puede evidenciar es que para los jueves y viernes se presenta el mayor volumen de mensajes (53,93%). Por otro lado, Enel presenta un comportamiento estable, los días que más tienen mensajes a través de Twitter son los días sábado, domingo y lunes (47.02%). Finalmente, para Vanti presenta un comportamiento más estable que el del Acueducto pero menos que el de Enel, los días que más tienen más tráfico en Twitter son los lunes, martes y miércoles (56.15%).

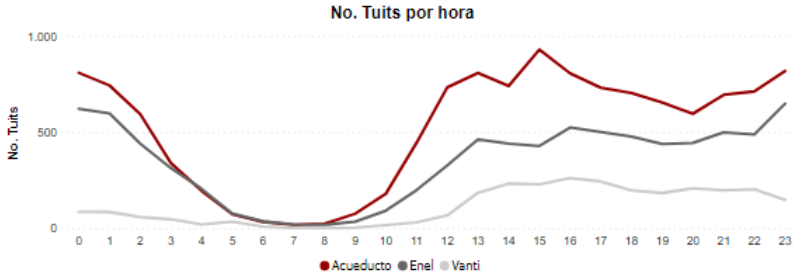
Figura 21. Cantidad de tuits por día de la semana para a) Acueducto, b) Enel y c) Vanti.



Fuente: Autoría Propia.

La Figura 22 corresponde al análisis por hora de cada uno de los tuits publicados para cada una de las entidades; se puede observar rápidamente las horas con mayor cantidad de interacciones en la red social. De acuerdo con los datos obtenidos, las horas en las que los usuarios más tuitean corresponden a las 0 horas hasta las 3 horas, desde las 10 horas hasta las 16 horas y desde las 20 horas hasta las 23 horas. Con lo anterior, se puede inferir que la mayor franja de tiempo de mensajes se ajuste acorde con los horarios de atención de los canales digitales de cada una de las empresas, además de presentar ciertas franjas nocturnas y de madrugada que demuestran la alta conectividad e interactividad que tiene la red social.

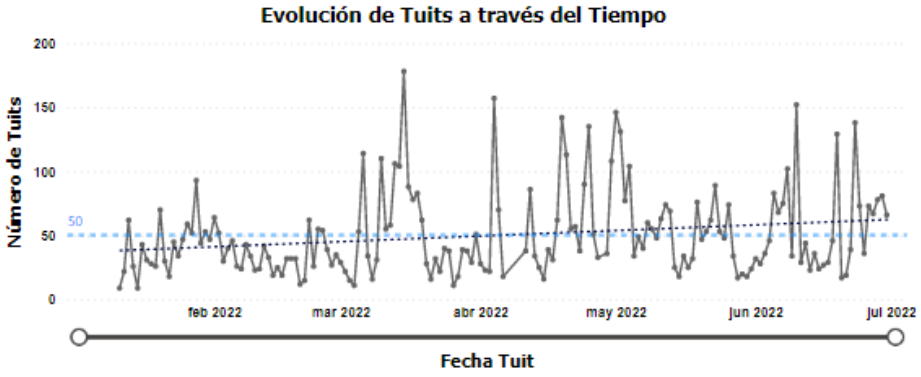
Figura 22. Evolución por hora del número de tuits por empresa



Fuente: Autoría Propia.

Al realizar el mismo análisis, pero llevado a nivel mensual, en el caso de Enel (Figura 23) se muestra que la media para el periodo de tiempo analizado está ubicada en 50 tuits diarios (línea azul celeste) con una leve tendencia al alza al finalizar el periodo de los datos obtenidos. En la serie temporal existen algunos valores muy por encima de la media durante los meses de abril y mayo (temporada de lluvias); los tuits correspondientes a esta franja se explican dada la alta cantidad de reportes relacionados con fallas masivas de la continuidad del servicio en localidades de Bogotá e incluso en los municipios aledaños a la ciudad. Además, se encuentran tuits asociados a las consecuencias y posibles efectos que puede tener la interrupción del servicio para conservar alimentos y funcionamiento de equipos vitalicios. Los siguientes puntos atípicos se presentan el 06 de junio, en el cual se reportaron intermitencias en el servicio en el municipio de Fusagasugá y en toda la zona occidental de Bogotá; con intervalos de corte de servicio superiores a 6 horas con fallas continuas. Además de percibir el malestar de los usuarios con estos eventos, se reportaban daños en los electrodomésticos y equipos electrónicos de los usuarios.

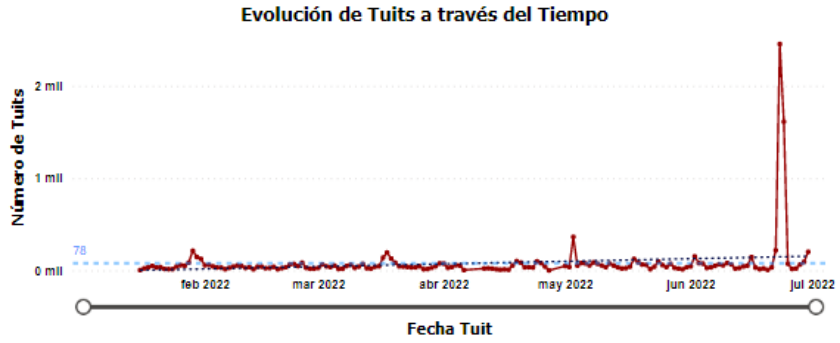
Figura 23. Evolución temporal de la cantidad de tuits recibidos por Enel



Fuente: Autoría Propia.

Como se evidencia en la Figura 24, el Acueducto presenta un par de picos de alta contactabilidad, los días 23 (2455 tuits) y 24 de junio (1613 tuits), un evento atípico ocasionado por el corte del servicio en la zona noroccidental de Bogotá donde los usuarios expresaban fallas del suministro del agua mayores a 72 horas y el déficit en la solución ofrecida por los carrotanques.

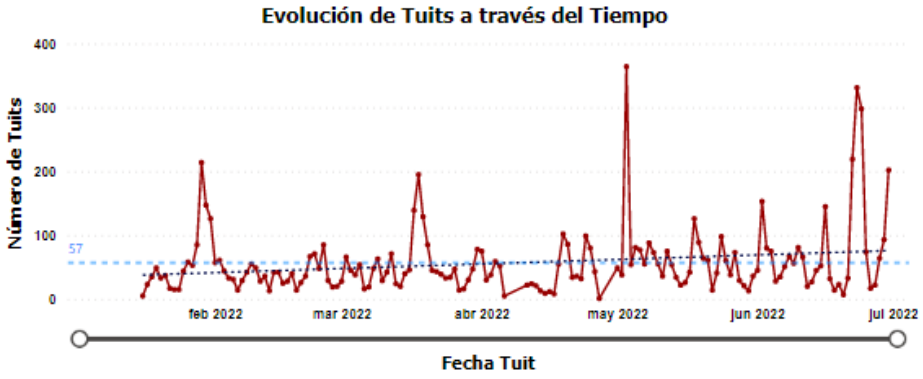
Figura 24. Evolución temporal de la cantidad de tuits recibidos por el Acueducto



Fuente: Autoría Propia.

Para poder trabajar con esta serie temporal, se realizó un promedio por derecha y por izquierda, el cual consistió en tomar la cantidad de tuits de los 7 días anteriores al 23 (incluyéndolo) para darle un valor normalizado; por otro lado, para el día 24 se tomó la cantidad de tuits de los 6 días posteriores a este, para asignarle un valor promedio al segundo día. Los resultados se pueden observar en la Figura 25:

Figura 25. Evolución temporal de la cantidad de tuits recibidos por el Acueducto ajustado



Fuente: Autoría Propia.

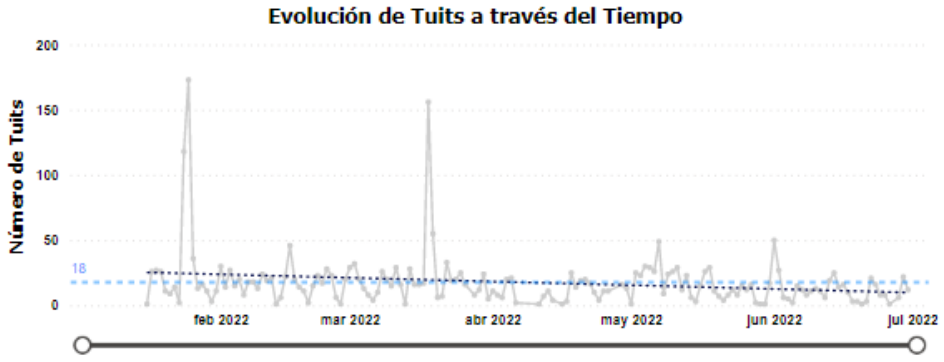
Con los datos normalizados, se puede visualizar la variación de la cantidad de tuits en el periodo de tiempo estudiado. El ajuste de los dos valores atípicos permitió disminuir la media de 78 a 57 tuits diarios, mostrando una leve tendencia al alza de tuits al finalizar el periodo. Un hallazgo importante a nivel visual está

relacionado con una posible estacionalidad en la que los usuarios tuitean más que los otros días, detallando más a profundidad, cada 47 días esta empresa tendrá un día de alto tráfico, superando las 200 comunicaciones en cada pico.

En la Figura 25, se pueden visualizar otros días de alta volumetría, el primer pico que corresponde al 29 de enero se presenta un problema de cortes y calidad de agua en Usme, además de comunicar problemas de movilidad dada la intervención por obras en las redes. Para el día 18 de marzo, existen inconvenientes con en el suministro del servicio por más de 72 horas en la localidad de Chapinero, haciendo énfasis en que los tiempos que prometía la empresa para reestablecer el mismo no los habían cumplido y las afectaciones estaban siendo críticas para los usuarios. Finalmente, para el día 03 de mayo, también hubo una problemática de cortes del servicio en la localidad de Fontibón, a lo cual los usuarios comunican que la empresa omitió realizar los avisos previos a la fecha de corte, ese mismo día, pero en Usaquén, se presentó dado los problemas a nivel de movilidad por los trabajos en mantenimiento de la red fluvial y de alcantarillado que no fueron avisados.

Finalmente, para Vanti se puede observar en la Figura 26 un promedio de 18 tuits diarios para el periodo analizado con una tendencia a la baja al finalizar el periodo estudiado. De acuerdo con el comportamiento de los datos recopilados, se puede inferir que el canal no tiene un alto uso a comparación de las otras dos empresas de servicios públicos de la ciudad.

Figura 26. Evolución temporal de la cantidad de tuits recibidos por Vanti



Fuente: Autoría Propia.

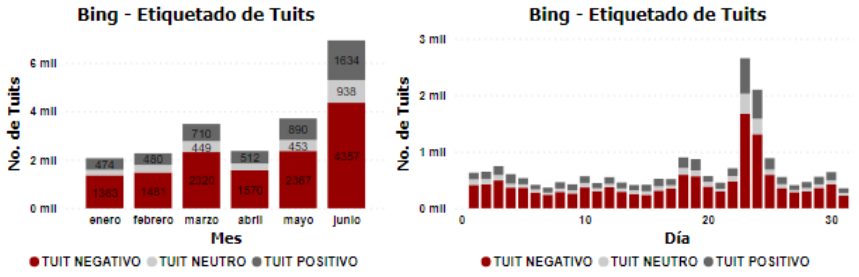
A partir de los puntos atípicos que se observan en la gráfica, se puede inferir que hay una leve estacionalidad promedio de 25 días calendario, en los que se esperaría recibir una demanda mayor de mensajes por este canal. Para el día 25 de enero, los tuits están relacionados con facturación, cortes de servicio sin previo aviso y costos asociados a la reconexión de este. Por otro lado, el día 18 de marzo se presenta una alta cantidad de solicitudes relacionadas con reconexión, visitas y respuestas a peticiones de

los clientes por vencimiento de tiempo; además de los costos que asumen los usuarios por estas causales, generando malestar general. Para el día 7 de mayo, se presenta una cantidad de mensajes atípica, relacionados con el servicio al cliente por medio de sus canales de comunicación y las visitas para reconexión e instalación de servicios, donde según lo comunicado por los usuarios el servicio al cliente es deficiente respecto a tiempos de espera y respuestas obtenidas, además del incumplimiento de las visitas domiciliarias para revisión o instalación del servicio.

5.6.2 Evolución de sentimientos mediante BING

En términos generales, de acuerdo con los resultados obtenidos en la Figura 27sobre la evolución temporal de los sentimientos aplicados al diccionario BING, se obtuvo que para todo el corpus analizado se obtuvieron 20817 tuits categorizados (88.67%) y 2659 (11,32%) tuits sin asignación, donde los primeros se distribuyen así: 13458 tuits negativos (64,65%), 2659 tuits neutros (12,77%) y 4700 tuits positivos (22,58%). Dentro del periodo de tiempo analizado, se puede observar que para cada uno de los meses la mayor proporción de tuits se perciben como negativos y esto es ocasionado por la cantidad de palabras que utilizan los usuarios de Twitter que están catalogadas que expresan sentimientos negativos. Analizando a detalle la evolución del diccionario a través de cada día calendario en que se pública, se puede identificar que desde los días 18 al 30 el tráfico de mensajes aumenta considerablemente respecto al resto de los días calendario, esto puede estar relacionado con los cortes de facturación y la suspensión de servicios. Tomando en cuenta lo anterior, también se puede identificar que la proporción de tuits negativos por día calendario son en promedio mayor al 50%, lo cual secunda que la percepción a nivel general de las empresas es negativa.

Figura 27. Evolución temporal de los tuits por mes y día calendario con la aplicación del diccionario BING consolidado para las tres empresas

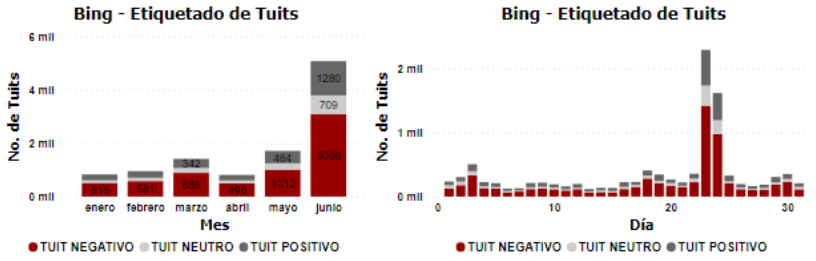


Fuente: Autoría Propia.

Para el Acueducto, durante el periodo analizado se obtuvieron 10802 tuits etiquetados por el diccionario (86,56%) y 1677 tuits descartados o sin categoría (13,43%). Dentro de los hallazgos, se puede identificar que el mes de junio dado que se duplican y hasta triplican las cifras de tuits negativos, lo cual está

asociado a cortes del servicio, lo cual se explica en el apartado anterior. Detallando los resultados de la Figura 25 por día calendario, el día 23 y 24 fueron los que más se tuiteo; además de evidenciar que la tendencia diaria es que los tuits negativos superen el 50% del valor total de los tuits, y esto predomina y es congruente con los datos obtenidos a nivel general.

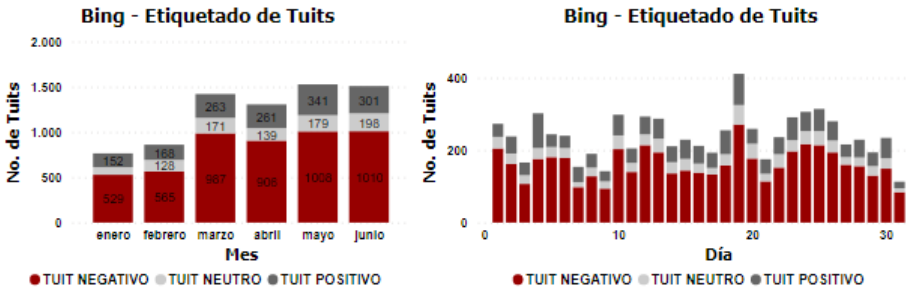
Figura 28. Evolución temporal de los tuits por mes y día calendario con la aplicación del diccionario BING al acueducto.



Fuente: Autoría Propia.

En el caso de Enel, los resultados obtenidos en la Figura 29, indican que durante el periodo analizado se obtuvieron 7390 tuits etiquetados por el diccionario (88,42%) y 968 tuits descartados (11,58%). Los tuits etiquetados se distribuyen de la siguiente forma: 5005 Negativos (67,72%), 899 tuits neutros (12,17%) y 1486 tuits positivos (20,11%). Dentro de los hallazgos, se puede identificar que durante los meses de marzo a junio la cantidad de tuits etiquetados como negativos es constante a nivel proporcional, ocupando en promedio 68% de la cantidad total de mensajes por mes. En cuanto a los resultados obtenidos por día calendario, se observa que el comportamiento de las comunicaciones corresponde a una funcional senoidal normalizada con baja variación, donde los días que contienen la mayor cantidad de tuits son del día 18 al 26 de cada mes.

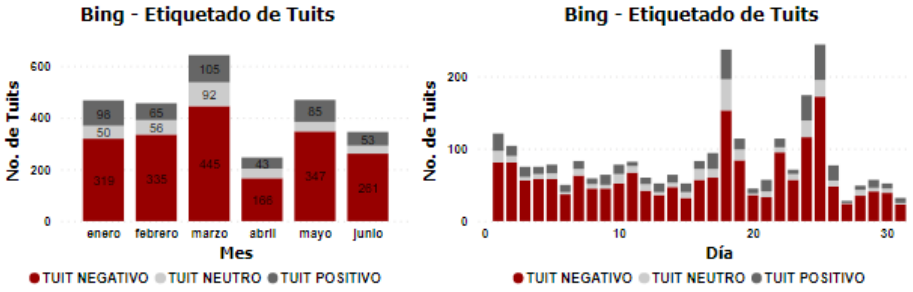
Figura 29. Evolución temporal de los tuits por mes y día calendario con la aplicación del diccionario BING a Enel.



Fuente: Autoría Propia.

Para Vanti se obtuvieron los resultados que se muestran en la Figura 30, durante el periodo analizado se obtuvieron 2623 tuits etiquetados por el diccionario (94,6%) y 150 tuits descartados (5,4%). Los tuits etiquetados se distribuyen de la siguiente manera: 1873 tuits negativos (71,41%), 303 tuits neutros (11,55%) y 447 tuits positivos (17,04%). Dentro de los hallazgos evidenciados, se puede identificar que el mes que más obtuvo tuits negativos fue el mes de marzo y que el mes que presenta una menor cantidad de tuits negativos e inclusive a nivel general para las 3 compañías es el mes de abril. Los resultados obtenidos por día calendario, muestran que los días 18, 24 y 25 son los que se presenta mayor cantidad de tuits y que el resto de las fechas se presentan un comportamiento en general estable, con baja variabilidad, pero donde predomina la cantidad de tuits negativos. Eliminando los datos atípicos del día calendario, se obtiene un comportamiento senoidal de los datos, con una estacionalidad de 8 a 10 días y una tendencia a la baja que puede observar a final del último mes, pero identificando claramente los días de alto y bajo tráfico con una alta carga de mensajes negativos.

Figura 30. Evolución temporal de los tuits por mes y día calendario con la aplicación del diccionario BING a Vanti.



Fuente: Autoría Propia.

5.6.3 Evolución de sentimientos mediante AFFIN

Para analizar el resultado del diccionario Affin, se obtiene por promedios el valor de asignación de cada palabra para consolidar el puntaje de cada tuit, paso seguido, se promedian las puntuaciones de cada mensaje por fecha de publicación y así finalmente poder obtener la valoración final diaria que cuantifica la intensidad de los sentimientos.

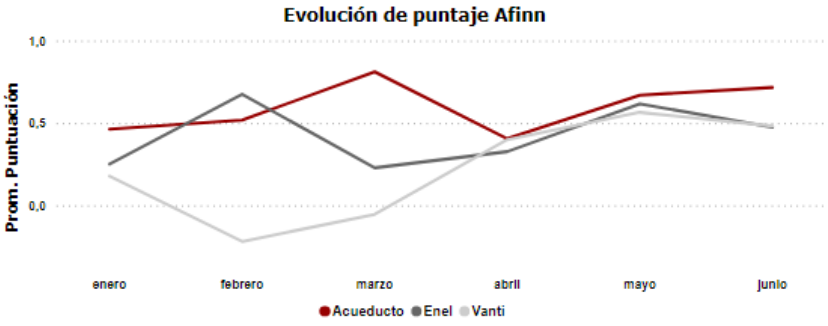
Los hallazgos obtenidos que se visualizan en la Figura 31 muestran que, en el caso del acueducto, la cuantificación de la intensidad de los sentimientos presenta resultados positivos a lo largo del periodo analizado con: promedio superior al 0.5, un máximo en marzo de 0.81 y en los meses de mayo y junio una

tendencia al alza (de acuerdo con la escala, la intensidad de sentimiento al ser positiva indica que la percepción de estas son palabras positivas).

En el caso de Enel, la intensidad de los sentimientos muestra resultados positivos a lo largo del periodo analizado con un promedio superior a 0.4 y un valor máximo del 0.68 en el mes de febrero, resaltando que de marzo a mayo este puntaje tiende al alza y disminuye en junio. En el caso de Vanti, la intensidad de los sentimientos muestra un puntaje promedio de 0.2, febrero y marzo relacionan valoraciones negativas; además que, desde el segundo mes del estudio hasta el quinto, existe una tendencia al alza que mejora el resultado final para esta entidad.

Otro de los hallazgos a resaltar es la relación inversa que presenta los valores obtenidos por Enel y Vanti en el mes de febrero, donde la primera empresa obtiene su mejor puntaje y la segunda su peor puntuación en todo el periodo analizado. De igual manera sucede con el Acueducto y Enel para el mes de marzo, donde el Acueducto obtiene su mejor puntaje y Enel obtiene el peor puntaje. Estas relaciones inversas están asociadas con la cantidad de comunicaciones que se obtuvieron de cada entidad mensualmente, además de la cantidad de palabras que logró identificar el diccionario para asignar la puntuación a cada palabra de cada tuit. Por último, cabe resaltar que para las tres empresas se tiene una tendencia al alza durante los meses de abril a junio terminando con un puntaje favorable cercano a 0.5 o mayor.

Figura 31. Evolución temporal de las tres empresas aplicando la intensidad de sentimientos del diccionario Affin



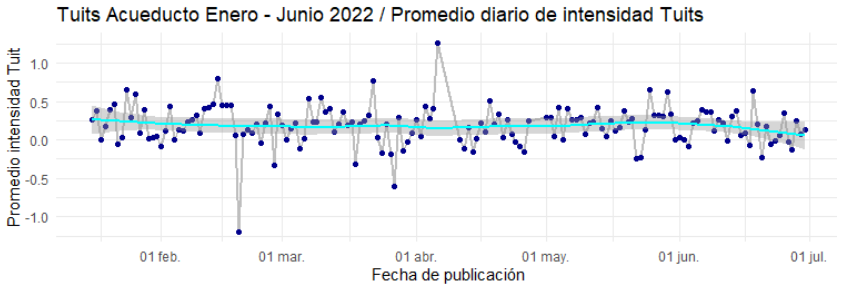
Fuente: Autoría Propia.

Para realizar un análisis temporal de la evolución de la intensidad de las palabras a través del diccionario Affin, se extrae el promedio diario de las palabras usadas para cada una de las empresas. Además, a cada una de las gráficas obtenidas se les incluyeron la suavización de curvas de datos mediante el método *"Loess",* la cual es un tipo de regresión local que se utiliza para para ajustar regresiones múltiples mediante vecinos cercanos donde se caracteriza dado que entre más grande sea el intervalo de datos analizados, más suave será la curva resultante (R Core Team, 2021).

Para el Acueducto, en la Figura 32 se puede observar a través de la serie temporal que el valor promedio diario de la intensidad de las palabras no presenta un patrón establecido, existiendo una variabilidad moderada dado los valores atípicos de la serie a nivel positivos como negativos, obtenidos exactamente en la siguientes fechas: el 15 de febrero con +0.76, el 21 de febrero con -1.16, el 1 de marzo con -0.36, 24 de marzo con -0.71 y el 8 de abril con +1.26.

Como lo muestra la serie, estos datos son altamente variables hasta mediados del mes de abril y de allí en adelante se comportan de una manera estable e incluso disminuyendo la desviación conforme avanzan los días. Con la curva suavizada, se puede observar que su tendencia es hacia el 0 iniciando en el intervalo positivo y terminando a la baja incluso pasando al intervalo negativo; teniendo baja desviación de acuerdo con lo que se observa en los límites máximos y mínimos de la curva y observando leves variaciones en la misma cuando se encuentra con los datos atípicos de la serie.

Figura 32. Evolución temporal del puntaje promedio diario del diccionario Affin para Vanti

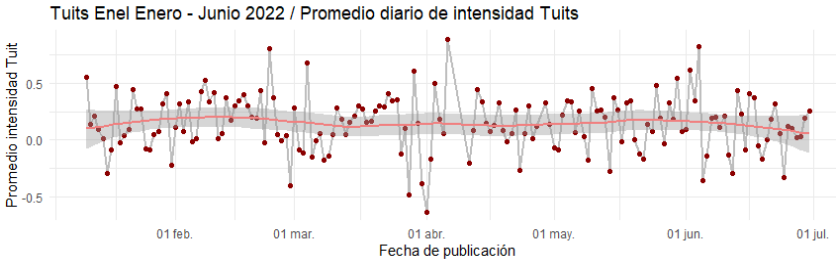


Fuente: Autoría Propia.

En la Figura 33 se puede observar el comportamiento diario del puntaje promedio de la intensidad de las palabras usadas en los tuits de la empresa Enel. Se puede ver que, en la mayoría de los días, se obtiene un puntaje superior a 0; lo cual indica que las palabras usadas por los usuarios son levemente positivas, aunque están cercanas al 0 por lo cual no se puede decir que los mensajes analizados sean positivos.

La suavización de la serie de datos permite mostrar la línea promedio y los intervalos máximos y mínimos para los datos mediante lo cual, se concluye que se presenta baja variabilidad y pocos valores atípicos, existiendo un valor máximo de 0.82 el día 6 de abril y un valor mínimo de -0.64 el día 1 de abril.

Figura 33. Evolución temporal del puntaje promedio diario del diccionario Affin para Enel

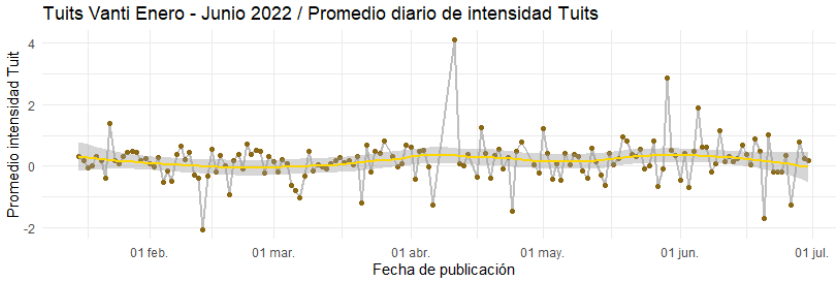


Fuente: Autoría Propia.

Finalmente para Vanti, en la Figura 34 se puede observar que dentro de los resultados obtenidos hay una variabilidad baja, a excepción de algunos valores atípicos como el 2 de abril en el cual este día se obtiene la puntuación máxima en la escala Affin positiva, así como el 26 de mayo que presenta una puntuación de +3,05 y el 3 de junio con una puntuación de +2,1. Analizando los valores negativos de la serie, se puede observar una estacionalidad promedio de 20 días en que se encontrará un valor máximo local en los datos; por lo que se puede decir que, los mensajes que contienen un alto contenido de palabras con intensidad negativa se repite a través del tiempo.

Finalmente, dentro de la curva suavizada, se encuentran valores en promedio cercanos al 0 pero con tendencia hacia el intervalo positivo además de evidenciar una baja desviación, con lo cual se infiere que, los datos tienen un comportamiento estable y neutral con el mensaje que transmiten los usuarios a través de esta red social.

Figura 34. Evolución temporal del puntaje promedio diario del diccionario Affin para Vanti



Fuente: Autoría Propia.

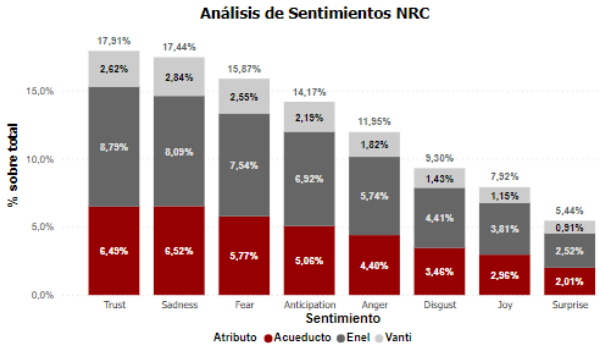
5.6.4 Evolución de sentimientos mediante NRC

Para analizar la evolución a través del tiempo para el diccionario NRC, se totaliza la cantidad de palabras asignadas a cada uno de los 8 sentimientos a partir de la frecuencia de aparición de cada vocablo

en cada tuit. Obteniendo un porcentaje de asignación de cada mensaje en cada categoría predefinida por el diccionario.

En la Figura 35 se observa el resultado de la participación porcentual de cada sentimiento por empresa, donde se resalta que el sentimiento con mayor participación en general es *Trust*, esto dado que para Enel la mayoría de las palabras utilizadas transmiten sentimientos asociados a confianza de acuerdo con los tuits analizados; caso contrario sucede con el Acueducto y Vanti, donde el sentimiento con una participación mayor para cada una es *Sadness*. El tercer sentimiento con mayor participación a nivel general corresponde a *Fear*, estas tres categorías corresponden al 51,22% de comunicaciones, con lo que se puede inferir que los usuarios que tuitean a las empresas demuestran sentimientos negativos, asociando el contenido de los mensajes; se infiere que estos sentimientos fueron generados por todos los problemas de servicio, cortes y reconexiones que comunican los usuarios a través de esta red social.

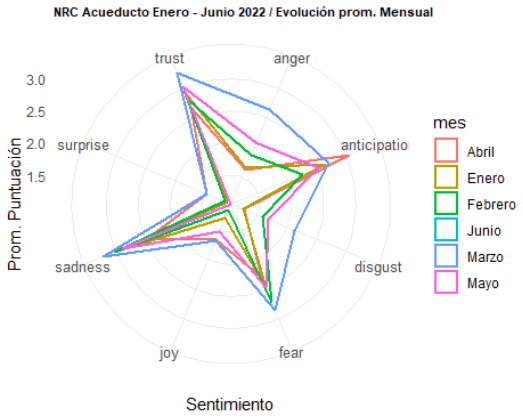
Figura 35. Participación de las tres empresas sobre el total de cada sentimiento del diccionario NRC



Fuente: Autoría Propia.

Como se muestra en la Figura 36, para el Acueducto se puede identificar que *Trust, Sadness* y *Fear* representan en general los sentimientos de los usuarios hacia la empresa; esto dado que mes a mes son los que tienen un puntaje superior. Un aspecto importante para resaltar se presenta en abril y mayo con el sentimiento *Anticipation*, ya que se obtiene un valor atípico de acuerdo con el comportamiento y su distribución mensual. Finalmente, los sentimientos positivos como son *Surprise* y *joy* son las categorías más bajas para el periodo de tiempo analizado y esto permite inferir que los mensajes transmiten sentimientos negativos y se asocian al malestar que la empresa le genera a los usuarios en la prestación del servicio.

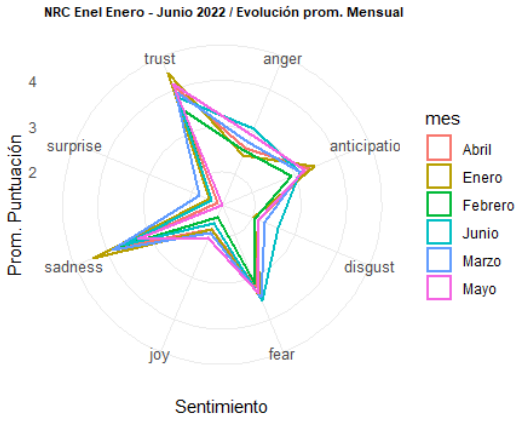
Figura 36. Evolución mensual de los sentimientos asociados al diccionario NRC aplicado al acueducto



Fuente: Autoría Propia.

En la Figura 37 se puede observar la evolución de los sentimientos NRC para Enel. Se evidencia un comportamiento estable mes a mes durante el periodo de tiempo analizado, a partir de lo cual se identifican que los sentimientos predominantes son: *Trust, Sadness, Fear, Anger* y *Anticipation.* Dentro de los anteriores, hay un sentimiento positivo que es la confianza; lo cual se asocia a que los usuarios esperan y confían en que la empresa les brinde una pronta solución, sin embargo; al predominar *Sadness y Fear* como las demás categorías que componen el top 3 de sentimientos, se infiere que existe una insatisfacción por parte de los usuarios, lo que ocasiona un aumento en la cantidad de los mensajes recibidos en los que la percepción a nivel general no es muy buena.

Figura 37. Evolución mensual de los sentimientos asociados al diccionario NRC aplicado a Enel.

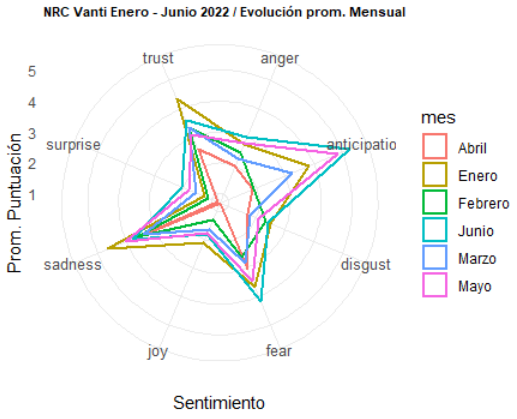


Fuente: Autoría Propia.

Para Vanti, en la Figura 38 se puede observar la evolución de cada uno de los sentimientos NRC a través del tiempo; donde se identifica un comportamiento variable para las categorías de: *Anticipation, Anger* y *Trust* en los meses: enero, marzo, mayo y junio. Los sentimientos con mayor puntuación para todos los meses son *Trust, Anticipation, Fear* y *Sadness*; y esto permite inferir que los usuarios sienten insatisfacción con el servicio prestado, además con todos los temas relacionados con pagos, facturación y reconexión, lo cual es secundado al revisar lo concluido a partir de los demás diccionarios.

Por otro lado, estos sentimientos indican que los usuarios al enviar los mensajes tratan de ser directos y concisos con la información suministrada para evitar demoras en la respuesta que da la empresa a sus solicitudes; lo que deja ver que esta red social como canal de atención de usuarios para la empresa presenta altas deficiencias y no es efectiva para su objetivo.

Figura 38. Evolución mensual de los sentimientos asociados al diccionario NRC aplicado a Vanti.



Fuente: Autoría Propia.

5.7 Topic modeling

En apartados anteriores se han presentado las palabras con una mayor frecuencia de aparición para cada uno de los corpus, sin embargo, se considera altamente relevante para la consecución del objetivo, identificar con mayor detalle los temas sobre los cuales los clientes de las compañías en cuestión están hablando en Twitter. Para llegar a esto, se propone la aplicación del algoritmo LDA, el cual permite tener un listado de palabras por cada temática basado en la probabilidad de que dichos términos aparezcan juntos en un tuit.

Este análisis se realiza de manera masiva y automática (sin entrar a detallar cada tuit de cada cliente), encontrando en primera medida, la cantidad de tópicos; luego, los centroides alrededor de los cuales se agrupan los diferentes grupos de temas y finalmente las listas para dar el nombre a cada grupo de palabras.

Se utilizan entonces las librerías: "*reshape2" y "topicmodels"* para realizar este procedimiento del software R (R Core Team, 2021).

Lo anteriormente descrito, puede encontrarse en las secciones del código: *k optimo* y *topic modeling* respectivamente. La primera etapa involucra el paquete "ldatuning" para determinar de manera grafica un numero optimo - k - de tópicos a trabajar por cada corpus, seguidamente, se utiliza la función "LDA" para encontrar el listado de palabras asociado a cada tema, es importante mencionar que, los principales insumos de esta son: el "dtm" de cada corpus (descrito en la sección 5.3.2.5) además del número k de temas que se definen previamente. Finalmente, a partir de las listas de palabras se identifica en conceso por los investigadores el tema del cual se está refiriendo, asignándole un nombre a cada una.

5.7.1 Numero óptimo de tópicos por empresa

En esta sección se presentan los resultados para cada una de las empresas relacionadas a la investigación, aplicando la función "FindTopicNumbers" del paquete "ldatuning" de R (R Core Team, 2021). Es importante aclarar que el método utilizado para determinar el número óptimo de tópicos es uno de los tantos que existe (algunos de los que se encuentran en la literatura son elbow/knee point, perplexity entre otros); sin embargo se ha definido utilizar este dado su facilidad para interpretar los resultados, además de involucrar 4 métricas de comparación: "Griffiths2004", "CaoJuan2009", "Arun2010", "Deveaud2014"; donde el resultado obtenido son un par de graficas que relacionan las métricas anteriormente mencionadas y que corresponda al valor mínimo en uno de los esquemas (en donde se representan las métricas Arun2010 y CaoJuan2009) y al máximo en el otro (en donde se representa las métricas Griffiths2004 y Deveaud2014).

Un aspecto importante para tener en cuenta al momento de evaluar los resultados es que se opta por definir un numero de tópicos (k) bajo en el cual no se presenten "cruces de tópicos"; es decir, que las palabras con una probabilidad de aparición mayor sean excluyentes en cada tema, para que cada conjunto tenga una interpretación sencilla y clara; además de garantizar que los tópicos tengan sentido y por ende las palabras que componen cada uno tengan relación y conexión.

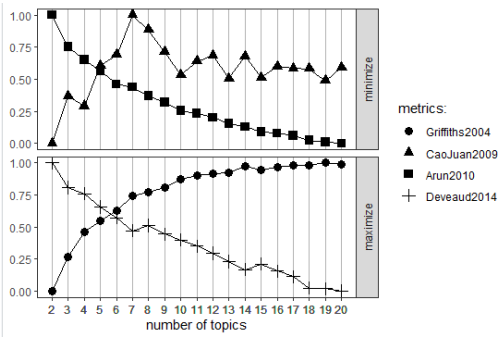
5.7.2 Resultados Tópicos

Encontrar los resultados de los tópicos óptimos para cada empresa, permite identificar los temas de los cuales está hablando cada audiencia a través de esta red social; y con esto relacionarlos a aspectos positivos y negativos sobre las necesidades transmitidas.

5.7.2.1 Acueducto

Como se puede visualizar en la Figura 39, la cantidad de tópicos que minimiza las métricas de CaoJuan2009/Arun2010 y máxima a su vez Griffiths2004/Deveaud2014 es 5 o 6.

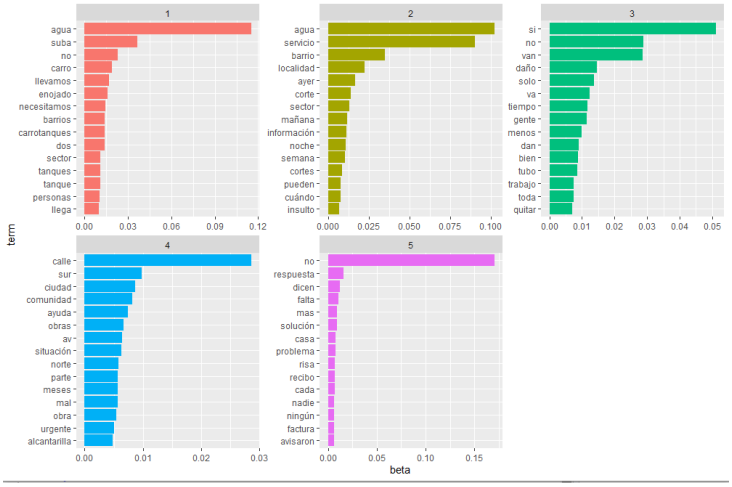
Figura 39. Número de tópicos – Acueducto



Fuente: Autoría Propia.

Al ejecutar el algoritmo con un k=6 se encuentra que dos tópicos se parecen mucho, puntualmente en los que se habla sobre cortes en el servicio, por lo cual, se opta por realizar la división en 5 temas diferentes. Estos se muestran en la Figura 40.

Figura 40. Top words por tópico – Acueducto



Fuente: Autoría Propia.

En este orden de ideas, y después de realizar un análisis de los listados de las palabras de cada tópico, se definen los siguientes nombres u etiquetas para cada grupo. En este caso, se resalta que

aparecen temas “técnicos” tales como: obras y daños en infraestructura. Así mismo, hay un tópico uno para solicitud de carrotanques derivado de los cortes en el servicio.

Tabla 14. Asignación de nombres por tópico - Acueducto

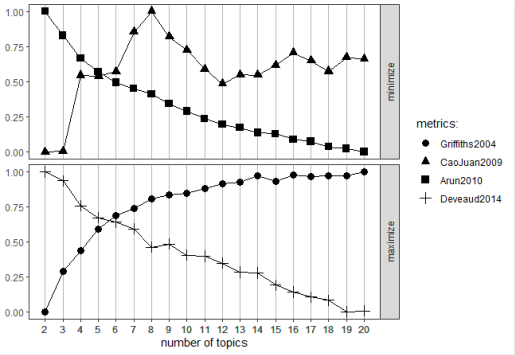
Tópico	Etiqueta
1	Solicitud de Carrotanques
2	Cortes en el servicio
3	Daños en Infraestructura
4	Obras en vía publica
5	Solución de caso e Información de solicitudes

Fuente: Autoría Propia.

5.7.2.2 Enel

Como se puede observar en la Figura 41, la cantidad de tópicos que minimiza las métricas de CaoJuan2009/Arun2010 y máxima a su vez Griffiths2004/Deveaud2014 es 5 seguidamente de 6.

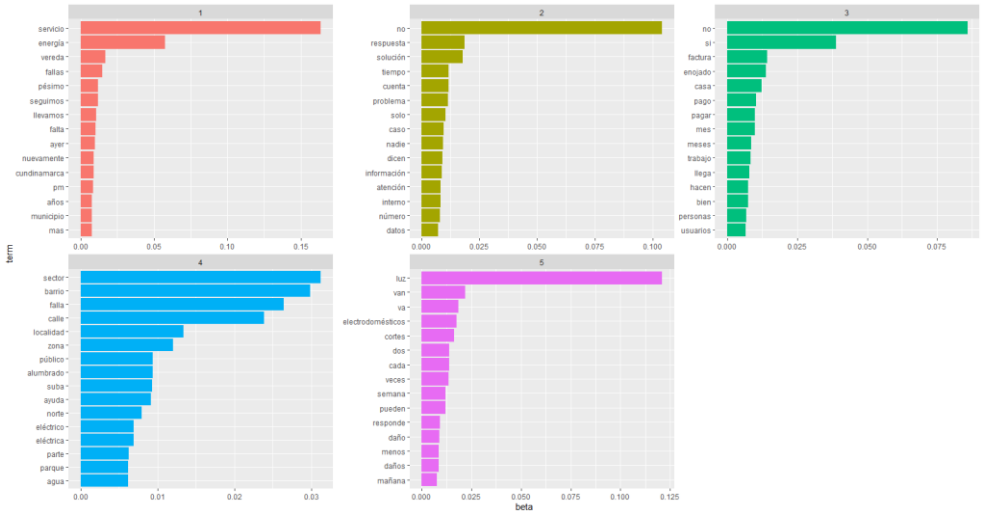
Figura 41. Número de tópicos – Enel



Fuente: Autoría Propia.

Sin embargo, al compararse los listados de palabras; con 6 tópicos se evidencia que un grupo no tiene mucho sentido dado que las palabras no tienen relación evidente y, por lo tanto, se opta por utilizar un k igual a 5. El resultado de las 15 palabras con una probabilidad mayor de aparición en conjunto se muestra en la Figura 42:

Figura 42. Top words por tópico – Enel



Fuente: Autoría Propia.

Finalmente, a partir del análisis de cada grupo de palabras identificados anteriormente se definen los siguientes tópicos de interés, asignándole a cada uno su etiqueta respectiva:

Tabla 15. Asignación de nombres por tópico - Enel

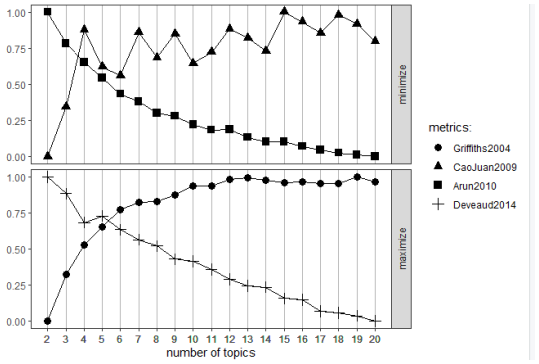
Tópico	Etiqueta
1	Cortes en el servicio
2	Solución de caso e Información de solicitudes
3	Facturación
4	Alumbrado publico
5	Daños en Electrodomésticos

Fuente: Autoría Propia.

5.7.2.3 Vanti

Como se puede observar en la Figura 43, la cantidad de tópicos que minimiza las métricas de CaoJuan2009/Arun2010 y máxima a su vez Griffiths2004/Deveaud2014 es 5 o 6.

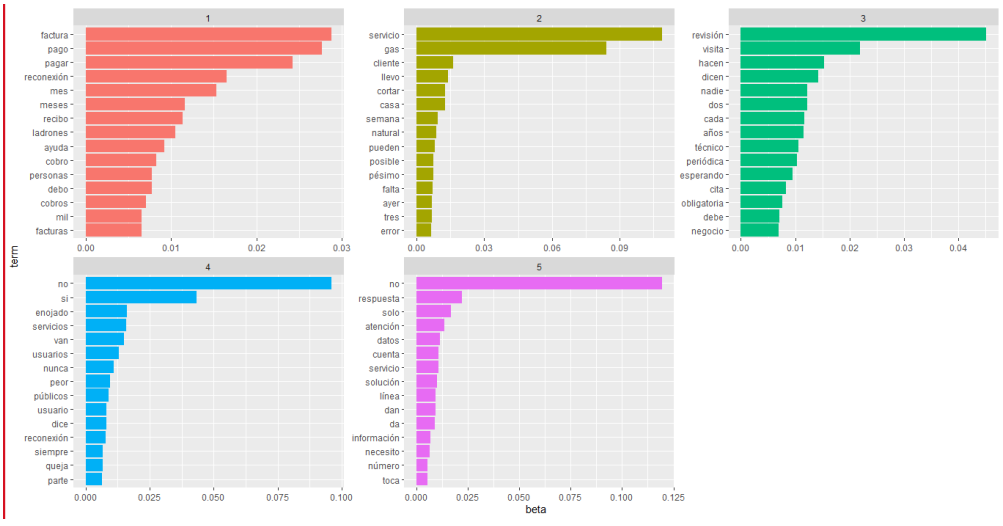
Figura 43. Número de tópicos – Vanti



Fuente: Autoría Propia.

Al realizar el análisis del listado de palabras al ejecutar la función LDA, se encuentra que con un k=6 existen tópicos que pueden compartir palabras y, por tanto, tener una interpretación similar, por lo cual se realiza el ejercicio con 5 tópicos. El resultado de las 15 palabras con una probabilidad mayor de aparición en conjunto se muestra en la Figura 44; donde posteriormente y a partir de estas se definirán las etiquetas sobre los temas de interés que habla la audiencia.

Figura 44. Top words por tópico – Vanti



Fuente: Autoría Propia.

A partir de lo anterior, se definen los nombres de los grupos que se muestran en la Tabla 16. Es importante aclarar la diferencia entre el tópico 4 y 5, la cual radica en que el conjunto 4 incluye palabras como: “petición”, “problema”, “denuncia”, “super intendencia”, “razón” entre otros, lo cual hace pensar que se asocia a Peticiones, quejas y reclamos (en especial del tema de reconexión), por su parte, el tópico 5 incluye términos como: “información” , “atención” , “solución” , “numero” , “clientes” , “responden” y “oficina”, por lo cual, se piensa que se vincula más a temas de información de solicitudes y atención de las mismas.

Tabla 16. Asignación de nombres por tópico - Vanti

Tópico	Etiqueta
1	Facturación
2	Cortes de servicio
3	Revisión en terreno
4	Mala atención – PQRS
5	Solución de caso e Información de solicitudes

Fuente: Autoría Propia.

5.7.3 Asignación de tópico a cada tuit de cada empresa

Una vez se tienen todas las listas de palabras por tópico, se realiza un consolidado por cada una de las empresas en un archivo .xlsx, de manera que pueda ser cargado en R (R Core Team, 2021) con el fin de realizar la asignación por tema a cada uno de los tuits de los usuarios. De esta manera y como se explica más adelante, se podría evidenciar la evolución de las temáticas en el tiempo. A la par de lo anterior, es necesario tener también otro archivo .xlsx con las probabilidades de aparición de cada palabra en cada tópico.

Cada tuit puede tener una gran variedad de palabras, pudiendo no todas pertenecer al mismo tópico, por lo que, para efectos del presente análisis, se realiza una ponderación entre la cantidad de palabras en cierto tuit multiplicada por la probabilidad de aparición de cada una de ellas en cada tópico. De esta manera, se puede obtener la prevalencia de un único tema en cada tuit (o el tópico dominante), es decir, determinar cuantitativamente a que temática pertenece cada mensaje de cada usuario, basado en la probabilidad de aparición de las palabras utilizadas en cada tópico.

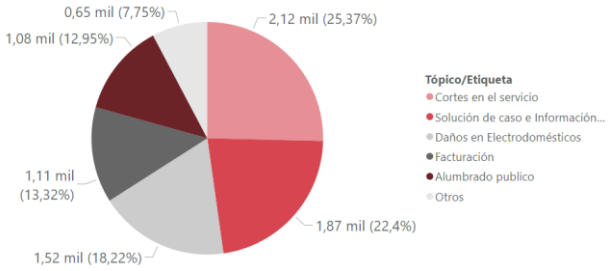
Finalmente, una vez se tiene el tópico dominante de cada tuit, es necesario realizar un cruce entre la etiqueta y el nombre que previamente se asignó a cada conjunto de temas de cada empresa. Todo lo anterior se puede encontrar en los apartados: “Listado de palabras”, “Ciclo para asignación a tópicos”, “Ajuste asignación de frecuencia de términos”, “agrupación de puntajes por tópico” y “Asignación de etiquetado”,

permitiendo consolidar una única base de datos con los campos de interés. (En el código se encuentra como: df_Cruzado).

Es importante mencionar que existen ciertos tuits, en los cuales los términos o palabras utilizadas por el cliente no se encuentran dentro de los listados de términos asociados a cada tópico, por lo cual, no pueden ser categorizados dentro de las 5 temáticas respectivas de cada empresa. Estos mensajes, se asignan a la clasificación "otros" (con el fin de continuar con la premisa de analizar la totalidad de las comunicaciones enviadas por los clientes a través de la red social). Por lo anterior, en los tableros que se presentan en las siguientes secciones, se encuentra una categoría más adicional a las 5 definidas previamente.

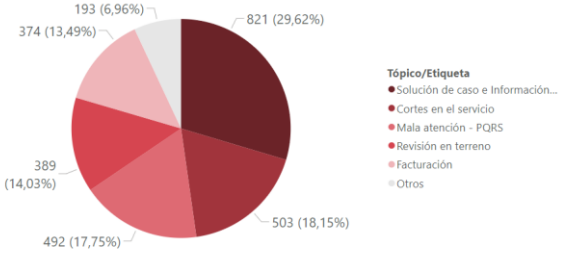
Volviendo a los resultados obtenidos, se evidencia que para las tres compañías una gran cantidad de tuits se categoriza dentro de alguno de los cinco tópicos correspondientes, en promedio menos del 11% de las comunicaciones se asignan a la etiqueta "otros". Por otro lado, para las 3 compañías, hay un tema predominante que permite clasificar cerca de un 25% del total de tuits, para Enel corresponden a cortes en el servicio, para Vanti corresponde a solución de caso e información de solicitudes finalmente para el Acueducto corresponde a solicitud carrotanques.

Figura 45. Distribución de tópicos - Enel



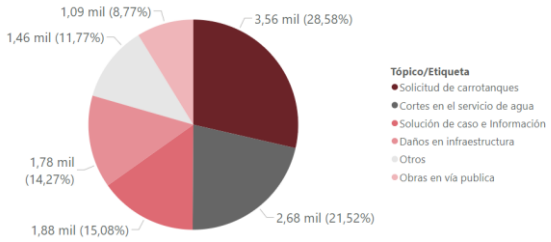
Fuente: Autoría Propia.

Figura 46. Distribución de tópicos - Vanti



Fuente: Autoría Propia.

Figura 47. Distribución de tópicos - Acueducto



Fuente: Autoría Propia.

A nivel general las 3 empresas presentan una alta asignación de tuits relacionados con Solución de caso e información de solicitudes, lo que corresponde a lo analizado previamente en las nubes de palabras. Se resalta también que, otro tópico con una participación bastante marcado corresponde a cortes en el servicio, por lo cual se puede pensar que los usuarios de las 3 organizaciones utilizan el canal precisamente para reportar sus fallas y presentar su mal estar por estas situaciones.

En el Anexo 3. Nubes de palabras en cada tópico, se detalla una nube de palabras con los términos de mayor frecuencia de aparición por cada uno de los cinco tópicos de cada empresa, se infiere que, los términos utilizados en los tuits corresponden al título y presentan una relación directa con el mismo.

5.7.4 Tópicos en el tiempo

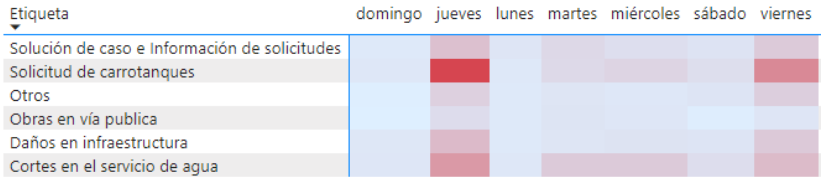
Se considera también de una alta relevancia presentar la evolución y comportamiento de los tópicos en el tiempo, con el fin de identificar si hay algún tipo de tendencia o relación entre posibles aumentos o disminuciones de los tuits asociados a estos. Para hacer esto, se construye un mapa de calor por día de la semana y una gráfica que permite visualizar el comportamiento mensual por cada tópico de cada empresa. Para realizar estas visualizaciones, se requiere contar con la base de datos o *dataframe* final (df_cruzado) en donde se encuentra: el tuit, el tópico asociado y su fecha de creación y posteriormente cargarlo en Ms Power BI (Microsoft Power BI, 2022) para realizar las visualizaciones correspondientes.

5.7.4.1 Acueducto

A nivel general se puede apreciar que para esta compañía los días con mayor cantidad de tuits asociados a solicitud de carrotanques (tópico dominante) son los jueves y viernes, seguido por los cortes en estos mismos dos días (lo cual respalda lo anteriormente expuesto en la evaluación de los tópicos). Por

ello, es recomendable que al interior de la empresa se revise si estos días se está realizando algún tipo de programación de cortes de agua que no están siendo informados a los clientes, y que, por tanto, generan un incremento relevante en los tuits no solo de reporte de fallas sino de requerimientos de carros para suplir dichas interrupciones en el servicio.

Figura 48. Días de la semana y tópicos – Acueducto

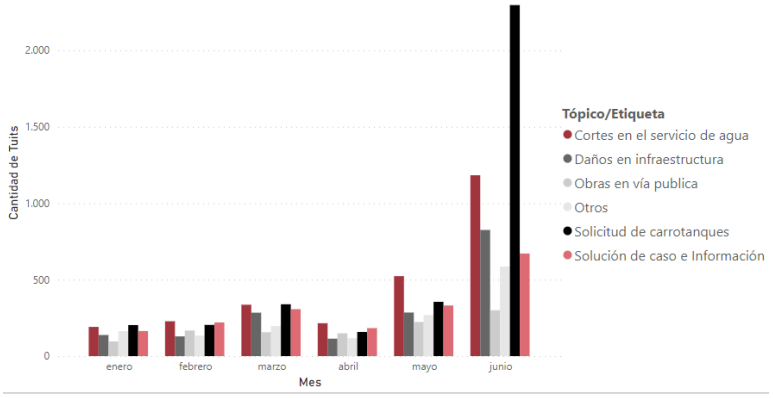


Fuente: Autoría Propia.

Por otro lado, al analizar el comportamiento mensual de las 5 temáticas propuesta, se evidencia un pico de más de 2000 tuits durante junio. Como ya se mencionó, esto se debe a que los días 23 y 24 de dicho mes se presentó una falla masiva en el servicio en diferentes barrios ubicados en sector noroccidental de Bogotá, los cuales conjuntamente reportaron la situación y al ser una falla de tal magnitud se presentó escases en los carrotanques, razón por la cual los temas: solicitud de carrotanques, cortes en el servicio y daños en infraestructura tienen un incremento de más del 50%. Sin embargo, si se revisan los meses anteriores, se evidencia que en su gran mayoría que son estos mismos tópicos que presentan una mayor volumetría de tuits.

A nivel general, para esta compañía se observa que el canal tiene una tendencia a ser más utilizado para reportar situaciones asociadas con los temas anteriormente expuestos, únicamente en de marzo a abril se presenta un decrecimiento, de resto mes a mes los reportes por Twitter aumentan.

Figura 49. Meses y tópicos – Acueducto

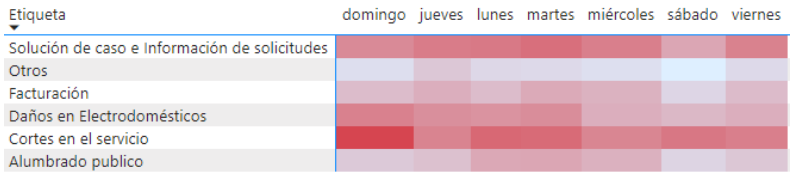


Fuente: Autoría Propia.

5.7.4.2 Enel

Para la empresa Enel se exhibe un comportamiento uniforme en la semana en los tópicos de solución de caso e información de solicitudes y cortes en el servicio, este último presenta una alta volumetría los domingos (al igual que para el acueducto) sin embargo, durante los demás días también se puede decir que tiene una gran cantidad de reportes. Por lo cual, debe evaluarse y orientar al canal a que acoja estrategias que permitan volverse más resolutivo y de orientación para el cliente, además de revisar que sucede con las interrupciones en el servicio, en especial las asociadas a Cundinamarca.

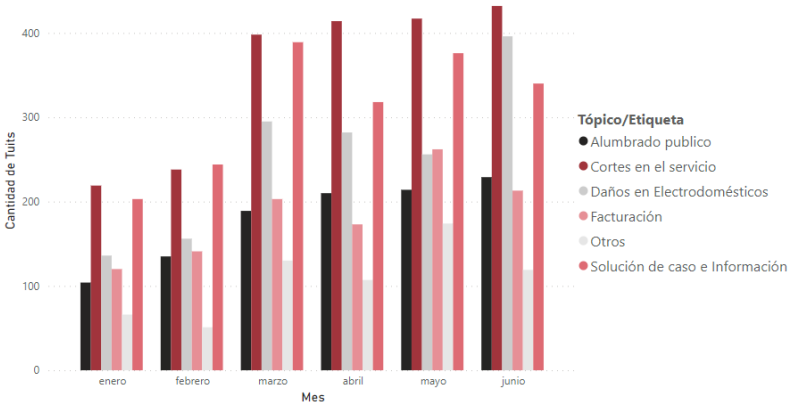
Figura 50. Días de la semana y tópicos – Enel



Fuente: Autoría Propia.

En cuanto al comportamiento mensual, se puede ver en la gráfica que hay un incremento importante en el uso del canal durante el mes de marzo y a partir del cual se mantiene constante. A diferencia de las demás empresas, los tópicos en general tienen un comportamiento estable. En donde los cortes en el servicio repuntan mes a mes, es decir, el canal es mayormente utilizado para solicitar información sobre qué pasa con las interrupciones y para reportar estas mismas.

Figura 51. Meses y tópicos – Enel



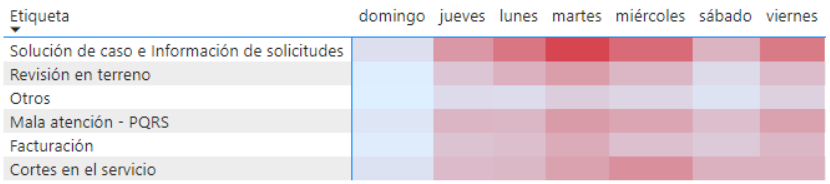
Fuente: Autoría Propia.

Seguido de esto, aparecen los grupos de tuits vinculados a la solución de caso e información de solicitudes y daños en electrodomésticos. Estos tres tópicos están estrechamente relacionados, dado que al existir cortes las personas reportan por este medio la falla, solicitan información de avances de solución de estas y derivado de esto, cuando se dañan algunos electrodomésticos, el cliente decide informarlo directamente por el canal.

5.7.4.3 Vanti

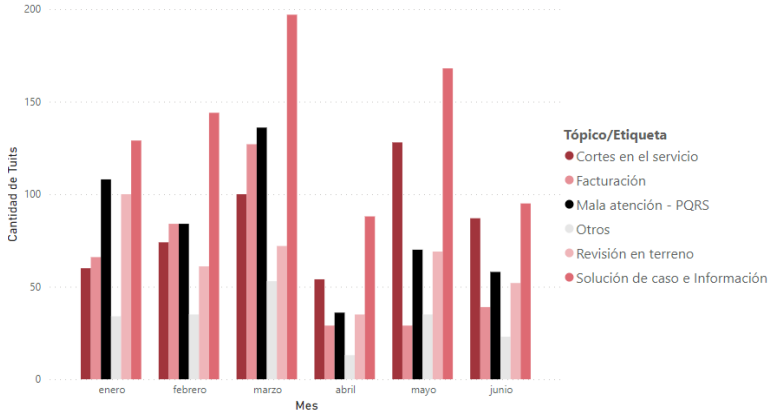
A partir de los tuits de Vanti, se obtiene que lo martes es el día con mayor tráfico en el canal en cuanto a solución de caso e información de solicitudes, se puede decir que este tema es el que presenta un comportamiento más marcado, dado que los 4 restantes son estables durante todos los días. Los domingos a diferencia de la compañía Enel y al igual que el acueducto, son días donde no hay una alta usabilidad del canal y por lo tanto no se pueden catalogar tuits en ninguno de los 5 ejes definidos. Se debe entonces revisar algún tipo de estrategia para reforzar la atención en el canal durante el martes. En este caso no se encuentra una relación directa entre los días de aparición entre los distintos tópicos.

Figura 52. Días de la semana y tópicos – Vanti



Fuente: Autoría Propia.

Figura 53. Meses y tópicos – Vanti



Fuente: Autoría Propia.

Durante los seis meses que en los que se realiza el análisis, se evidencia que el tópico dominante es el asociado a solución de casos e información de solicitudes, se puede ver que hay un leve pico durante el mes de marzo y que después de este mes (en especial en abril) se va disminuyendo paulatinamente. A diferencia de las demás compañías, el segundo tópico de mayor aparición no es constante, durante algunos meses se encuentra que es mala atención – PQRs y otros cortes en el servicio.

5.7.5 Análisis de sentimientos por tópico en el tiempo

Para identificar las áreas que presentan una mayor dolencia para el cliente, se propone realizar un análisis temporal por tópico de la categorización – Tuit positivo, negativo, neutro – obtenida de la aplicación del diccionario Bing (se opta por utilizar este lexicón, a razón de que es el que mayor porcentaje de asignación a nivel general presenta como se explica en la sección 5.5.4 Análisis de sentimientos).

En general los cortes en el servicio y la solicitud de información y de casos son las áreas con mayores oportunidades de mejora, como se ha mencionado, se relaciona con que no se avisan de los cortes y con la poca información que se obtiene sobre los casos mediante esta red social. A continuación, se detalla lo obtenido por cada empresa enfatizando en el análisis de las comunicaciones negativas, para facilitar este análisis se generan los mapas de calor que se encuentran en los anexos 5, 6 y 7.

5.7.5.1 Acueducto

El mapa de calor que representa el comportamiento de los tópicos por categoría para esta compañía se encuentra en el Anexo 4. Mapa de calor por tópico mensual – Acueducto. De allí se puede inferir lo siguiente:

Esta empresa presenta un comportamiento atípico en el mes de junio, justificado en lo que se explicó previamente (más exactamente en la sección 5.6.1 Líneas de tiempo; Figura 24. Evolución temporal de la cantidad de tuits recibidos por el Acueducto). Por lo tanto, si se deja de lado este mes se evidencia que para esta compañía (al igual que para Enel) el mayor dolor de los clientes corresponde a cortes en el servicio de agua, seguido de solución de caso e información de solicitudes. El primer grupo iniciando con menos de 150 tuits negativos y finalizando con más de 300 a pesar de presentar un quiebre en el mes de abril, por su parte, el tópico de solicitud de caso e información inicia con 108 tuits y termina en poco menos de 200. Ambos presentan un incremento considerable.

Al incluir en el análisis el tópico de solicitud de carro tanques, este pasa a ocupar el primer lugar con casi 1162 comunicaciones en junio (referente a tuits negativos), sin embargo, en meses pasados se

puede decir que este grupo estuvo controlado dado que siempre se mantuvo con menos de 200 tuits en la categoría de negativos.

A nivel general se evidencia un comportamiento en donde hay una tendencia al alza en los tuits negativos. Todos los tópicos a excepción del que se asocia a la solicitud de carrotanques presentan un punto de quiebre en marzo, en donde se incrementan tanto los positivos como los negativos y en abril se da una tendencia a la baja.

A diferencia de Enel, todos los 5 grupos tienen un comportamiento similar. En enero se presenta un promedio de 82 tuits negativos por tópico, para el último mes del análisis, este promedio se aumenta a 165, se puede pensar que la buena percepción ha ido disminuyendo. La tasa de tuits negativos en enero era en promedio del 51% (en donde Solución de caso e Información de solicitudes presenta una participación del 11.1%)

5.7.5.2 Enel

Al analizar la tabla del Anexo 5. Mapa de calor por tópico mensual – Enel, se identifica fácilmente que el tópico con mayor cantidad de tuits negativos corresponde a cortes en el servicio y que en el periodo de evaluación tiene una tendencia constante durante los últimos meses, presentándose una subida puntual en el mes de marzo y estabilizándose cerca de 300 tuits negativos por mes. Por su parte, el único grupo de comunicaciones que tiene una leve disposición a la baja. corresponde al grupo de facturación.

A nivel general, hay un crecimiento para todos los tuits negativos, por lo cual puede pensarse que el grado de satisfacción ha tendido a la baja (además de haber un incremento en la contactabilidad del canal), lo que se secunda al revisar que en enero en promedio existían 85 tuits negativos por cada tópico, pero al finalizar el análisis esta relación casi que se duplica (llegando a 166).

La tasa promedio de tuits negativos para todos los tópicos es del 60% (de la cual cerca del 18% corresponde al tópico de cortes (el más crítico)). Se recalca que el grupo: otros únicamente presentan un crecimiento importante al evaluar los tuits categorizados como “no etiquetados” dado que pueden existir otras temáticas que no se contemplan al realizar el algoritmo LDA.

5.7.5.3 Vanti

Para los tuits negativos de esta compañía, se presenta un comportamiento similar al del acueducto, en donde los 3 primeros meses tienen una tendencia al alza y en abril se da un punto de inflexión, un aumento nuevamente en mayo y por último en junio, algunos tópicos tienden a estabilizarse y a bajar la cantidad de comunicaciones negativas de los clientes.

Dentro de esta categoría, resalta el comportamiento de los mensajes asociados a: solución de caso e información de solicitudes, las cuales inician con 90 tuits en el mes de enero y finalizan junio con 64. Por otro lado, se resaltan dos categorías:

- **Facturación:** A pesar de presentar un incremento en los 3 primeros meses, a partir de abril, se da una tendencia a la baja que se mantiene constante, siendo el tópico que finaliza con menor cantidad de tuits en la categoría de tuits negativos (30 en total).
- **Revisión en terreno:** A diferencia de la anterior; presenta una tendencia al a la baja notable en enero, febrero, marzo y abril, sin embargo, a partir de este último periodo, se da un incremento considerable, finalizando con 40 tuits después de haber tenido un mínimo de 18 previamente.

El promedio de tuits negativos mensual corresponde al 62%, de este porcentaje el 20% se asocia a: solución de caso e información, en cuanto a la categoría de tuits positivos, esta corresponde al 20% de las comunicaciones promedio por mes, y de este, el 4.4% está relacionado con el tópico: solución de caso e información de solicitudes.

5.8 Encuesta CIER – Aplicación Enel

La encuesta CIER dirigida por la Comisión de Integración Energética Regional – CIER - (Innovare Pesquisa CIER, 2021), es una investigación que se realiza a nivel internacional enfocada en determinar, comparar y encontrar oportunidades de mejora para las empresas distribuidoras del sector energético residencial. Se lleva a cabo anualmente y los resultados se emiten al año siguiente de haber sido realizada.

Como lo indican en su sumario ejecutivo del año 2021, es una investigación que pretende apoyar a las empresas para mejorar sus procesos “La encuesta CIER se realiza anualmente desde 2003, y ofrece a las distribuidoras instrumentos e incentivos destinados a mejorar su desempeño. Entre los objetivos del trabajo, se destacan:

- Medición del nivel de satisfacción de los clientes con respecto a la calidad del producto y de los servicios prestados por la distribuidora.
- Generación de índices que permitan la comparación de los resultados entre todas las distribuidoras.
- Generación de matrices de apoyo a la definición de acciones de mejora” **(Innovare Pesquisa CIER, 2021)**.

Se enfoca principalmente en ciertas áreas que, de cara al cliente, tienen un impacto bastante relevante y son las siguientes: servicio de energía, factura, atención al cliente, comunicaciones e imagen de la empresa. Cada una de las cuales se subdivide a su vez en máximo 5 o 6 atributos, sobre los cuales los clientes responden a la pregunta: "Que tan satisfecho se encuentra con este atributo" y califican de 1 a 10. Una vez se tiene la puntuación de cada atributo, se clasifica según la siguiente escala y se obtiene el porcentaje respectivo que corresponde a cada intervalo:

Figura 54. Intervalos de clasificación encuesta CIER

Fuente: (Innovare Pesquisa CIER, 2021).

De esta manera se obtiene el IDAT y el IDAR (índice de desempeño del área), el cual como se menciona en (Innovare Pesquisa CIER, 2021) corresponde al porcentaje de clientes cuya evaluación haya sido igual o mayor que 7 para todos los atributos de un área de evaluación (por ejemplo: atención al cliente, factura, entre otros), sin tomar en consideración a aquellos que no supieron o se negaron a contestar.

5.8.1 Homologación de etiquetas y escala

Para efectos de la presente investigación, se opta por tener en cuenta la clasificación a nivel IDAR (no por atributos), dado que se ha asociado cada tuit de cada usuario a un tema en específico, el cual para alcance de la presente investigación no llega a dividirse en más atributos, por ejemplo, el tópico: "Factura" no cuenta con subdivisiones para saber los subtemas que se encuentran dentro de este. Por lo anterior, se propone es una homologación entre los 5 tópicos encontrados y las áreas de evaluación propuestas por (Innovare Pesquisa CIER, 2021). A continuación, se presenta esta validación a partir de los tópicos del corpus de Enel:

Tabla 17. Atributos encuesta CIER vs Tópicos

Atributo encuesta CIER	Tópico / Etiqueta
Suministro de energía	Cortes en el servicio
Atención al cliente	Solución de caso e Información de solicitudes
Factura de energía	Facturación
Alumbrado publico	Alumbrado publico
Atención al cliente	Daños en Electrodomésticos

Fuente: Autoría Propia.

Los siguientes atributos no coinciden con ninguno de los tópicos determinados después de aplicar LDA: Información y comunicación, responsabilidad socioambiental e Imagen. Así mismo, los tuits que fueron clasificados con la etiqueta otros, se dejan como otros dado que no se pueden clasificar y aun menos homologar con la encuesta CIER.

En cuanto a la homologación de la escala numérica referida en la Figura 54, primero; se toma como referencia el valor cuantitativo de los tuits calculado mediante el aplicación del Diccionario Affin (el cual se explica en la sección: 5.4.2 Indicadores Diccionario AFINN), después se encuentra que el 98% de los datos obtienen una calificación que varía entre -0.5 y 0.5, de esta manera se toma la puntuación máxima y mínima de los tuits que existe en este intervalo para calcular el rango. Se encuentra lo siguiente:

Tabla 18. Puntuación máxima, mínima y rango - Homologación CIER

Puntuación Máxima	0,489
Puntuación Mínima	-0,5
Rango	0,989

Fuente: Autoría Propia.

Finalmente, para definir los distintos intervalos se va sumando en un 20% el valor del rango, por ejemplo: para el intervalo tres que corresponde a los tuits que deben ser clasificados como “regulares”, se suma un 60% del valor del rango a la puntuación mínima encontrada en el intervalo entre – 0.5 y 0.5. En la se encuentra la división numérica para categorizar cada tuit de acuerdo con la escala CIER:

Tabla 19. Categorización tuits – Homologación CIER

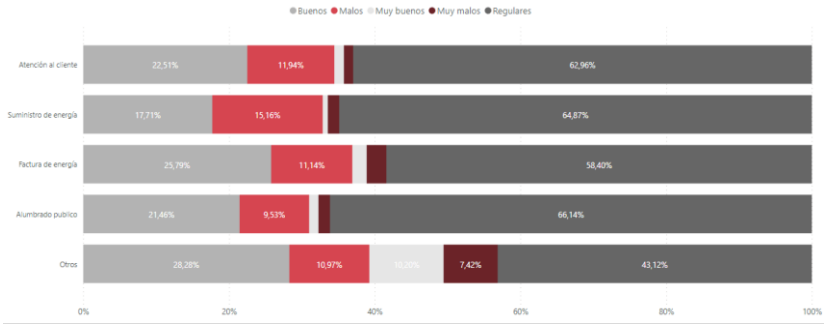
Categoría CIER	Límite inferior	Límite superior	Cantidad de tuits categorizados
Muy malos		-0,3022	171
Malos	-0,3022	-0,1044	1025
Regulares	-0,1044	0,0934	5158
Buenos	0,0934	0,2912	1842
Muy buenos	0,3		162
	Total		8358

Fuente: Autoría Propia.

5.8.2 Resultados

A nivel general el intervalo que contiene una mayor cantidad de tuit equivale a la categoría de regulares, con cerca de un 62%, seguido de los buenos y malos. Se observa que los valores extremos (muy malos y muy buenos) tienen una asignación muy pequeña, únicamente resaltando en la categoría otros.

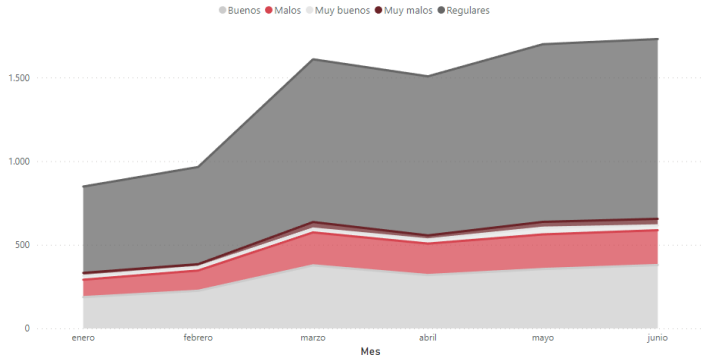
Figura 55. Participación categorías homologadas encuesta CIER



Fuente: Autoría Propia.

Como se explica más adelante, la gran mayoría de tuits se puntúan con un valor cercano a cero, razón por la cual esta categoría es la mayor. Se resalta que, según esta escala de homologación, los tuits buenos superan en casi 900 a los negativos, esto se refleja en los porcentajes por categoría que se aprecian en la Figura 55. Participación categorías homologadas encuesta CIER. Este comportamiento se presenta durante todos los meses, como se puede ver en la siguiente gráfica, en donde mes a mes los tuits categorizados como positivos superan a los negativos:

Figura 56. Categorización CIER por mes



Fuente: Autoría Propia.

Con lo anterior y según el procedimiento que se planteó, el IDAR basado en el *topic modeling* y el análisis de sentimientos, se presenta en la Tabla 20. El tema más crítico corresponde al suministro de energía, se podría pensar que se debe principalmente a que a pesar de tener la mayor cantidad de tuits asociados mes a mes, la categoría de “regulares” es la mayor de los tópicos homologados, además de ser la segunda

que menos comunicaciones "positivas" tiene. Por último, se debe mencionar que las comunicaciones categorizadas como negativas en factura y atención al cliente, son muy similares.

Tabla 20. IDAR por categorías de topic modeling homologadas con encuesta CIER

Categoría CIER	Atención al cliente	Factura de energía	Suministro de energía	Alumbrado Publico
Positivos	23,83%	27,77%	19,88%	22,76%
Negativos	13,21%	13,84%	14,81%	11,10%
IDAR	10,62%	13,93%	5,07%	11,66%

Fuente: Autoría Propia.

5.8.3 Consideraciones

- **Privacidad de la información:** En la página de la organización CIER se encuentra disponible el informe de 2021 **(Innovare Pesquisa CIER, 2021)** sin embargo, los resultados que allí se exponen se encuentran a nivel general y no se cuenta con el detallado para ninguna empresa (incluyendo Enel), por lo cual, la comparación con los resultados obtenido en el presente análisis no se puede realizar. Se opta entonces por proponer el método de homologación y en caso de requerirse, la empresa internamente podrá contrastar los resultados.
- **Horizonte de tiempo:** Como ya se mencionó, el informe con el que se cuenta es del año 2021 y los tuits con los que se realiza la presente investigación son de 2022 por lo cual, aun si se contara con el resultado CIER para la compañía en cuestión, no se podría realizar la comparación dado que los resultados aún no han sido publicados. Por ello, se hace énfasis en que el objetivo de esta etapa es proponer un posible método para llegar a obtener un resultado que permita identificar oportunidades de mejora en un breve espacio de tiempo.
- **Granularidad encuesta CIER:** Para cubrir lo anteriormente expuesto, se puede volver a realizar una interacción con el algoritmo LDA pero para cada conjunto de tuits de cada tópico por aparte, esto para buscar la homologación a nivel IDAR, esto se propone como trabajos futuros dado el alcance y tiempo de la presente investigación.
- **Homologación:** A pesar que se trataron de utilizar otros métodos para homologar la escala (por ejemplo normalizando o tomando el valor mínimo como el más pequeños de los valores asignados a los tuits y el máximo como el mayor de las puntuaciones), muchos valores se aproximan al 0 por lo cual se recarga el intervalo de las comunicaciones que deben ser asignadas como normales, se recomienda incluir y enriquecer los diccionarios para que se reconozcan más palabras y se pueda aumentar o disminuir el valor asignado a cada tuit.

5.9 Métricas

Teniendo en cuenta todo lo anterior, se puede llegar a preguntar sobre que indicadores pueden presentarse con el fin de realizar un seguimiento a la percepción del servicio de las empresas, en esta sección se proponen algunas métricas que pueden llegar a aportar en esta cuestión y que evaluadas cada cierto periodo de tiempo pueden llevar a tomar acciones que permitan una mejor operación del canal y por tanto pueden aportar una mejora en la calidad del servicio (visto inclusive desde los distintos tópicos de cada compañía).

Las métricas propuestas se encuentran en la Tabla 21, las cuales en su mayoría están basadas en el diccionario BING (dado que fue el mejor diccionario que otorgó el mayor rendimiento a los datos analizados) , no se incluyen muchas métricas que se pueden calcular a partir del diccionario Affin; ya que de acuerdo a los resultados obtenidos, la medición podría tomar valores extremos para un alto porcentaje del total de datos analizados, ocasionando un análisis de percepción con alta variabilidad y margen de error.

Tabla 21. Métricas propuestas de evaluación

Nombre	Objetivo	Temporalidad Propuesta	Descripción	Detalle
Cantidad de tuits por hora por tópico	Identificar tópicos que puedan estar siendo tendencia dentro de cada compañía de manera temprana	Un corte cada hora	A través de esta métrica se propone identificar los tópicos de los que más se está hablando en cada compañía, lo que permite revisar rápidamente si se tiene alguna novedad con alguna temática y poder tomar acciones correctivas, por ejemplo, informar a los clientes sobre algún corte no esperado en el servicio. Este indicador se calcula basado en el máximo histórico de la cantidad de tuits de cada compañía por cada tópico a cada hora del día.	Anexo 7
Tuits negativos acumulados por hora	Identificar una alta volumetría y afluencia de clientes insatisfechos	Un corte cada hora	Este indicador se centra en reconocer una gran cantidad de tuits acumulados por hora, lo que permite identificar si alguna empresa en particular está presentando algún tipo de situación que genere un mal estar general. Se utiliza el diccionario BING para clasificar los tuits, además de la cantidad acumulada por hora de cada cuenta de la categoría: tuits negativos.	Anexo 8
Promedio esperado de tuits positivos en 6 meses	Identificar acciones de mejora que se hayan implementado que mejoren la satisfacción y percepción el cliente,	Semestral	Este indicador podría pensarse que se utiliza para medir el grado de existo de acciones que se hayan tomado en el mediano plazo para la mejora de la percepción del servicio de las empresas, se basa la utilización de la categoría Tuits positivos del diccionario Afinn. Se utilizan los datos del inicio	**Metas:** Acueducto: 57,95% (7113 tuits) Enel: 43,41%

Nombre	Objetivo	Temporalidad Propuesta	Descripción	Detalle
	enfocado en el aumento de comunicaciones positivas		hasta el fin del estudio para determinar el ratio de crecimiento o decrecimiento de esta categoría y sin tener un objetivo a cumplir en el semestre.	(3628 tuits) Vanti: 57% (1581 tuits)
Cantidad de tuits negativos en el tópico critico por empresa	Identificar acciones que permitan mejorar la percepción de cada tópico y garantizar una mejora basada en la voz del cliente	Diario	A partir de la identificación de los tópicos con mayor cantidad de tuits negativos (categorizados a través del diccionario Bing), se propone realizar un seguimiento diario a la cantidad de comunicaciones que se cataloguen dentro de este grupo. Al igual que en casos anteriores, se pueden identificar acciones que hayan generado un decrecimiento en este indicador y que puedan mantenerse en el tiempo para generar una mejora continua. Se calcula a partir del promedio mensual histórico de las comunicaciones recibidas de cada tópico crítico.	Anexo 9
Tuit *polarity* score (tps)	Determinar la cantidad de tuits positivos y negativos	Un corte cada hora	Esta métrica es la base para realizar el análisis de sentimientos, lo cual permite identificar y categorizar los tuits en positivos, negativos o neutros a través de la aplicación del diccionario Bing. A su vez, es el fundamento para poder hacer análisis más profundos en cuanto a la criticidad de los temas, especialmente la medición del desempeño de cada tópico de cada compañía en el tiempo	Ver sección 5.4.1

Nombre	Objetivo	Temporalidad Propuesta	Descripción	Detalle
Pn tuit polartiy score (pn tuit)	Determinar la relación entre la totalidad de tuits positivos y negativos	Un corte cada hora	El PN tuit es un indicador que siempre debe tener como objetivo ser mayor o igual a uno, esto debido a que es la razón entre la cantidad de tuits positivos y negativos de cada empresa en cierto periodo de tiempo, encontrando si la totalidad de comunicaciones positivas es mayor que la negativa. Posibilita de una manera rápida realizar el seguimiento a la percepción general del usuario.	PN Tuit >= 1 Ver sección 5.4.1
User *polarity* score (up)	Identificar la categoría (positiva, negativa, neutra, n/a) de los usuarios a partir del etiquetado de sus tuits	Un corte cada hora	Mediante este indicador se realiza la categorización a la totalidad de los usuarios que se comunican con cada compañía, centrándose en identificar cuáles de estos tienen una mayor cantidad de tuits positivos o negativos para asignar una etiqueta a cada uno. Este indicador se basa en el TPS para cumplir con su objetivo principal.	Ver sección 5.4.1
Pn user *polarity* scorte (pn user)	Determinar la relación entre usuarios categorizados como positivos y negativos	Un corte cada hora	El PN user permite realizar un seguimiento rápido a la percepción centrada en los usuarios de Twitter (Twitter, S.F) , al igual que el PN Tuit debe tener como objetivo lograr valores mayores o iguales a uno para cada empresa, ya que presenta la relación entre los usuarios que tienen una mayor cantidad de tuits positivos versus los que tienen una mayor cantidad de tuits negativos.	PN User >= 1 Ver sección 5.4.1

Fuente: Autoría Propia.

5.10 Dashboard

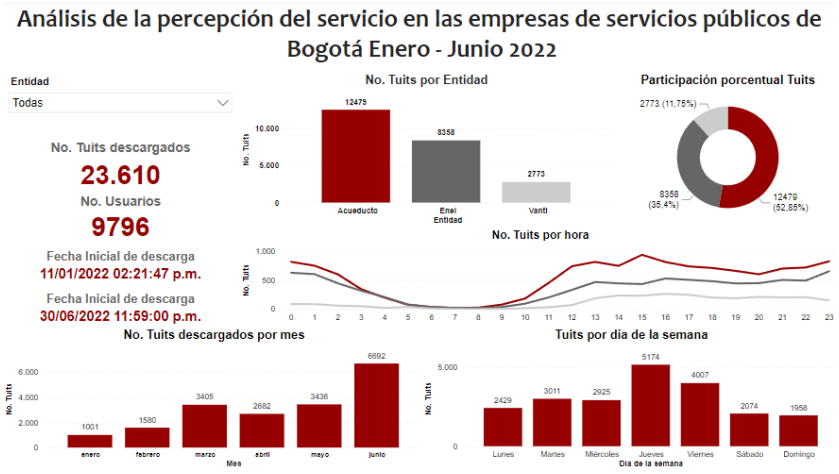
Uno de los objetivos de esta investigación consiste en proponer algunos tableros de control que permitan realizar un seguimiento y un análisis del comportamiento en general de los tuits de los usuarios, no solo en términos de volumetría y contactabilidad del canal, sino también vistas que permitan identificar rápidamente los tópicos y temas en general de los que la gente está hablando en Twitter.

Las visualizaciones que a continuación se describen, se desarrollan en Ms Power BI (Microsoft Power BI, 2022) precisamente porque se enfocan en que de una manera dinámica y rápida un director o jefe pueda identificar rápidamente el estado del canal. Se aclara que, si bien los tableros no están diseñados en tiempo real, pueden servir como base para posibles futuras implementaciones.

5.10.1 Resumen - General

La primera vista llamada Resumen – General, permite identificar fácilmente cuál de las 3 empresas presenta un mayor uso del canal en términos de cantidad de tuits descargados y de número de usuarios únicos que se han comunicado en el periodo de enero a junio de 2022 (como se mencionó previamente, esto puede variar si es que alguna compañía opta por hacer una implementación en tiempo real). Así mismo, a través de las gráficas de barras y de líneas se identifican fácilmente días, meses y horas pico, con el fin de saber en qué franjas el canal puede estar más congestionado y plantear tácticas y estrategias que permitan responden a la situación. Es importante aclarar que a través del filtro se puede ver el detalle de cada entidad.

Figura 57. Vista: Resumen – General

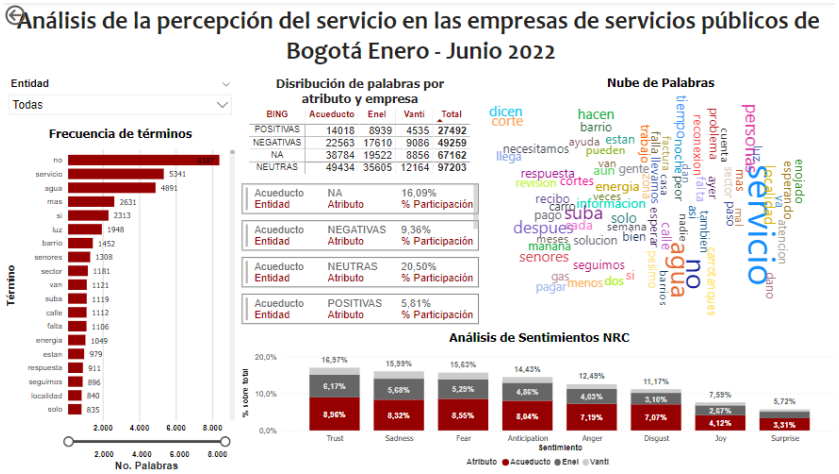


Fuente: Autoría Propia.

5.10.2 Resumen – Corpus

Esta vista corresponde a la visualización que se plantea para explorar rápidamente los corpus de cada compañía, como ya se ha mencionado previamente, estos corpus se forman a partir de los tuits y de la aplicación de los distintos niveles de limpieza, se puede decir que son las palabras que realmente generan valor. A través de la gráfica de frecuencia y las nubes de palabra se puede tener una idea sobre qué términos son los más empleados en los tuits. Así mismo se propone un resumen genérico de la aplicación de los 3 diccionarios mencionados previamente, a través de Bing se identifican los términos positivos negativos y neutros (se propone una visualización para ver valores porcentuales y cantidad de tuits) y mediante la gráfica de barra se identifica fácilmente los sentimientos que están asociados a los tuits de cierta entidad y en general a cada compañía.

Figura 58. Vista: Resumen - Corpus



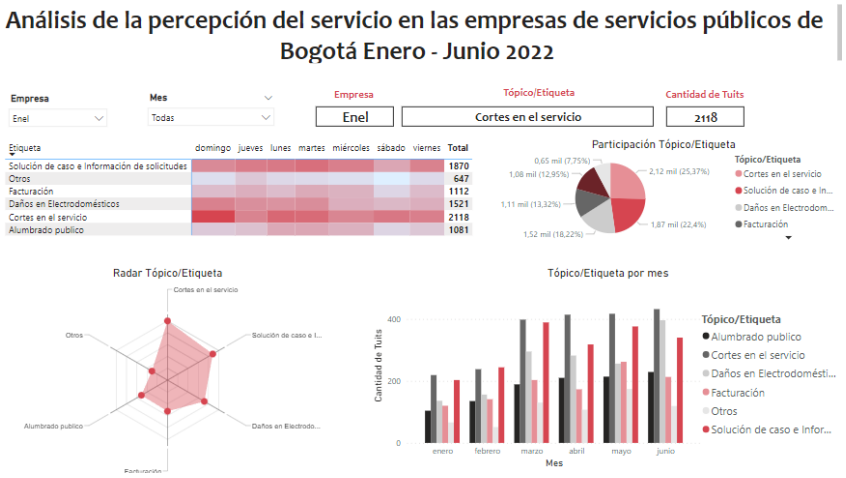
Fuente: Autoría Propia.

5.10.3 Topic modeling

Para visualizar fácilmente los temas de los cuales se están hablando en los tuits, se propone la siguiente visualización, en donde lo primero que se observa es la empresa filtrada y de la cual se requiera la información, seguido del tópico que más tuits tiene asociado además de la cantidad exacta de estos. Adicionalmente, se tiene un mapa de calor con el cual se reconoce fácilmente el día en el que más se habla sobre cierta temática. Seguido de esto, un gráfico circular para conocer el comportamiento en términos porcentuales de cada grupo de tuits y así analizar las proporciones entre temáticas más

rápidamente, lo que se puede reconocer también con el radar. Finalmente, es interesante analizar el comportamiento de los tópicos en el tiempo, a través del esquema de barras.

Figura 59. Vista: Topic modeling

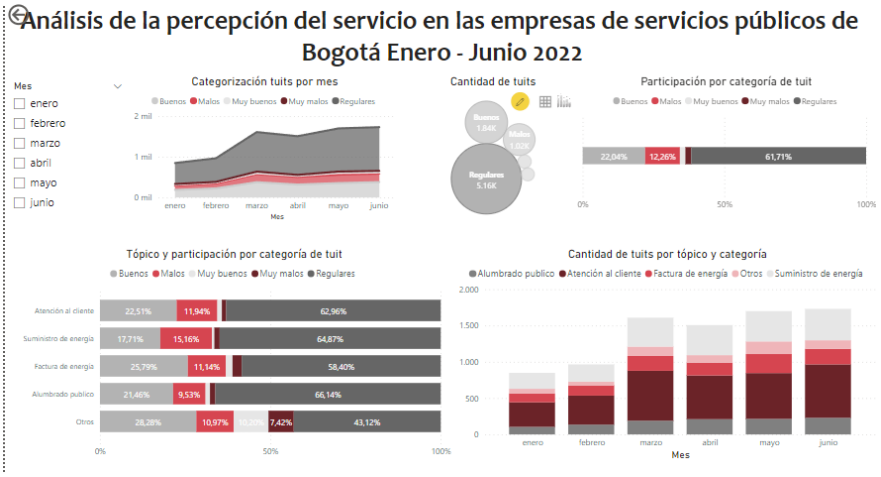


Fuente: Autoría Propia.

5.10.4 Encuesta CIER – Enel

Esta vista, como su nombre lo indica, únicamente se centra en los tuits de Enel, esto a razón de que la encuesta CIER solo aplica para empresas del negocio de energía (como ya se explicó previamente).

Figura 60. Vista: Topic modeling



Fuente: Autoría Propia.

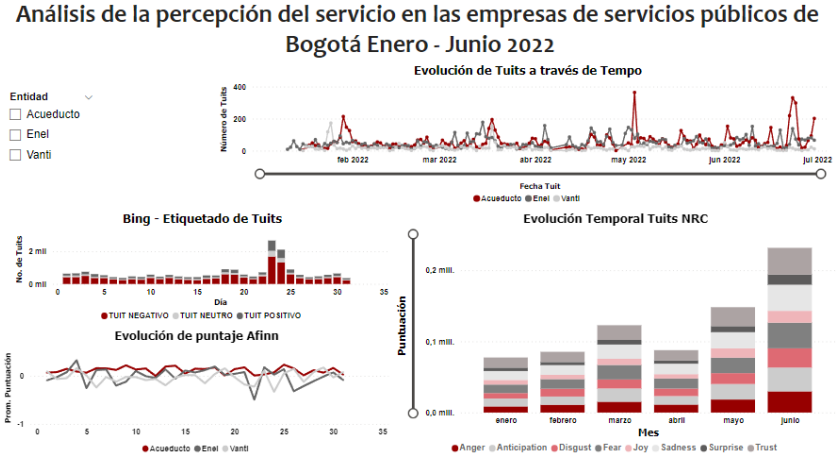
Es aquí donde se presentan los tópicos homologados con los diferentes pilares (alumbrado público, atención al cliente, factura de energía, otros, suministro de energía) o atributos de la encuesta y su calificación u categorización (bueno, malo, muy bueno, muy malo y regular) de cada comunicación.

Al igual que en el tablero anterior, la evolución en el tiempo es bastante relevante por lo cual, se encuentran diferentes tipos de graficas (barras y áreas) que permite evidenciar la evolución de las categorías y atributos en el tiempo, así como su participación porcentual. Por último, se presenta la cantidad de tuits en cada categoría y a nivel general, esto último mediante un gráfico de burbujas.

5.10.5 Análisis de sentimientos – Series Temporales

La penúltima visualización que se propone está enfocada en reconocer fácilmente el detalle de la aplicación de los diccionarios en el tiempo. Para realizar esto, se considera de alto valor en primera medida, mostrar el comportamiento general de la cantidad de tuits de cada empresa, para así reconocer por ejemplo que en mayo se presentó un pico para el acueducto y posibles estacionalidades bimestrales para esta misma empresa. Por otro lado, la gráfica de barras del diccionario Bing, muestra de una manera visual cuantos tuits – positivos, negativos y neutros- se presentaron en el tiempo. Por su parte, el diccionario Afinn se representa mediante graficas de líneas para conocer el intervalo dentro del cual los puntajes de cada tuit se encuentran mes a mes, y reconocer que a nivel general ninguna compañía es estable en sus valores obtenidos. Por último, el diccionario NRC muestra un detalle mayor que el presentado previamente, a través de un gráfico apilado que ilustra los sentimientos con mayor intensidad mes a mes.

Figura 61. Vista: Análisis de tiempo – series temporales

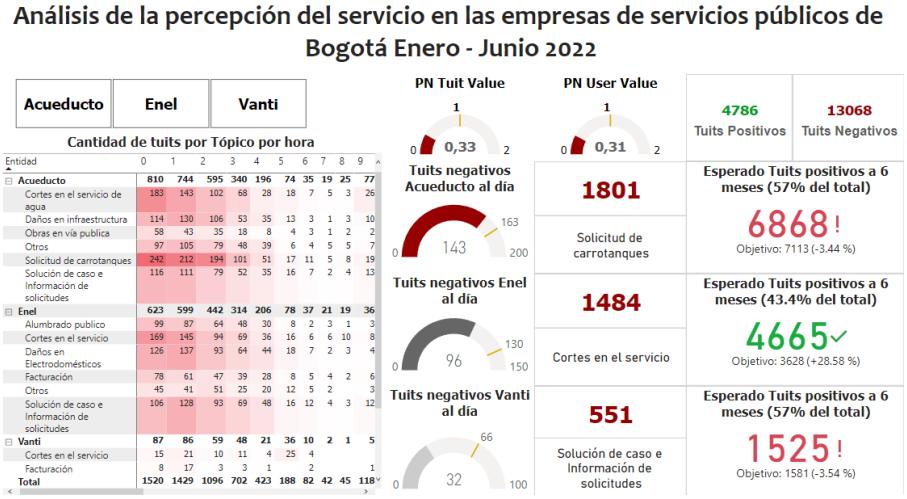


Fuente: Autoría Propia.

5.10.6 Seguimiento a métricas

Finalmente, la última visualización que se plantea permite evidenciar el comportamiento de las diferentes métricas que se desarrollaron en la sección 5.9 Métricas. Se plantea un mapa de calor para visibilizar la cantidad de tuits por tópico hora a hora y así identificar los principales temas de los que se están hablando en la red social. Se encuentra también la cantidad de tuits negativos al día para realizar un seguimiento a las temáticas críticas, las cuales se pueden ubicar junto con la cantidad de mensajes asociadas a cada una de ellas. Por otro lado, se muestra la categorización de los tuits positivos/negativos y las métricas que se desprenden de esta: "PN *Tuit Value*" y "PN *User value*", los cuales siempre deben tomar un valor mayor o igual a uno. Para precisar si las acciones que se han tomado en pro de la mejora de la percepción del servicio han tenido algún impacto, se expone el esperado de tuits positivos en los 6 meses, resaltando en rojo o en verde si se está o no en senda de cumplimiento.

Figura 62. Vista: Métricas



Entidad	0	1	2	3	4	5	6	7	8	9
Acueducto	**810**	**744**	**595**	**340**	**196**	**74**	**35**	**19**	**25**	**77**
Cortes en el servicio de agua	183	143	102	68	28	18	7	5	3	26
Daños en infraestructura	114	130	106	53	35	13	3	1	3	10
Obras en vía publica	58	43	35	18	8	4	3	1	2	2
Otros	97	105	79	48	39	6	4	5	5	7
Solicitud de carrotanques	242	212	194	101	51	17	11	5	8	19
Solución de caso e Información de solicitudes	116	111	79	52	35	16	7	2	4	13
Enel	**623**	**599**	**442**	**314**	**206**	**78**	**37**	**21**	**19**	**36**
Alumbrado publico	99	87	64	48	30	8	2	3	1	3
Cortes en el servicio	169	145	94	69	36	16	6	6	10	8
Daños en Electrodomésticos	126	137	93	64	44	18	7	2	3	4
Facturación	78	61	47	39	28	8	5	4	2	6
Otros	45	41	51	25	20	12	5	2		3
Solución de caso e Información de solicitudes	106	128	93	69	48	16	12	4	3	12
Vanti	**87**	**86**	**59**	**48**	**21**	**36**	**10**	**2**	**1**	**5**
Cortes en el servicio	15	21	10	11	4	25	4			
Facturación	8	17	3	3	1		2			1
Total	**1520**	**1429**	**1096**	**702**	**423**	**188**	**82**	**42**	**45**	**118**

Fuente: Autoría Propia.

6. Conclusiones y recomendaciones

6.1 Conclusiones

La minería de texto ha tomado gran relevancia para las organizaciones, puesto que permite explotar los datos textuales y convertirlos en información de interés para la toma de decisiones. Las redes sociales, no están aisladas de este fenómeno, por lo cual las empresas las están aprovechando para: promocionar sus productos o servicios, posicionar su marca y como un canal de atención con sus usuarios, sin la necesidad de realizar altas inversiones en desarrollo y despliegue tecnológico para lograr tal fin.

Por otro lado, realizar un análisis cuantitativo y cualitativo a las empresas de servicios públicos de Bogotá usando sus cuentas oficiales de Twitter, permitió la personalización del modelo KDD y la evaluación del rendimiento de cada conjunto de datos para realizar: el análisis de sentimientos, la evaluación cualitativa a partir de las palabras que se usan con más frecuencia y el análisis estadístico que muestra la composición y comportamiento de las variables a nivel general y su evolución a través del periodo de tiempo.

Se destaca que una correcta aplicación de la fase de procesamiento permitió tener resultados de calidad y reflejar la realidad objeto de análisis. Cabe resaltar que los resultados están influenciados por algunos aspectos como: la limpieza de datos, transformación de los datos, estructuración de los datos, modelos elegidos para cuantificar y cualificar los textos, entre otros.

Uno de los hallazgos realizados en esta etapa es la limitante de Twitter respecto a la cantidad de caracteres que puede tener un mensaje (tuits de máximo 280 caracteres), por lo que cada uno de los hilos de usuario se consolidaron en un solo texto estructurado sobre el tuit inicial con la finalidad de tener el contexto general de la información que se quiere transmitir hacia cada una de las empresas.

La extracción de datos a lo largo de los 6 meses permitió obtener una muestra aceptable y suficiente de datos para hacer un análisis de sentimientos robusto, a pesar de reducir la cantidad de tuits al aplicar una limpieza divida en tres fases. Precisamente esta arquitectura de procesamiento permitió que los resultados consideraran netamente los tuits de los usuarios de las empresas seleccionadas y así estos no presentaran alteraciones u anomalías, al eliminar en la primera etapa, todos los mensajes relacionados a cuentas: corporativas, gubernamentales y de entidades sin ánimo de lucro. La segunda fase de limpieza consideró todas aquellas palabras que no aportaban valor dentro de cada uno de los mensajes y esto reduce el tamaño de los corpus a analizar. Finalmente, con el tercer nivel se retiraron las palabras que no estaban consideradas dentro del procedimiento de limpieza de R (R Core Team, 2021) a causa del idioma y esto

permitió obtener corpus con alta calidad para su posterior análisis. Dentro de los hallazgos relacionados con la limpieza de los datos, se resaltan los problemas de ortografía, las letras repetidas al final de cada vocablo, abreviaciones y palabras unidas en cada mensaje enviado por cada usuario y que ninguna de las fases dos y tres puede identificar plenamente para eliminarlas.

De acuerdo con lo evidenciado en la literatura revisada; el análisis de sentimientos se puede realizar con la aplicación de alguno de los diccionarios con los que cuenta R (R Core Team, 2021) para esta finalidad, en esta investigación el considerar los 3 (Bing, Afinn y NRC) para cada corpus permitió integrar los diferentes resultados que arroja cada y obtener resultados más detallados sobre los sentimientos asociados de cada usuario con cada empresa.

Por otro lado, se debe considerar que las investigaciones de minería de texto aplicadas en lenguaje español "latino colombiano" son escasas, dado que no se consideran modismos ni regionalismos y los software que permiten realizar el análisis de datos textuales, no consideran esto por su desarrollo en países con lengua inglesa; lo que conlleva a que dos de los tres diccionarios (BING y Afinn) fueran enriquecidos con diferentes variantes semánticas y sinónimos para que el corpus tuviera mayor cobertura al momento de cuantificar y cualificar cada uno de los términos. Lo que se ratifica con la prueba aplicada sobre la muestra; en la cual obtuvo un mejor rendimiento mediante asignación que usando el modelo predefinido en R (R Core Team, 2021) para los diccionarios anteriormente mencionados.

Otro hallazgo importante para resaltar en esta fase de procesamiento, es la conversión de los emojis en palabras; ya que posibilito que todos los símbolos y caracteres especiales fueran puntuados de manera correcta y así aumentaran la cobertura de los diccionarios hacia cada uno de los corpus, logrando tener un contexto más preciso sobre el mensaje enviado, dada la importancia de los emojis en la transmisión de sentimientos y emociones en el mensaje que el usuario quiere comunicar a través de la red social.

En este sentido, una aplicación notable es la homologación del diccionario Afinn con la escala NPS, ya que esto permitió ajustar la puntuación promedio de cada tuit al estándar y determinar la cantidad de los usuarios y tuits promotores, neutros y detractores de cada empresa. Con lo cual se pudo conocer la proporción y evaluar con un método adicional la percepción del servicio a nivel de tuit y usuario con su respectivo puntaje. Dentro de los resultados obtenidos, se destaca que en promedio el 66% de mensajes y usuarios de cada una de las empresas son detractores de estas; esto a causa del uso de esta red social como canal de atención al cliente relacionadas con solicitudes de información, estados de casos y PQRs. Mediante el diccionario BING, se logró identificar que en promedio el 63% de los tuits para cada empresa están etiquetados como negativos, lo que secunda que los mensajes analizados corresponden a una no

muy buena percepción. Respecto al diccionario NRC, más del 50% de los tuits asocian sentimientos relacionados a tristeza, odio y confianza (sentimientos que no son buenos según la definición del diccionario); lo cual apoya la teoría de que la percepción del servicio es negativa.

Con los resultados obtenidos a partir de los gráficos de frecuencia y nubes de palabras, se logró identificar los términos por empresa que utilizan los usuarios al momento de comunicarse a través de esta red social, destacando que el top 3 de las mismas se asocian al servicio; lo cual combinado con los resultados de las nubes de palabras, permitió establecer relaciones entre estos vocablos con los barrios y municipios, además de identificar otras posibles áreas de servicios como por ejemplo: facturación, cortes, reconexión y mantenimiento de infraestructura pública (lo cual se confirma más adelante).

Uno de los grandes aportes de esta investigación está relacionado con la exploración temporal de cada corpus y la manera en que el análisis de sentimientos realizado cambia a través de los meses del estudio, permitiendo encontrar el flujo de mensajes, tráfico sobre cada empresa y la posibilidad de que exista alguna tendencia o estacionalidad de estos. Un hallazgo importante es que la cantidad de mensajes para cada empresa es alta los lunes, martes, viernes y sábado, además que las horas que presentan mayor tráfico están comprendidas entre las 08:00 horas y 00:00.

Por otro lado, cada empresa presenta un comportamiento particular; y por ende una estacionalidad diferente que se puede evidenciar en los diferentes puntos máximos donde se reciben los mensajes; se resalta el ajuste realizado a los datos del acueducto para normalizar los valores atípicos y así visualizar la serie temporal de manera detallada.

La evolución de sentimientos aplicado al diccionario BING, permitió identificar que en promedio el 13% de los tuits de cada corpus fue descartado de la asignación; esto a razón de los términos que contiene cada tuit y donde se resalta que los meses de mayo y junio presentan la mayor cantidad de comunicaciones negativa; mas exactamente los días 23 y 24. Respecto al diccionario NRC, para cada una de las empresas los sentimientos asociados a: tristeza, odio, confianza y anticipación son los que tienen los valores más altos mes a mes, con lo cual se deduce que estos sentimientos son constantemente para todo el periodo analizado. Finalmente, analizando el resultado de la evolución mediante el diccionario Afinn; se encontró que por medio del método de suavización la mayor cantidad de días de la investigación, tienen puntuación negativa y no presentan un comportamiento estacional que permita identificar la frecuencia con la cual se podría repetir dicho evento. En general, los resultados al usar los tres diccionarios dejan en claro que la percepción del servicio durante el periodo de tiempo analizado fue negativa y progresivamente los sentimientos asociados de manera poco positiva son crecientes al finalizar el mismo.

Una de las variables que juega un papel importante en el desarrollo de la investigación, correspondió a la definición de la cantidad de temas -k- con los que se aplica el algoritmo LDA para la consecución del resultado del *Topic modeling*, lo que, a su vez, posibilito sentar las bases para la construcción de los indicadores y métricas con las que se propone el análisis de la percepción de servicio.

Como se mencionó previamente, existen varios procedimientos para obtener el numero "óptimo de tópicos" (5.7.1 Numero óptimo de tópicos por empresa), uno de ellos, corresponde al método gráfico, implementado en esta investigación y a partir del cual, se concluyó que la utilización de la función "*FindTopicNumbers*" es bastante útil y puede ser una aproximación adecuada. Sin embargo, para las tres empresas consideradas, se tuvo que realizar un análisis posterior a cada grupo de vocablos para verificar que cada tópico tenga sentido, por lo que se infirió que el método gráfico, debe acompañarse de la interpretación de quien desarrolle este tipo de procedimientos, además de ser necesario hacer varias iteraciones, las cuales involucren números de tópicos cercanos al encontrado, para así obtener agrupaciones apropiadas.

Mediante el resultado del *Topic modeling*, se concluyó que existen algunas temáticas comunes para las tres compañías, lo que indico que los clientes del sector de servicios públicos de la ciudad de Bogotá utilizaron este canal para escalar solicitudes relacionadas con: Cortes en el servicio y solución de caso e informaciones de solicitudes. La primera presento una participación general de 4.47 mil tuits (20% de todo el corpus) y la segunda 5.1 mil comunicaciones (23% de todo el corpus). Por otro lado, la etiqueta "otros" únicamente se asignó a 2249 tuits, menos del 10% de la descarga de mensajes, esto reafirma lo anteriormente expuesto, en cuanto a que el rendimiento del modelo (algoritmo LDA) es una buena aproximación.

A nivel general el tópico dominante: solución de caso e información de solicitudes, tiene una tendencia a mantenerse constante y con un alto tráfico durante todos los días de la semana, sin embargo, se recalca que el martes y el lunes son los días en que los clientes mayor uso dan al canal para escalar sus consultas sobre este tema, otro tópico que también tuvo un comportamiento marcado es el de cortes en el servicio, por lo que se puede pensar que están estrechamente relacionados. Puntualmente, para el acueducto los días de mayor tráfico son jueves y viernes, para Enel domingo lunes y martes y para Vanti, martes y miércoles.

En cuanto al comportamiento mensual de cada tópico de cada empresa, se evidencio que los dominantes tienen tendencia no solo a aumentar en términos de contactabilidad, sino también, a mantenerse en el tiempo, es decir, mes a mes los grupos de tuits que mayor cantidad tiene siempre fueron los mismos,

por lo cual, se sugiere hacer análisis por parte de las empresas dado la temporalidad (seis meses) del presente trabajo, en donde puedan identificar más a fondo (a través de la aplicación del *topic modeling* pero en una capa adicional) las temáticas puntuales dentro de cada tópico dominante, para el acueducto: Cortes en el servicio, para Enel y Vanti: solución de caso e información de solicitudes.

La aplicación de los mapas de calor, posibilito la identificación de las diferentes temáticas y dolores del cliente, a través de estas, se evidencia que la percepción general no es la mejor, dado el incremento de los tuits negativos (como ya se ha explicado). Para el acueducto el tópico dominante fue: cortes en el servicio (dejando de lado solicitud de carrotanques por su comportamiento atípico) tiene una tendencia a aumentar sus tuits negativos, con una tasa mayor al 50%. En cuanto a Enel, se pudo decir que el mayor dolor se relaciona con cortes en el servicio, las comunicaciones negativas presentaban un comportamiento al alza en los primeros meses y luego a mantenerse constante, lo que indica también que la percepción no tiende a mejorar. Finalmente, para Vanti, el mayor inconveniente tuvo que ver con solución de caso e información de solicitudes, sin embargo, esta compañía es la que termino con una tendencia a la baja en cuanto a tuits negativos, por lo cual puede decirse que su percepción es la que puede mejorar.

En este sentido, se demostró que es posible plantear un modelo de métricas que permitan mejorar el seguimiento, basado en la obtención de la clasificación y categorización de los tuits según el tópico y la utilización del diccionario Bing, lo cual puede generar mucho valor dado que se está utilizando la voz de cliente como insumo principal para el seguimiento del desempeño de algún canal (en este caso Twitter), lo cual permite tomar acciones tácticas que alineen la operatividad con estrategias *customer centricity*.

Por otra parte, se concluyó que es viable utilizar el modelo propuesto para realizar una homologación con los principales pilares de la encuesta CIER, según los resultados, se logró equiparar 3 de los 5 principales atributos, además del plus de incluir el alumbrado público. Adicionalmente, se determinó que el área con más oportunidades de mejora correspondía a suministro de energía y la que mejor percepción tenía podría decirse que es factura de energía. Por último, cabe resaltar que, para mejores resultados, se puede investigar una homologación distinta en cuanto a la escala, así como enriquecer los diccionarios para que las puntuaciones de los tuits puedan variar y bajar los que se etiquetan como regulares.

Finalmente, se probó que a través el método utilizado en la investigación es posible transformar los datos en información, permitiendo asignar valores cuantitativos a datos que son meramente de tipo cualitativo y no estructurados. Esta conversión y asignación se realizó a través del uso de herramientas como R (R Core Team, 2021) y Power BI (Microsoft Power BI, 2022), con las cuales se construyeron

visualizaciones para un nivel gerencial, que permitan tomar decisiones basadas en números (*data driven*) y que, a su vez, estén orientadas en la mejora de la satisfacción y experiencia del cliente.

Esta investigación, permitió sentar una base sobre un posible método en que la información que se puede obtener a través de los canales digitales, en este caso Twitter, pudo ser explotada al aplicarse los métodos adecuados para aprovechar su máximo valor, transformándola en activos estratégicos de alta relevancia para las áreas comerciales, áreas de marketing y de *Business Intelligence*.

6.2 Recomendaciones y trabajos futuros

A continuación, se presentan algunos aportes relacionados con la minería de texto en las redes sociales, el etiquetado de datos, la evolución de sentimientos en series temporales que son de gran interés para futuras investigaciones en la generación de nuevo conocimiento apoyado en el manejo y explotación de este tipo de datos.

- **Idioma:** Realizar una investigación sobre la inclusión de términos propios del lenguaje español latino, permitirá tener una mayor cobertura de los corpus a analizar ante diferentes aplicaciones en las que se use minería de texto y análisis de sentimientos en este idioma.
- **Fuentes de información:** Considerar otras fuentes de información suministradas por cada una de las empresas, pueden incrementar el tamaño del corpus, aportando en la generación de un modelo automatizado de gestión de experiencia; en donde se involucren otros canales de atención, para identificar temas de interés general además de nuevo conocimiento que no solo las redes sociales pueden otorgar. Por otro lado, usar datos abiertos u otras redes sociales para este tipo de trabajos; permite comparar e integrar múltiples fuentes de información a partir de diversos canales de comunicación para contrastar los resultados obtenidos y expuestos en el presente trabajo.
- **LDA para cada tópico:** Con el fin de poder identificar con un mayor detalle las oportunidades de mejora y así mismo, poder hacer una homologación más detallada con la encuesta CIER (para el caso de Enel), se propone que se tome cada corpus de tuits de cada empresa por aparte (después de haber pasado por los diferentes niveles de limpieza) y realizar el procedimiento para encontrar el número de tópicos y la aplicación del algoritmo LDA para cada conjunto de comunicaciones. Se recomiendo que para trabajos futuros se pueda implementar este procedimiento y así tener un mayor detalle de los principales subtemas a

los que los clientes hacen referencia. En este orden de ideas y mediante la aplicación de lo anteriormente expuesto, se pueden llegar a realizar homologaciones bastante claras como la de CIER con los atributos IDAR.

- **Cuidades Inteligentes:** El uso de la minería de texto en redes sociales puede ser de uso potencial para proyectos relacionados con cuidades inteligentes; ya que al tomar el etiquetado de estos datos se pueden identificar problemáticas sociales relacionadas con la prestación de los servicios básicos y así generar proyectos de impacto que mejoren la calidad de vida de los usuarios. Así mismo, puede aplicarse a productos de consumo y de segunda o tercera necesidad, permitiendo determinar necesidades a satisfacer o deseos que los posibles consumidores puedan tener.
- **Machine Learning:** Considerar la automatización de los modelos y técnicas de minería de texto a un nivel no supervisado, permitirá crear algoritmos de autoaprendizaje que puedan interpretar de acuerdo con el contexto el mensaje que transmite cada usuario a cada organización y crear flujos de atención personalizada. Así mismo, esto puede ser usado para enriquecer los diccionarios usados para el análisis de sentimientos, entrenando los algoritmos para refinar y mejorar los resultados el *topic modeling.*
- **Identificar la ironía, sarcasmo y burla dentro de los textos:** Un avance importante que se puede dar en materia de minería de texto para precisar oraciones, tuits y mensajes en general es lograr identificar la ironía con la que el emisor del mismo se expresa para así obtener un contexto objetivo de las intenciones del mensaje y lo que realmente quiere transmitir al receptor; y que esto actualmente con el análisis de sentimientos es difícil de conocer, ocasionando que los resultados que se puedan obtener distorsionen la realidad de la finalidad de la comunicación.

7. Referencias

Ali, T., Ahmad, I., Ur Rehman, A., & Kamal, S. (2018). Understanding Customer Experiences through Social Media Analysis of Three Giants of Soft Drink Industry. *2018 5th International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC)*. doi:10.1109/BESC.2018.8697304

Ávila Rodríguez, M. (Septiembre de 2020). Análisis de tweets y su influencia en los seguros de vida en el ámbito colombiano. 1-121. Bogotá, Cundinamarca, Colombia: Escuela Colombiana de Ingeniería Julio Garavito. Recuperado el 09 de Octubre de 2021, de https://repositorio.escuelaing.edu.co/bitstream/handle/001/1237/%c3%81vila%20Rodr%c3%adguez%2c%20Maria%20Paula-2020.pdf?sequence=2&isAllowed=y

Ba, Y., & Lee , H. (20 de Noviembre de 2012). Sentiment Analysis of Twitter Audiences: Measuring the Positive or Negative Influence of Popular Twitterers. (ASIS&T, Ed.) *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY*, 2522-2535. doi:10.1002/asi

Bello-Orgaz, G., Menéndez, H., Okazaki, S., & Camacho, D. (2014). Combining social-based data mining techniques to extract collective trendsfrom twitter. *Malaysian Journal of Computer Science, 27*(2), 95-111. Recuperado el 10 de Septiembre de 2021, de http://ejum.fsktm.um.edu.my/ArticleInformation.aspx?ArticleID=1475

Bogotá Como vamos. (2021). *Resultados 2 Fase: Encuesta Virtual #miVozmiCiudad.* Encuesta, Bogotá. Recuperado el 18 de Septiembre de 2021, de https://s3.documentcloud.org/documents/20425406/bogota-resultados-2da-fase-mivozmiciudad.pdf

Chamlertwat, W., Bhattarakosol, P., Rungkasiri, T., & Haruechaiyasak, C. (2018). Discovering Consumer Insight from Twitter via Sentiment Analysis. *Journal of Universal Computer Science*, 973-992.

Cortez Reyes, R. A. (Junio de 2018). Extracción de conocimiento a partir de textos obtenidos de Twitter. (U. T. Salvador, Ed.) *Entorno*(65), 30-41. doi:http://dx.doi.org/10.5377/entorno.v0i65.6048

DANE. (31 de Diciembre de 2018). *Departamento Administrativo Nacional de Estadisticas*. Recuperado el 20 de Septiembre de 2021, de https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/censo-nacional-de-poblacion-y-vivenda-2018

Empresa de Acueducto y alcantarillado de Bogotá. (16 de Septiembre de 2021). *EAAB*. Recuperado el 16 de Septiembre de 2021, de https://www.acueducto.com.co/wps/portal/EAB2/Home/la-empresa/informacion-general/!ut/p/z0/04_Sj9CPykssy0xPLMnMz0vMAfljo8zizQKdDQwtDIz8DEyMnA0CgwOcgvxDnQ19jMz0C7IdFQFA0Q31/

Empresa de Acueducto y Alcantarillado de Bogotá. (18 de Septiembre de 2021). *EAAB.* Recuperado el 18 de Septiembre de 2021, de https://www.acueducto.com.co/wps/wcm/connect/EAB2/ec8d1be2-2b2a-40a3-96b7-0d3ffcefddce/7.+JULIO+2021+-PQR%C2%B4S.pptx?MOD=AJPERES&CACHEID=ROOTWORKSPACE.Z18_K862HG82NOTF70QEKDBLFL3000-ec8d1be2-2b2a-40a3-96b7-0d3ffcefddce-nJwBCdx

Enel - Codensa. (16 de Septiembre de 2021). *Enel*. Recuperado el 16 de Septiembre de 2021, de https://www.enel.com.co/es/las-companias/codensa.html#:~:text=Con%20un%2024%25%20de%20participaci%C3%B3n,el%20manejo%20de%20sus%20operaciones

Enel. (2020). *Informe de sosteniblidad 2020.* Bogotá: N.E. Recuperado el 18 de Septiembre de 2021, de https://www.enel.com.co/content/dam/enel-co/espa%C3%B1ol/sobre_enel/informes_sostenibiidad/2020/informe-de-sostenibilidad.pdf

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (Noviembre de 1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. *COMMUNICATIONS OF THE ACM, 39*(11), 27-34. Recuperado el 10 de Octubre de 2021, de https://dl.acm.org/doi/abs/10.1145/240455.240464

Innovare Pesquisa CIER. (25 de Octubre de 2021). *COCIER Juntos Progresamos.* Obtenido de COCIER Juntos Progresamos: COCIER Juntos Progresamos

Kuffo, L., Vaca, C., Izquierdo, E., & Bustamante, J. C. (2018). Know your customer: Detection of Customer Experience (CX) in Social Platforms using Text Categorization. *2018 IEEE International Conference on Big Data (Big Data).* Seattle, WA, USA: IEEE. doi:10.1109/BigData.2018.8622556

Kuo, W.-K., Riantama, D., & Chen, L.-S. (30 de Diciembre de 2020). Using a Text Mining Approach to Hear Voices of Customers from Social Media toward the Fast-Food Restaurant Industry. (MDPI, Ed.) *Sustainability*, 268-285. doi:https://doi.org/10.3390/su13010268

Microsoft Power BI. (2022). *Microsoft Corporation*. Obtenido de https://powerbi.microsoft.com/es-es/

Ngaboyamahina, M., & Sun, Y. (2019). The Impact of Sentiment Analysis on Social Media to Assess Customer Satisfaction: Case of Rwanda. *International Conference on Big Data Analytics* (págs. 356-400). Kobee: IEEE.

Niño Martinez, N., Vaca, C., Rios, B., & Rey, L. (2020). Minería de textos y análisis de redes sociales en twitter. En *La industria 4.0 desde la perspectiva Organizacional* (págs. 85-105). Bogotá. doi:http://dx.doi.org/10.47212/industria4.0-6

Ogudo, K., & Dahj Muwawa Jean, N. (2019). Sentiment Analysis Application and Natural Language Processing for Mobile Network Operators' Support on Social Media. *2019 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)*, (págs. 1-10). Winterton, South Africa. doi:10.1109/ICABCD.2019.8851052

Pro Colombia: Exportaciones, turismo, inversión y marca país. (18 de Septiembre de 2021). Recuperado el 18 de Septiembre de 2021, de https://www.colombiatrade.com.co/noticias/5-empresas-colombianas-que-se-han-reinventado-debido-la-situacion-actualhttps://www.colombiatrade.com.co/noticias/5-empresas-colombianas-que-se-han-reinventado-debido-la-situacion-actual

R Core Team. (2021). *R: A language and environment for statistical computing*. Obtenido de R foundation for statistical computing: https://www.R-project.org/

Ranjan, S., Sood, S., & Verma, V. (2019). Twitter Sentiment Analysis of Real-time Customer Experience Feedback for Predicting Growth of Indian Telecom Companies. *2018 4th International Conference on Computing Sciences (ICCS)*, (págs. 166-174). doi:10.1109/ICCS.2018.00035

Salaberry, N. (2020). ANÁLISIS DE CONTENIDO EN TWITTER Y EL AISLAMIENTO SOCIAL OBLIGATORIO. *REVISTA DE INVESTIGACIÓN EN MODELOS MATEMATICOS APLICADOS A LA GESTION Y LA ECONOMIA*, 1-15. Recuperado el 2021 de Septiembre de 21, de http://www.economicas.uba.ar/institutos_y_centros/revista-modelos-matematicos/

Sari, E. Y., Wierfi, A. D., & Setyanto, A. (2019). Sentiment Analysis of Customer Satisfaction on Transportation Network Company Using Naive Bayes Classifier. *2019 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM).* Surabaya, Indonesia: IEEE. doi:10.1109/CENIM48368.2019.8973262

Songpan, W. (7-9 de Junio de 2017). The Analysis and Prediction ofCustomer Review Rating Using Opinion Mining. *IEEE Computer Society*, 71-77. Obtenido de https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7965709

Super Intendencia de Servicios Publicos Domiciliarios. (1 de Enero de 2021). En 2020, Superservicios recibió más de 260 mil trámites y solicitudes de usuarios de los servicios públicos domiciliarios. Bogota, Cundinamarca, Colombia.

Superintendencia de Servicios Públicos Domiciliarios. (29 de Abril de 2020). CIRCULAR EXTERNA No.20201000000204. Bogotá, Cundinamarca, Colombia.

Superintendencia de Servicios Públicos Domiciliarios De Colombia. (2015). *Régimen Básico: Ley 142 de 1994, Ley 143 de 1994, Decreto 990 de 2002.* Bogotá: Imprenta Nacional de Colombia.

Twitter. (S.F de S.F de S.F). *Centro de Ayuda Twitter*. Recuperado el 30 de Septiembre de 2021, de https://help.twitter.com/es/rules-and-policies/twitter-api

Vanti. (2020). *Informe de Sostenibilidad 2020.* Bogota: Una Tinta Medios SAS. Recuperado el 18 de Septiembre de 2021, de https://www.grupovanti.com/wp-content/uploads/2021/05/Informe-de-Sostenibilidad-Vanti.pdf

WeAreSocial. (16 de Septiembre de 2021). *Digital 2021*. Recuperado el 16 de Septiembre de 2021, de https://wearesocial.com/digital-2021

Zhan, Y., Han, R., Tse, M., Helmi Ali, M., & Hu, J. (2021). A social media analytic framework for improving operations and service management: A study of the retail pharmacy industry. *Technological Forecasting & Social Change*, 1-14. doi:https://doi.org/10.1016/j.techfore.2020.120504

Zhiwei, L., & Park, S. (2015). What makes a useful online review? Implication for travel product websites. *Tourism Mangment 47*, 140-151. Obtenido de https://www.sciencedirect.com/science/article/abs/pii/S0261517714001903?via%3Dihub

8. Anexos

A. Anexo 1. Glosario

- Algoritmo LDA: Es un código con ciertos procesos matemáticos complejos, que permite aplicar el *topic modeling.* Como se mencionó previamente, para esta tesis se aplica mediante una función de R (R Core Team, 2021).
- Análisis de sentimientos: Método que se utiliza para asignar valores cuantitativos y cualitativos a un conjunto de términos y que permite saber que sentimientos, se asocian a un texto analizado. Puede ser: negativo, positivo, enojo, alegría o un valor cuantitativo que se mueva entre -1 a 1. Se apoya en el uso de diferentes diccionarios desarrollados para lograr asociar a cada palabra a cada etiqueta definida.
- API: Conjunto de códigos y contraseñas que permite conectar Twitter con determinado software de programación especializada, con el fin de poder descargar datos directamente de la red social.
- Corpus: Conjunto de tuits a analizar organizados de forma matricial, los cuales ya han pasado por cierta etapa de limpieza.
- Cuentas: Es el reflejo de la creación del perfil digital de los usuarios de Twitter. En este pueden involucrarse imágenes, descripciones entre otros. Aplica para cualquier tipo de persona natural o jurídica.
- Dashboard: Es una representación gráfica de algún conjunto de datos, que permite apoyar la toma de decisiones a partir de representaciones visuales y utilización de diferentes tipos de gráfico. Los cuales generan valor al explotar representaciones dinámicas y sencillas.
- *Dataframe* (DF): Es un conjunto de datos organizados matricial y estructuradamente que se utiliza en el lenguaje de programación de R (R Core Team, 2021).
- Diccionario AFINN: Es un método para realizar análisis de sentimientos que asigna a cada palabra un valor entre -1 y 1, en donde -1 es un valor muy negativo y +1 un valor muy positivo.
- Diccionario BING: Es un método para realizar análisis de sentimientos que se basa en asignar a cada palabra una categorización: positiva, negativa o neutra.
- Diccionario NRC: Es un método para realizar análisis de sentimientos que permite relacionar uno o más de siete sentimientos (*anger, anticipation, disgust, fear, joy, asdness, surprise, trust*) a cada termino de un texto analizado.

- Document Term Matrix (DTM): Es un conjunto de datos ya procesados, organizados de forma matricial en donde cada fila corresponde a un tuit y cada columna a una palabra. Si el vocablo se encuentra en la columna del tuit correspondiente, se diligenciará con 1 sino con 0.
- Documento: Para el presente análisis, un documento corresponde a un tuit.
- Encuesta CIER (Comisión de Integración Energética Regional): Como su nombre lo indica, es una encuesta netamente del sector energético, que permite conocer la satisfacción de los usuarios con respecto a ciertas temáticas relacionadas con calidad de servicio, interrupciones, infraestructura, medio ambiente, financiera, entre otros.
- Estacionalidad: Es un comportamiento de algún conjunto de datos que se repite cada cierto periodo de tiempo.
- Etiquetado: Para este escrito, se define como la categorización textual o asignación de algún tipo de valor ya sea a un documento o a un término.
- Formato TIDY: Es un tipo de formato de datos básico para la realización de cualquier tipo de análisis de datos no estructurados, especialmente procesamiento de análisis de sentimientos y LDA. Permite contabilizar rápidamente las palabras con mayor frecuencia a partir del DTM.
- IDAR: Conjunto de atributos o pilares principales que se utilizan en la encuesta CIER (se detallan en la sección 5.8 Encuesta CIER – Aplicación Enel).
- K – Numero de tópicos: Es la cantidad óptima de temáticas a evaluar al aplicar un modelo de *topic modeling.*
- Librería: Conjunto de funciones desarrolladas por usuarios que permite realizar acciones en R Studio; las cuales pueden ser, por ejemplo: visualización, análisis estadísticos, algoritmos complejos entre otras.
- NPS – Net Promoted Score: Es una de las principales escalas que se utilizan para cuantificar la satisfacción del cliente, se basa en determinar la cantidad de clientes que son negativos, positivos y neutros a partir de la pregunta: “En una escala de 1 a 10, ¿Qué tan probable es que el cliente recomiendo la empresa a un amigo o colega?”. Donde los detractores son los que contestan: seis o menos, los neutros siete u ocho y los promotores más de nueve.
- Nube de palabras: Es una forma de visualización basada en la cantidad de palabras que se repiten en cierto conjunto de datos. Las de mayor frecuencia, aparecerán con un tamaño más grande, mientras que las de menor frecuencia, se visualizarán con un tamaño pequeño.

- Numero de documento: En esta investigación, se identifica con un numero único que se asigna a cada tuit en cada corpus de cada empresa.
- Power BI: Software perteneciente a la empresa Microsoft, que permite la extracción, transformación y cargue de datos para posibles visualizaciones.
- Preprocesador: Herramienta desarrollada para facilitar la primera limpieza de los datos al eliminar los mensajes que puedan alterar el objetivo de la investigación, consolidar los tuits de los usuarios, además de posibilitar la estructuración de una base de datos a analizar.
- Promedios: Método estadístico que se obtiene al realizar una suma de cierta cantidad y luego dividirlo por el número de sumandos.
- R studio: Es un software especializado de código abierto que permite analizar, transformar, visualizar y crear algoritmos para transformar y procesar grandes volúmenes de datos.
- *Stopwords*: Son las palabras que no tienen ningún tipo de valor para análisis textuales y que vienen predefinidas en software como R (R Core Team, 2021).
- Tendencia: Es el comportamiento en el tiempo de cierto conjunto de datos, ya sean valores numéricos dados por la frecuencia de repetición de ciertos términos, o por valores cuantitativos determinados a través de distintos métodos.
- Término: Para la presente investigación, corresponde a las palabras o vocablos de cada tuit de cada usuario.
- *Topic modeling*: Es un algoritmo que permite identificar cuáles son las principales temáticas involucradas en algún conjunto de datos textuales. Se basa en la probabilidad de aparición de cada termino junto a otro.
- Tópicos: Corresponde a los principales temas a los que se hace mención en cierto conjunto de datos textuales.
- Tuits: Mensaje escrito de máximo 280 que los usuarios de la red social Twitter pueden publicar, para transmitir sus emociones, opiniones o pensamientos. Involucran el uso de emojis, caracteres especiales, ubicación entre otros.
- Twitter: Red social caracterizada por que los usuarios pueden transmitir sus pensamientos, opiniones, peticiones o solicitudes mediante tuits para comunicarse entre conocidos, amigos, familiares, entidades gubernamentales, empresas etc.
- Usuarios: Son las personas que utilizan ya sea un servicio (en el caso de la investigación público o algún tipo de red social) o algún producto de consumo.

B. Anexo 2. Resumen estado del arte

Tabla 22. Principales artículos encontrados en la construcción del estado del arte

Aplicación	Autor	Sector	Breve Descripción	Fuente
Framework Referencia	**Ngaboyamahina & Sun, 2019**	**Servicios y Turismo**	Modelo básico de referencia para las soluciones basadas en la utilización de datos abiertos	International Conference IEEE on Big Data Analytics
Framework Referencia	**Ávila Rodríguez, 2020**	**Seguros de Vida**	Modelo de redes neuronales y aprendizaje automático que permite entender y personalizar las tarifas asociadas a los seguros de vida	Tesis de grado Escuela Colombiana de Ingeniería Julio Garavito
VOC en medios digitales	**Songpan, 2017**	**Postventa Turismo**	Comparación reseñas de hoteles y su calificación mediante *Naive Bayes* y árboles de decisión	Journal Commputer Society
VOC en medios digitales	**Bello-Orgaz, Menéndez, Okazaki, & Camacho, 2014**	**Muebles y Decoración**	Comparación de las reseñas de clientes mediante la utilización de C.4.5 Tree, y *clustering* a través de: *Dirichlet Process Algorithm*	Malaysian Journal of Computer Science
VOC en medios digitales	**Kuo, Riantama, & Chen, 2020**	**Comidas Rápidas**	Principales tópicos en las reseñas de los clientes antes y después de la pandemia, comparación mediante *TD-IDF y LASSO*	Sustainability Journal
VOC en medios digitales	**Kuffo, Vaca, Izquierdo, & Bustamante, 2018**	**Postventa en diversas empresas**	Comparación y utilización de distintas técnicas para encontrar la mejor que permita realizar *clustering* de las reseñas de experiencia mediante *Random Forrest* y SVM	2018 IEEE International Conference on Big Data (Big Data)
Percepción de cliente	**Ba & Lee, 2012**	**Percepción figuras publicas**	Percepción de las figuras públicas, métricas de comparación y análisis en el tiempo. Utilización de análisis estadísticos de regresión y comparación.	Journal of the American society for information science and technology

Aplicación	Autor	Sector	Breve Descripción	Fuente
Percepción de cliente	**Niño Martínez, Vaca, Ríos, & Rey, 2020**	**Campañas Publicas**	Determinación de la efectividad de Campañas del Ejercito nacional de Colombia mediante Python y *Vader*	La industria 4.0 desde la perspectiva Organizacional
Percepción de cliente	**Salaberry, 2020**	**Salud Publica y Medidas del Gobierno**	Percepción de los ciudadanos argentinos respecto a las medidas de prevención del COVID por el gobierno mediante R	Revista de investigación en modelos matemáticos aplicados a la gestión y economía
Percepción de cliente	**Cortez Reyes, 2018**	**Tecnología y avances en IA**	*Data Mining* en R para conocer *trending topics* y la opinión de los temas más hablados relacionados con nuevas tecnologías, por ejemplo, inteligencia artificial.	Revista Entorno
Benchmarks	**Chamlertwat, Bhattarakosol, Rungkasiri, & Haruechaiyasak, 2018**	**Telefonía Móvil**	Productos de Tecnología: Comparación características básicas para conocer la percepción de los clientes en cuanto a los nuevos productos.	Journal of Universal Computer Science
Benchmarks	**Ali, Ahmad, Ur Rehman, & Kamal, 2018**	**Bebidas Gaseosas**	Análisis de competitividad de marcas de gaseosa basados en las reseñas y opiniones de los clientes en redes sociales.	5th International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC)
Benchmarks	**Ranjan, Sood, & Verma, 2019**	**Telecomunicaciones de India**	Estudio del impacto en los seguidores de las páginas de un nuevo competidor en el mercado, aplicando análisis de correlación y *TD-IDF*	4th International Conference on Computing Sciences (ICCS)
Benchmarks	**Sari, Wierfi, & Setyanto, 2019**	**Transporte Publico**	Identificación de la polaridad de estos sobre dos empresas de transporte privado de la India a través de plataformas Online, mediante *Naive Bayes, TF-IDF y RapidMinker*	International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM)

Aplicación	Autor	Sector	Breve Descripción	Fuente
Benchmarks	**Zhan, Han, Tse, Helmi Ali, & Hu, 2021**	***Retail* de Farmacias en *UK***	Integración de los dolores de los clientes con la operación, basado en el análisis de sentimientos (Lexicón), *topico modeling* (LDA)y modelos visuales (Heat Map)	Technological Forecasting & Social Change
Benchmarks	**Ogudo & Dahj Muwawa Jean, 2019**	**Redes Móviles en Sudáfrica**	Determinación y definición del nivel de detracción potencial y de promoción en tres operadores de redes móviles en Sudáfrica a través de EDA y R	International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)

Fuente: Autoría propia.

C. Anexo 3. Nubes de palabras en cada tópico

Tabla 23. Nubes de palabras por tópicos – Acueducto

N°	Tópico	Nube de palabras	Análisis
1	Solicitud de Carrotanques		A partir de la nube de palabras se identifica que las palabras con una frecuencia de aparición más alta corresponden a la etiqueta asignada. Se concluye que las solicitudes de “carrotanques” se concentran en “suba” y se le pide a la empresa para “barrios” completo. Palabras como "llega” y “llevamos” indican que se escala esta petición dado que hay cortes en el servicio (Tópico 2).
2	Cortes en el servicio		Teniendo en cuenta la gráfica se evidencian términos como “corte”, “cortes”, “llevamos”, lo que da a entender que se está hablando de una intermitencia en el servicio. Para este tópico, las localidades que más reportan son “Engativá” y “suba”; lo que coincide con las peticiones de carrotanques del tópico 1.
3	Daños en Infraestructura		Este tópico presenta palabras como “tiempo”, “trabajo” y “daño” lo que puede indicar que los daños en la infraestructura toman mucho tiempo para ser reparados. Una palabra interesante es “vergüenza” la cual indica malestar por parte de los clientes. Este tópico puede relacionarse con el 2, por lo que los daños en infraestructura generan intermitencia en el servicio.

N°	Tópico	Nube de palabras	Análisis
4	Obras en vía publica		De este grupo de tuits se puede identificar que los clientes dicen que hay una “mala” “atención” y ejecución de las obras en la ciudad. Las cuales se relacionan con: “tapas”, “alcantarillas”. Por su parte la palabra: “av”, “barrio” y “calle” indica que se reporta la ubicación de las obras y que estas pueden generar cortes.
5	Solución de caso e Información de solicitudes		El tópico 5 se relaciona con que no se emite “respuesta”, inclusive se piensa que puede ser que “nadie” “responde” tanto por temas asociados a “cortes” en el “servicio”, temáticas asociadas a “pagos” y “factura”. En general este conjunto de tuits indica que puede haber oportunidades de mejora en la atención que se brinda por el canal.

Fuente: Autoría Propia.

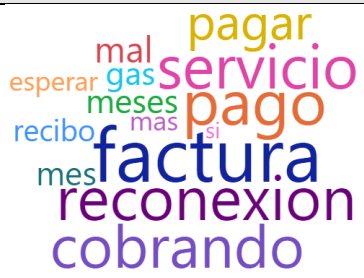
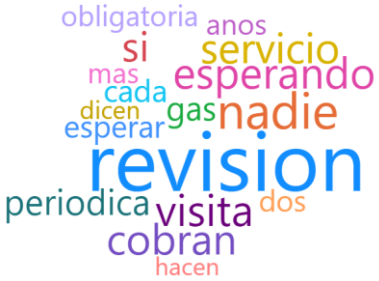
Tabla 24. Nubes de palabras por tópicos – Enel

N°	Tópico	Nube de palabras	Análisis
1	Cortes en el servicio		Se recalca que para este tópico en particular aparecen las palabras: “Cundinamarca”, “calera”, “municipio” y “vereda” lo que indica que a nivel rural se utiliza este canal como medio para reportar fallas también, en especial el municipio de la calera. Aparecen términos como “fallas”, “nuevamente”, “mantenimiento” y “servicio” lo que quiere decir que hay interrupciones en el servicio.
2	Solución de caso e Información de solicitudes		Para el tópico 2 se evidencian palabras como “respuesta”, “interno”, “necesito”, “mensaje”, “respuesta” lo que quiere decir que los clientes están esperando sus respuestas, que han remiten los datos de sus “casos” y que puede ser que requiera que se dé una respuesta efectiva frente a lo que indican mediante el canal.
3	Facturación		Para el tópico de facturación, se da la aparición de vocablos como “pagar”, “recibo”, “llega” y “factura”, los cuales giran alrededor de la temática de pagos y facturación. Se hace evidente que los clientes preguntan por el pago de sus “recibos” mediante esta red y que pueden tener inconvenientes dado que hay que términos de molestia como “enojado”.

N°	Tópico	Nube de palabras	Análisis
4	Alumbrado publico		Este tópico se relaciona con los cortes de luz, pero directamente en los “postes” y “luminarias”. Términos como “parque”, “urgente”, “Zona” y “publico” permiten pensar que los clientes reportan fallas de alumbrado público en la red social y que, en algunos casos, pueden ser de carácter “urgente”. Términos como “barrio” y “localidad” demuestran que se comparte la dirección de la novedad
5	Daños en Electrodomésticos		Como se puede apreciar este tópico está estrechamente relacionado con el tópico 1, dado que aparecen vocablos como: “cortes2, “daños”, “servicios” entre otros. Lo que indica que, dados los micro cortes e interrupciones en el fluido eléctrico, los “electrodomésticos” de los clientes se queman, por lo cual, el usuario opta por usar este canal como medio de resolución.

Fuente: Autoría Propia.

Tabla 25. Nubes de palabras por tópicos – Vanti

N°	Tópico	Nube de palabras	Análisis
1	Facturación		Dentro del tópico 1 se evidencian términos como “cobrando”, “pago”, “recibo” y “pagar” además de “factura” y “reconexión” lo que permite inferir que los clientes tienen novedades con el pago no solo de su factura mensual, sino de reconexiones que posiblemente se cobran sin haber sida requeridas por el cliente.
2	Cortes de servicio		A nivel general la palabra que más resalta es “servicio” al igual que “pésimo” y “cortar” lo que indica que evidentemente los usuarios reportan sus fallas a través del canal digital. Por otro lado, la aparición de “reconexión” nace como respuesta a que las personas solicitan que se normalice el servicio.
3	Revisión en terreno		De este conjunto de tuits se resaltan términos como “nadie”, “revisión”, “periódica” y “visita”, de esto se puede concluir que existe un incumplimiento en las visitas que se realizan “periódicamente” a los predios de los clientes. Vocablos como “esperando” puede ser un indicio de que los clientes reportan cuando se quedan con una revisión en terreno pendiente.

N°	Tópico	Nube de palabras	Análisis
4	Mala atención - PQRS		En este tópico se destaca que aparecen términos como “peor” y “enojado” lo que indica una molestia por parte del cliente. Por otro lado, se puede pensar que la mala atención en PQRs se relaciona directamente con temas de facturación, pagos y reconexiones. Se evidencia que este tópico involucra términos ubicados en los 3 grupos anteriores.
5	Solución de caso e Información de solicitudes		De este último grupo de tuits se identifican palabras como “abusivos” y “pésimos” lo que indica una molestia por parte del usuario en cuanto a que no se da respuesta sobre las posibles incidencias que haya reportado, esto se secunda por la aparición de términos como “información”, “respuesta” “cuenta”, “datos”, numero”, “línea” dado que son datos básicos para la atención de cualquier incidencia.

Fuente: Autoría Propia.

D. Anexo 4. Mapa de calor por tópico mensual – Acueducto

Tabla 26. Mapa de calor por tópico Acueducto

Tópico / Etiqueta	Mes						Total
Categorización Bing	Enero	Febrero	Marzo	Abril	Mayo	Junio	
Solicitud de carrotanques	204	204	342	158	355	2308	3571
Tuit Negativo	107	105	170	83	174	1162	1801
Tuit Neutro	46	55	86	33	91	594	905
Tuit Positivo	49	44	83	38	90	509	813
Tuit No Etiquetado	2	0	3	4	0	43	52
Cortes En El Servicio De Agua	191	228	336	215	523	1180	2673
Tuit Negativo	114	135	211	124	325	659	1568
Tuit Neutro	40	56	75	52	120	287	630
Tuit Positivo	36	37	46	34	77	217	447
Tuit No Etiquetado	1	0	4	5	1	17	28
Solución De Caso E Información De Solicitudes	164	220	307	183	331	668	1873
Tuit Negativo	108	138	191	98	193	389	1117
Tuit Positivo	28	43	60	50	78	133	392
Tuit Neutro	27	39	53	32	58	137	346
Tuit No Etiquetado	1	0	3	3	2	9	18
Daños En Infraestructura	138	129	283	113	285	823	1771
Tuit Negativo	63	54	183	61	134	381	876
Tuit Positivo	47	47	56	28	96	227	501
Tuit Neutro	28	27	40	22	52	196	365
Tuit No Etiquetado	0	1	4	2	3	19	29
Otros	163	134	200	120	269	616	1502
Tuit Negativo	55	51	49	55	58	234	502
Tuit Neutro	41	32	75	25	89	159	421
Tuit Positivo	41	29	55	31	61	150	367
Tuit No Etiquetado	26	22	21	9	61	73	212
Obras En Vía Publica	96	169	155	147	223	299	1089
Tuit Negativo	46	77	56	71	102	141	493
Tuit Positivo	25	53	44	31	74	76	303
Tuit Neutro	24	39	55	43	46	77	284
Tuit No Etiquetado	1	0	0	2	1	5	9
Total	**956**	**1084**	**1623**	**936**	**1986**	**5894**	**12479**

Fuente: Autoría Propia.

E. Anexo 5. Mapa de calor por tópico mensual – Enel

Tabla 27. Mapa de calor por tópico Enel

Tópico / Etiqueta Categorización Bing	Mes						Total
	enero	febrero	marzo	Abril	mayo	Junio	
Cortes en el servicio	219	238	398	414	417	432	2118
Tuit Negativo	149	155	278	282	311	309	1484
Tuit Neutro	36	48	64	70	51	70	339
Tuit Positivo	34	34	53	61	55	53	290
Tuit No Etiquetado	0	1	3	1	0	0	5
Solución de caso e Información de solicitudes	203	244	393	319	376	340	1875
Tuit Negativo	123	139	246	187	230	228	1153
Tuit Neutro	34	53	83	75	75	55	375
Tuit Positivo	46	52	64	57	71	57	347
Daños en Electrodomésticos	136	156	295	284	256	397	1524
Tuit Negativo	82	97	185	167	157	205	893
Tuit Neutro	36	42	70	80	64	138	430
Tuit Positivo	18	16	39	37	35	52	197
Tuit No Etiquetado	0	1	1	0	0	2	4
Facturación	120	141	204	173	262	213	1113
Tuit Negativo	76	80	121	96	134	118	625
Tuit Neutro	27	38	45	51	85	51	297
Tuit Positivo	17	23	38	26	39	44	187
Tuit No Etiquetado	0	0	0	0	4	0	4
Alumbrado publico	104	135	189	210	214	229	1081
Tuit Negativo	61	72	99	118	105	109	564
Tuit Neutro	27	38	50	56	66	60	297
Tuit Positivo	16	25	38	36	43	58	216
Tuit No Etiquetado	0	0	2	0	0	2	4
Otros	66	51	130	107	174	119	647
Tuit Negativo	17	26	47	31	41	28	190
Tuit Neutro	22	10	34	33	43	26	168
Tuit Positivo	18	8	29	22	59	31	167
Tuit No Etiquetado	9	7	20	21	31	34	122
Total	**848**	**965**	**1609**	**1507**	**1699**	**1730**	**8358**

Fuente: Autoría Propia.

F. Anexo 6. Mapa de calor por tópico mensual – Vanti

Tabla 28. Mapa de calor por tópico Vanti

Tópico / Etiqueta Categorización Bing	Mes						Total
	Enero	Febrero	Marzo	Abril	Mayo	Junio	
Solución De Caso E Información De Solicitudes	**129**	**144**	**197**	**88**	**168**	**95**	**821**
Tuit Negativo	89	100	132	52	114	64	551
Tuit Neutro	22	25	36	22	30	15	150
Tuit Positivo	18	19	29	14	24	16	120
Cortes En El Servicio	**60**	**74**	**100**	**54**	**128**	**87**	**503**
Tuit Negativo	44	53	73	42	89	69	370
Tuit Neutro	5	10	18	8	16	12	69
Tuit Positivo	11	11	9	4	23	6	64
Mala Atención – Pqrs	**108**	**84**	**136**	**36**	**70**	**58**	**492**
Tuit Negativo	68	60	98	26	51	44	347
Tuit Neutro	17	12	19	6	12	8	74
Tuit Positivo	23	12	19	4	7	6	71
Revisión En Terreno	**100**	**61**	**72**	**35**	**69**	**52**	**389**
Tuit Negativo	52	42	41	18	48	39	240
Tuit Positivo	29	10	14	10	17	10	90
Tuit Neutro	19	9	17	7	4	3	59
Facturación	**66**	**84**	**128**	**29**	**29**	**39**	**375**
Tuit Negativo	43	59	81	18	17	30	248
Tuit Neutro	11	15	27	5	5	2	65
Tuit Positivo	12	10	20	6	7	7	62
Otros	**34**	**35**	**53**	**13**	**35**	**23**	**193**
Tuit Negativo	16	16	12	2	13	9	68
Tuit Neutro	7	8	15	3	7	6	46
Tuit Positivo	6	3	12	6	9	6	42
Tuit No Etiquetado	5	8	14	2	6	2	37
Total	**497**	**482**	**686**	**255**	**499**	**354**	**2773**

Fuente: Autoría Propia.

G. Anexo 7. Métrica: Cantidad de tuits por hora por tópico

Tabla 29. Cantidad de tuits por hora y tópico

Empresa	Hora																							
Tópico/Etiqueta	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**
Acueducto	62	51	34	21	13	4	3	2	3	8	16	33	37	42	53	62	55	39	40	37	30	47	60	102
Cortes en el servicio de agua	31	28	15	11	8	4	1	2	1	8	16	33	29	30	41	33	40	33	32	22	29	20	20	41
Daños en infraestructura	28	28	24	12	4	3	1	1	1	2	5	6	10	14	18	20	22	17	25	15	11	14	24	29
Obras en vía publica	10	5	7	3	2	2	1	1	1	1	2	6	6	10	10	14	11	7	8	8	7	20	8	10
Otros	24	21	18	9	6	1	1	2	2	2	2	6	7	10	9	13	14	9	8	8	14	16	16	15
Solicitud de carrotanques	62	51	34	21	13	4	2	1	3	4	10	26	37	40	53	62	55	39	40	37	30	47	60	102
Solución de caso e Información de solicitudes	23	25	18	8	5	3	3	1	1	3	7	6	23	42	21	29	24	14	22	15	22	19	16	22
Enel	**23**	**25**	**16**	**16**	**8**	**5**	**4**	**2**	**4**	**3**	**7**	**14**	**17**	**20**	**17**	**31**	**19**	**27**	**24**	**27**	**21**	**23**	**20**	**34**
Alumbrado publico	11	12	8	6	3	2	1	2	1	1	2	12	9	8	8	8	9	6	9	8	8	8	7	15
Cortes en el servicio	23	25	14	16	5	2	1	1	4	2	4	14	17	20	17	31	19	21	24	27	21	23	20	34
Daños en Electrodomésticos	17	16	13	7	5	5	2	1	1	1	5	4	11	10	9	11	17	27	17	24	10	13	8	21
Facturación	14	7	7	7	4	2	2	2	1	2	2	4	7	8	13	7	10	12	16	9	8	7	6	10
Otros	5	6	5	3	3	2	1	1	0	2	3	4	3	4	8	4	12	5	6	4	5	3	5	3
Solución de caso e Información de solicitudes	17	20	16	8	8	2	4	1	1	3	7	6	13	17	14	15	14	16	13	9	18	20	15	20
Vanti	**7**	**6**	**7**	**3**	**2**	**21**	**2**	**1**	**1**	**1**	**2**	**3**	**4**	**11**	**14**	**12**	**13**	**13**	**12**	**11**	**17**	**13**	**13**	**6**
Cortes en el servicio	5	4	3	2	1	21	2	0	0	0	1	3	4	7	11	9	11	9	12	6	5	7	6	5
Facturación	2	6	1	1	1	0	1	0	0	1	2	2	2	5	7	8	13	7	6	6	6	6	4	3
Mala atención – PQRS	3	5	7	3	2	1	1	1	0	1	1	2	3	3	8	10	10	10	8	5	9	6	5	6
Otros	4	1	1	1	1	1	1	0	0	1	1	0	2	3	3	5	2	4	3	3	4	3	2	3
Revisión en terreno	3	3	5	2	1	2	1	0	0	0	1	1	3	3	5	3	8	5	5	7	8	8	7	6
Solución de caso e Información de solicitudes	7	4	2	2	2	2	1	1	1	1	1	3	3	11	14	12	13	13	12	11	17	13	13	6

Fuente: Autoría Propia.

H. Anexo 8. Métrica: Cantidad de tuits negativos acumulados por hora por empresa

Tabla 30. Cantidad de tuits negativos acumulados por hora por empresa

Categorización Bing	**Cantidad de Tuits Negativos**																							
Empresa	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**
Acueducto	10	19	27	31	34	36	38	39	40	42	46	52	61	72	81	93	102	111	120	128	136	144	153	163
Enel	9	18	25	29	32	34	35	36	38	39	41	45	50	56	63	70	78	84	92	98	105	113	120	130
Vanti	2	5	7	8	9	14	15	15	16	17	18	20	21	25	29	33	39	43	47	51	55	59	63	66

Fuente: Autoría Propia

I. Anexo 9. Métrica: Cantidad de tuits negativos en el tópico critico por empresa

Tabla 31. Cantidad de tuits negativos en el tópico crítico por empresa

Empresa	Tópico	Tuits negativos diarios
Acueducto	Solicitud de carrotanques	2
Enel	Cortes en el servicio	2
Vanti	Solución de caso e Información de solicitudes	1

Fuente: Autoría Propia.