Evaluación de la percepción de los servicios públicos en Colombia mediante Text Mining vía Twitter

Evaluation of the perception of public services in Colombia through Text Mining using Twitter

Carlos Eduardo Gomez Rivera
Maestría en Gestión de Información
Escuela Colombiana de Ingeniería Julio Garavito, Bogotá, Colombia
carlos.gomez-r@mail.escuelaing.edu.co

Dante Conti
Maestría en Gestión de Información
Escuela Colombiana de Ingeniería Julio Garavito, Bogotá, Colombia
dante.conti@escuelaing.edu.co

Juan Guillermo Jaramillo Yepes
Maestría en Gestión de Información
Escuela Colombiana de Ingeniería Julio Garavito, Bogotá, Colombia
juan.jaramillo-y@mail.escuelaing.edu.co

Victoria Eugenia Ospina Becerra
Maestría en Gestión de Información
Escuela Colombiana de Ingeniería Julio Garavito, Bogotá, Colombia
victoria.ospina@escuelaing.edu.co

doi: https://doi.org/10.36825/RITI.10.22.004

Recibido: Junio 07, 2022
Aceptado: Agosto 15, 2022

Resumen: En Bogotá, los prestadores de servicios básicos son: Enel, Vanti y El acueducto, grandes empresas del sector que deben caracterizarse por una operación *customer centricity*. Sin embargo, dado que estas compañías son las que mayor cantidad de quejas y reclamos presentan ante los entes regulatorios, la percepción del servicio puede no ser la mejor. Determinar la veracidad de esto, es el objetivo de la presente investigación, a través de la explotación de técnicas de minería de texto aprovechando la voz de cliente en los tuits de los usuarios; aplicando la metodología *Knowledge Discovery in Databases*, para generar la base de datos compuesta por 9071 tuits de las tres empresas. En la fase de limpieza de datos, se establecen pasos adicionales para refinar dicha base y consolidar los tuits de interés para la investigación. Lo anterior, permite obtener una explotación de los datos explicando los resultados a través de nubes de palabras, diagramas de frecuencia, análisis de sentimientos y los ratios entre la polaridad de los tuits. Los resultados permiten inferir que, para la ventana de tiempo en el cual se realiza el análisis, la percepción del servicio no es buena y que existen oportunidades de mejora para las tres compañías.

Palabras clave: *Twitter, Minería de Texto, Percepción de Servicio, Diccionarios - Lexicones, Análisis de Sentimientos.*

Abstract: In Bogota, the public service providers are Enel, Vanti and Acueducto, big companies that are characterized by a customer centricity operation. However, these companies are the ones with the highest number of complaints and claims reported to regulatory entities. The perception of service may not be the best. Determining the veracity of this fact is the objective of this research, through the exploitation of text mining techniques by taking advantage of the voice of the customer in the user's tweets; applying the Knowledge Discovery in Databases methodology to generate the database composed of 9071 tweets of these three companies. In data cleaning phase, additional steps are established to refine the database and consolidate the tweets of interest for the research. This allows to obtain an exploitation of the data explaining the results through word clouds, frequency diagrams, sentiment analysis and the ratios between the polarity of the tweets. The results enable inferring that, for the time interval in which the analysis was performed, the perception of the service is not good and so, there are opportunities for improvement for the three companies.

Keywords: *Twitter, Text Mining, Service Perception, Dictionaries - Lexicons, Sentiment Analysis.*

1. Introducción

El desarrollo tecnológico global y su influencia en la vida de las personas, ha impulsado la creación de nuevos hábitos y grandes cambios en su conducta a pasos agigantados. La transformación de estas costumbres, como por ejemplo la utilización de plataformas digitales para facilitar la compra de bienes y servicios, ha llevado a que el comportamiento y las tendencias de consumo varíen, modificando drásticamente el medio donde se crean, nacen y se desarrollan las compañías.

Con el uso de la tecnología, los clientes se encuentran conectados y en línea la mayor parte del tiempo, por lo que los datos disponibles aumentan, posibilitando nuevas formas para que las organizaciones se adapten al cambio y transformen sus procesos tradicionales en digitales generando nuevas oportunidades de negocio; y de la misma manera los clientes cambian sus hábitos de consumo y la forma en la que se comunican con las empresas.

Y es que aprovechar dichos volúmenes para genera nuevo conocimiento, se ha convertido en uno de los principales objetivos de las empresas y de la humanidad en general. El interés en producir conocimiento y nueva información ha pasado las barreras de lo cuantitativo al lenguaje natural: extraer conocimiento de corpus textuales como libros, periódicos, informes técnicos, entre otros [1].

Procesar este tipo de información no es tarea sencilla y de esto es precisamente de lo que se encarga la minería de texto, la cual es definida por [2] como una aplicación de la lingüística computacional y del procesamiento de datos que facilita la identificación y extracción de nuevo conocimiento a partir de colecciones de documentos o de corpus textuales; mediante la aplicación de distintas técnicas y principios teóricos desarrollados en otras disciplinas, funciona como una aplicación complementaria a la minería de datos para identificar patrones y asociaciones entre temas que a simple vista no tienen relación directa o indirecta.

El objetivo del presenta artículo, es la aplicación de una serie de técnicas de minería de texto a los tuits de las empresas del sector público de la ciudad de Bogotá, para así realizar un diagnóstico de la percepción de servicio. Se considera de alta relevancia, dado que previamente no se identifican antecedentes que relacionen el sector público con estos nuevos tipos de análisis. Con esto se pueden aplicar ejercicios de vigilancia del servicio, percepción de las instituciones y procesos de mejora continua.

En la literatura científica se pueden encontrar referencias y marcos de trabajo similares a la investigación planteada en el caso de Colombia. Se puede citar por ejemplo a [3] donde se plantean indicadores para la cuantificación y cualificación de algunas figuras públicas mediante Twitter, así mismo lo propuesto por [4], en su evaluación del rendimiento de diferentes técnicas de clasificaciones y clusterización combinadas entre sí, sobre un conjunto de tuits de clientes de IKEA. Bajo estas premisas, el artículo se presenta con la siguiente estructura:

En primer lugar, un panorama referencial del estado del arte; seguidamente, se presenta la metodología aplicada, en este caso *Knowledge Discovery in Databases* (KDD) y una descripción genérica de los tuits que se analizaron; posteriormente, se expone el procedimiento de preparación y limpieza de datos, detallando cada uno de sus niveles además de la forma en que se aplica cada uno de los diccionarios seleccionados: Bing, Afinn y National Research Council Canadá (NRC), a continuación, se expone el resultado de adaptar e implementar dichos

diccionarios, además de las gráficas de frecuencia y nubes de palabras, por último, se plantean las conclusiones y las propuestas de trabajos futuros.

2. Estado del arte

Como es bien sabido, la cantidad de información y datos durante los últimos años ha estado en constante incremento, la alta conectividad con la que cuentan las personas a distintos tipos de plataformas es uno de los principales factores que explica dicha expansión. Según [5] "los datos estructurados solo representan alrededor del 20 por ciento de los datos en todo el mundo" lo que constituye un alto porcentaje de lo que se encuentra disponible, por lo cual, la búsqueda de técnicas que permitan explotar estos datos ha cobrado una alta relevancia para la generación de valor en las compañías de cualquier tipo.

Uno de los principales métodos que ha sido estudiado es la minería de texto o *Data Mining*, la cual es un conjunto de metodologías catalogadas como "no supervisadas" y que como menciona [6] es un área de conocimiento que comprende diversas técnicas para extraer información a partir de grandes volúmenes de datos textuales, encontrados generalmente en formato no estructurado, la cual se ha aplicado en diversidad de campos.

Un ejemplo de ello se presenta en el campo psicológico, donde [7] compara la efectividad del algoritmo *Kmeans* y la combinación entre *Kmeans* y enjambre de partículas (PSO por sus siglas en inglés) para predecir la tendencia al suicidio sobre 600 poemas de 12 autores reconocidos en la literatura actual (50 poemas por escritor, donde la mitad de ellos cometieron suicidio); definiendo 4 tópicos centrales sobre los cuales se clasifican los textos. Dentro de los resultados se resalta que la combinación de algoritmos permite tener un porcentaje de efectividad más alto para predecir la tendencia al suicidio y que la escritura de un poema permite identificar la propensión de un autor a cometerlo.

Sin embargo, el mayor campo de aplicación se relaciona con la exploración de los datos generados en las redes sociales, ya que la transformación digital ha llevado a las empresas a convertir estos medios virtuales en canales de comunicación y atención con sus clientes, creando espacios en donde el contenido generado (micro reseñas o micro blogs) nace a partir de experiencias, pensamientos y sentimientos relacionados a los individuos [8] además de aprovechar el gran volumen de información que generan sus usuarios a partir de los comentarios que expresan opinión, ideas, preferencias, recomendaciones, entre otros [4].

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)" [9]. Lo que convierte a esta red social en una fuente de datos abiertos, en donde las personas puedan expresar sus opiniones libremente con respecto a productos, servicios, figuras públicas, entre otros; y que pueden ser analizadas con fines investigativos aplicados a áreas como mercadeo, servicio al cliente entre otros.

Dentro de la literatura se pueden encontrar *frameworks* de referencia que permiten tener una idea general sobre cómo pueden adaptarse soluciones de *data mining*, un ejemplo de ello se encuentra en el artículo: "*The Impact of Sentiment Analysis on Social Media to Assess Customer Satisfaction: Case of Rwanda*" [10], cuyo principal objetivo es el diseño de un sistema que permita procesar y aprovechar las reseñas de las diferentes redes sociales; el cual consta de 4 etapas: primero, recolección de datos desde múltiples fuentes, tales como Facebook, Blogs, Trip Advisor, Twitter, aprovechando el uso de las APIs. Segundo, manejo de grandes cantidades de datos a través de *software* especializados como lo puede ser *hadoop*; en esta etapa se realiza la limpieza de palabras sin valor, url, entre otros. Tercero, una etapa de extracción y clasificación, en donde se implementa la tokenización y conteo a través *Term Frequency-Inverse Document Frequency (TF-IDF)*. Finalmente, la cuarta etapa en donde se explota la visualización.

En [11] se concluye que, del esquema que se sigue para la implementación de una solución que implica la minería de texto, la etapa correspondiente al procesamiento del texto es la más relevante; dado que es cuando se filtran y seleccionan los términos que realmente generan valor y que serán evaluados a través de otras técnicas. En especial cuando se trabaja con tuits como fuentes de datos, ya que se presentan: "abreviaturas, escritura incorrecta, uso de términos propios de cada país y la inclusión de caracteres especiales, como, por ejemplo, símbolos, URL, entidades de retuit, entre otros". Se advierte también que, técnicas como *stemming* o lemmatizacion pueden generar pérdidas de datos, influyendo así en los análisis posteriores y la veracidad del resultado final.

La implementación de soluciones haciendo uso de los datos generados a partir de las redes sociales, ha

permitido a las empresas acortar su distancia de comunicación con los usuarios de sus productos o servicios; promoviendo una imagen positiva y estableciendo estrategias de negocio para tomar decisiones importantes. Este es el caso de la investigación de [12]; donde realizan un análisis de competitividad de 3 marcas de gaseosas utilizando los datos no estructurados provenientes de cuentas oficiales de Twitter de cada una; generando un modelo de evaluación de contenido publicitario y que con la aplicación de técnicas de minería de texto (usando el software Weka), plantean un análisis exhaustivo para identificar los sentimientos de los consumidores y conocer los puntos fuertes de cada marca y las oportunidades de mejora.

Las redes sociales han servido también para entender y cuantificar la percepción de campañas publicitarias y su impacto en la audiencia objetivo, tanto para entidades gubernamentales como para compañías privadas. Esto es lo que realizan en la investigación exploratoria de [13]; quienes utilizan el análisis cuantitativo y cualitativo a partir de la minería de datos, el procesamiento del lenguaje natural y el análisis de las redes sociales sobre 5911 tuits relacionados con la cuenta "@Ejercito_CAAID". Para llevar a cabo el análisis de sentimientos, se utiliza el modelo VADER del lenguaje Python para asignar la polaridad de cada texto utilizando herramientas para traducir las palabras al lenguaje español sin sacrificar la calidad del análisis de textos. Dentro de los resultados obtenidos, se destacan los *hashtags* más relevantes y la evaluación del impacto a través de una línea de tiempo de los comentarios promotores y detractores que se utilizan en la publicidad de las campañas del Ejercito Nacional de Colombia.

Parte de las investigaciones que se han realizado en minería de texto, hacen énfasis en encontrar los métodos y técnicas; además de la combinación adecuada entre ambas para generar el mayor valor posible a esta información y esto es lo que plantea [4] en su investigación sobre una muestra de 100 tuits acerca de las opiniones de los clientes sobre la marca IKEA; donde se seleccionan cuatro técnicas de clasificación y tres de clusterización para determinar el mejor rendimiento sobre este conjunto de datos. El cual se mide mediante 6 algoritmos para cuantificar las medidas inter y extra-clúster. Dentro de los resultados, se logra generar una red de usuarios que influencian mediante sus reseñas positivas o negativas, a otros usuarios en sus publicaciones.

Otro estudio que hace uso de las reseñas online se evidencia en [14]; en donde demuestran un esquema para aprovechar la voz de cliente desde las opiniones online de 3 grandes empresas de comida rápida. El esquema propuesto es muy similar al mencionado anteriormente, comprendiendo también la medición de la eficiencia del modelo. Por otro lado, se incluyen algoritmos como LASSO y arboles de decisión, para definir los principales tópicos y palabras claves. Se concluye que, la información contenida en la web puede utilizarse como insumo básico para aprovechar la voz del cliente y evitar el gasto de recurso en la evaluación y envió de encuestas de satisfacción.

Una de las grandes aplicaciones que se tiene de la minería de texto es la evaluación de la percepción, a través del aprovechamiento del contenido de reseñas y tuits en gran diversidad de temáticas, permitiendo cuantificar y cualificar los sentimientos expresados por las personas. Esto se evidencia en [8], en donde se concluye que, de 3000 tuits analizados relacionados con la expansión y control de la pandemia en Argentina, el 49.8% corresponden a un sentimiento negativo, el 40.6% a uno neutro y solo el 9.6% a algo positivo. Por lo que se infiere, que las personas no se identifican con las medidas y con la situación generada por el virus.

Otros autores proponen análisis aún más específicos, como lo son [3], quienes a partir de los tuits de ciertas figuras públicas (por ejemplo políticos, músicos, dueños de grandes empresas, visionarios, entre otros) logran realizar un análisis temporal de la evolución del sentimiento que generan en la población, proponiendo métricas como: "*User sentiment score*, *tweet sentiment score*, *Positive-Negative tuit ratio*, *Positive-Negative user ratio*" y llegando a realizar análisis de correlación para conocer cómo situaciones del día a día en el mundo "real" pueden afectar la percepción en el mundo virtual. Se considera de alta relevancia, ya que es una base para empezar a tener métricas de comparación para evaluar la percepción de ciertos usuarios de Twitter (que puede ser inclusive una empresa) y compararlos con otros.

También se presenta un caso de ejemplo de la evaluación de percepción de las figuras públicas participantes en las elecciones de E.E.U.U 2020 en [15]. Se utiliza una API y un algoritmo de limpieza para descargar los tuits de los usuarios con los hashtags asociados a: Biden, Trump, Kamala y Pence. Una vez se tiene una muestra suficientemente grande, se toman 500 tuits aleatorios para definir los diccionarios y alimentar el algoritmo de clasificación, en este caso el tipo de tuit: noticia u opinión y si es adulación (positivo) o insulto (negativo). Así se aplica Naive Bayes para luego realizar técnicas de estadística inferencial a partir de una muestra de 30 comentarios, como pruebas hipótesis y de media de Fisher, para determinar que las medias de los elogios son iguales para todos los candidatos y por lo tanto no hay evidencia que los comentarios positivos favorezcan alguna figura política.

Se evidencia entonces que la minería de texto es una manera de aprovechar la voz del cliente. Sin embargo, no se encontraron evidencias en el sector de servicios públicos en Colombia (al menos en reportes o artículos de difusión pública y/o científica), lo que abre una oportunidad novel para esta investigación.

3. Materiales y métodos

Como se puede apreciar en la Figura 1, KDD (*Knowledge Discovery Databases*, por sus siglas en inglés) combina métodos automatizados para el modelado basado en datos [16]. Este método considera un proceso iterativo e interactivo con el conjunto de datos, que al aplicar los modelos adecuados que modelen los mismos; generarán conocimiento que puede ser usado para la toma de decisiones o simplemente estructurar esta información para que sea dispuesta en otra área o campo de conocimiento [17].

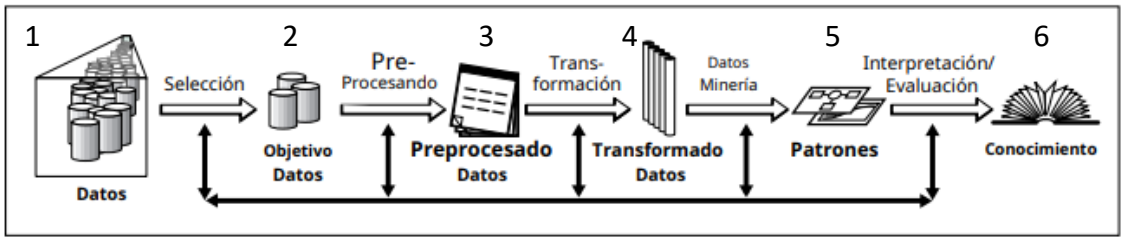

Figura 1. Proceso de descubrimiento de conocimiento método KDD. Fuente: [16].

Los datos con los que se realiza la investigación se extraen de las cuentas oficiales de cada una de las compañías prestadoras de servicios públicos; en la Tabla 1 se detalla la cantidad de tuits extraídos y los que se utilizan en el horizonte de tiempo de la investigación (1 de enero de 2022 hasta el 1 de marzo de 2022) la muestra que se toma corresponde a todos los tuits de los usuarios que escribieron a las cuentas por lo cual se considera conciso con el alcance de la investigación, dado que es un diagnóstico que permite tener un marco de referencia para aplicar soluciones que involucren *Data Mining* en este sector.

Tabla 1. Cuentas para realizar descarga de tuits.

Servicio Publico Básico	Empresa Prestadora	Cuenta(s) de Twitter	Cantidad de Tuits
Energía eléctrica	Enel Colombia	@ codensaservicio @EnelColombia @EnelClientesCo	3675
Gas natural	Vanti	@grupovanti	1619
Acueducto	Acueducto y Alcantarillado de Bogotá	@AcueductoBogota	3777

Fuente: Elaboración propia.

4. Resultados

La herramienta utilizada como soporte de cálculo, procesamiento y modelamiento de los datos es R [18].

1. Extracción de datos: Se realiza a través de la función *search_Tuits*, con ventanas temporales de 8 días.
2. Preparación de datos: Para mejorar el reconocimiento de las palabras en español "Latino Colombiano" en los tuits de los usuarios, se genera una etapa adicional como aporte novel dentro del caso de estudio.
   a) Preprocesador: Esta etapa permite organizar y estructurar los tuis de los usuarios de las diferentes cuentas, facilitando su posterior manipulación en R. Se resalta que, dada la eliminación de los mensajes de: cuentas corporativas, gubernamentales y distritales, los tuits estudiados por entidad se reducen en promedio 7.37%.
   b) Limpieza: Los tres niveles propuestos para limpiar los tuits, se desarrollan en su totalidad en R [18]:
      - Primer Nivel de Limpieza: A partir del *dataframe* consolidado en la etapa anterior, se eliminan: URL mencionadas por los usuarios, hashtag, espacios en blanco y doble

espacio. Los textos de los tuits se convierten en mayúsculas, con el fin de facilitar la eliminación de los *stopwords* en el siguiente nivel de limpieza.

- Segundo Nivel de Limpieza: Esta etapa corresponde a la mejora del reconocimiento de las palabras a nivel "Latino Colombiano". En un primer diagnóstico sobre los corpus, se realizaron graficas de frecuencia y nubes de palabra, encontrando una cantidad considerable de palabras sin valor, explicado en parte por el bajo reconocimiento de términos en español no castellano que tiene el software y el análisis de estas no genera valor real en términos de identificación de temas que tratan los clientes a través de Twitter. Para mitigar esto, se realizó una descarga de una muestra de 2000 tuits de la empresa Enel, las cuales se llevaron a formato *Tidy* y se exporta con su respectiva frecuencia de aparición. Se realiza la evaluación de cada palabra para definir el listado de términos sin valor. Por último, este listado se carga en R [18] y se eliminan los términos de los tuits de cada compañía.
- Tercer nivel de limpieza: Esta etapa corresponde a la limpieza final de datos que incluye las *stopwords* definidas por el sistema, números, puntuación y se convierten en minúsculas.

3. DTM – Formato Tidy: Se aplica la función *DocumentTermMatrix (DTM),* para generar la matriz de documentos y términos para cada empresa evaluada, tomando como argumento principal el corpus generado y teniendo algunos controles para no eliminar palabras de interés; específicamente el relacionado al *stemming* y a la longitud de ellas. Así mismo, se realiza la conversión a formato *Tidy*, permitiendo tener la frecuencia de cada palabra por tuit.

4.1. Análisis de sentimientos mediante aplicación de diccionarios

El software R [18] cuenta con la librería *Syuzhet,* la cual contiene la función *get_sentiment,* que como se menciona en [19] permite: *"Iterates over a vector of strings and returns sentiment values based on user supplied method…"*. A través de la aplicación de los métodos: Bing, Afinn y NRC en lenguaje español, permite realizar el análisis de sentimientos a un corpus definido mediante un algoritmo de búsqueda intensiva que compara los términos del diccionario con los del documento.

Los diccionarios base para la aplicación de la librería *Syuzchet*, pueden ser extraídos de la página: *http://saifmohammad.com/WebPages/lexicons.html* (el sitio contiene todas las fuentes idiomáticas con las cuales funcionan las librerías de R para la aplicación de minería de texto), donde se presentan como un conjunto de palabras sin modismos ni regionalismos y que se encuentra disponible en diferentes idiomas para su aplicación.

Para efectos de la presente investigación, se extrajeron estos diccionarios que utiliza R [18] y se realiza una revisión manual de las variaciones de las palabras que no son reconocidas (asignándole el valor de su palabra raíz o verbo); además de considerar las acentuaciones y puntuaciones de cada una para brindar un contexto correcto al ejercicio.

Por ultimo y con el fin de obtener una mayor asignación de las palabras en el corpus y un análisis de sentimientos más objetivo, se realizó una evaluación de la efectividad del método definido en R [18] (*Syuzhet*) contra un modelo de asignación que utiliza la función de R [18] *merge* para asignar y puntuar cada palabra de un corpus de 1928 tuits de la empresa Enel seleccionados aleatoriamente, basado en los diccionarios enriquecidos (Bing, NRC y Afinn), obteniendo los resultados en la Tabla 2.

Tabla 2. Resultado de la comparación de los métodos de análisis de sentimientos.

	Librería *syuzhet*		***Merge* (Modelo de Asignación)**	
	Palabras calificadas	Porcentaje calificación	Palabras calificadas	Porcentaje calificación
Bing	948	49.17 %	1367	70.90 %
Afinn	728	37.75 %	1226	63.59 %
NRC	845	43.82 %	732	37.97 %

Fuente: Elaboración propia.

Los porcentajes de calificación relacionan la cantidad de palabras calificadas por cada método sobre la totalidad de los términos en el corpus, basado en esto, se define aplicar el diccionario Bing y Afinn mediante *Merge* (Modelo de Asignación), mientras que para el diccionario NRC se utiliza la librería *Syuzhet* a través de la función *get_NRC_sentiment*. A continuación, en la Tabla 3, se define la manera de aplicar al corpus cada uno de los diccionarios:

Tabla 3. Ejecución de los diccionarios.

Diccionario	Objetivo	Criterios de ejecución
Bing	Clasificación de palabras de acuerdo con las categorías positivas, negativas y neutras.	Se priorizan los calificativos extremos (positivo y negativo). En caso de que el concepto no sea concluyente, se consulta el diccionario Affin para que, con sus puntajes, las palabras sean correctamente asignados
Afinn	Obtención *score polarity*, entre -1 (negativo) y 1 (positivo)	Se promedian los puntajes de cada palabra y así obtener un valor único para cada termino.
NRC	Clasificación de las palabras en una o varias de las siete categorías predefinidas.	Se aplica directamente la función *get_NRC_sentiment* sobre el corpus en lenguaje español.

Fuente: Elaboración propia.

Una vez finalizadas las aplicaciones de cada variación de las palabras de los tres diccionarios y teniendo en cuenta la utilización del modelo de asignación, se procede a crear un *dataframe* general con la asignación de los valores para cada palabra en cada tuit, el cual es la base para realizar el análisis de sentimientos.

4.1.1. Diccionario Bing y sus indicadores

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 relevante la propuesta ejecutada por los autores [3], los cuales proponen los siguientes indicadores.

- Para tuits:
  - TSS (*Tweet Sentiment Score*): Relación entre la cantidad de palabras positivas y negativas de cada tuit de cada usuario de cada una de las tres compañías.
  - TP (*Tweet Polarity*): Esta variable toma el valor de 1 o 0, siendo 1 si el TSS es mayor o igual a 1 o 0 si es menor a uno.
- Para usuarios:
  - USS (*User Sentimente Score*): Relación entre la cantidad de palabras positivas y negativas de cada usuario de cada una de las tres compañías.
  - UP (*User Polarity*): Esta variable toma el valor de 1 o 0, siendo 1 si el USS es mayor o igual a 1 o 0 si es menor a uno.
- Ratios:
  - PN_Tuit_Valor (*PN_TV*): Relación entre la cantidad de tuits positivos sobre la cantidad de tuits negativos.
  - Ratio_PN User_Valor (*PN_UV*): Relación entre la cantidad de usuarios catalogados como positivos sobre la cantidad catalogados como negativos.

Estos se acogen, adaptan e implementan en el corpus; partiendo de la asignación que se realizó con el diccionario Bing. Los resultados obtenidos se presentan más adelante en la seccion 4.2.1. Diccionario Bing.

4.1.2. Diccionario Afinn y sus indicadores

Este diccionario asocia a cada palabra un número que varía entre -1 y 1 dependiendo si se considera un término positivo o negativo. Esta cuantificación es importante en la investigación, dado que permite realizar la

homologación con la escala NPS (*Net Promoter Score*), indicador que se relaciona con la satisfacción del cliente y se basa en categorizar a los clientes en: promotores, neutros y detractores.

Con lo anterior, se proponen dos indicadores para aplicar la clasificación anterior a los tuits y a los usuarios. En la seccion de resultados se presentan los hallazgos de aplicar esta homologación.

- NPS por tuit: se realiza un conteo a nivel general de los tuits de cada categoría y se restan los promotores menos los detractores
- NPS por usuario: conteo de la cantidad de usuarios en cada categoría y se restan la cantidad de usuarios promotores menos los detractores.

4.1.3. Diccionario NRC

El diccionario NRC se basa en categorizar las palabras en uno o más de las 8 categorías de sentimientos que son: *anger, anticipation, disgust, fear, joy, sadness, surprise, trust.* En la sección de resultados, se presenta de manera gráfica los hallazgos para el corpus de cada empresa.

4.2. Resultado análisis de sentimientos

4.2.1. Diccionario Bing

Los resultados obtenidos de la aplicación de los indicadores y del diccionario en general se presenta en la Tabla 4.

Tabla 4. Aplicación Diccionario Bing Genérico.

	Palabras categorizadas				**Porcentaje de palabras categorizadas**			
Entidad	**Negativas**	**Neutras**	**Positivas**	**NA**	**Negativas**	**Neutras**	**Positivas**	**NA**
Acueducto	6660	13910	4258	9313	19,51%	40,74%	12,47%	27,28%
Enel	6878	14241	3717	4083	23,78%	49,24%	12,85%	14,12%
Vanti	4878	7288	2401	4025	26,24%	39,20%	12,91%	21,65%

Fuente: Elaboración propia.

A partir del análisis del resultado de los 3 corpus, se puede evidenciar que, la cantidad de términos negativos supera en un alto porcentaje a los positivos, lo que permite secundar que la percepción puede estar estrechamente relacionada con una percepción negativa hacia cada empresa. En promedio, el 43% de los términos no se asocian a un sentimiento, esto se debe a que se tuitean una gran cantidad de palabras que realmente no transmiten valor para ser consideradas en el análisis de sentimientos.

Por otro lado, el total de vocablos que no se encuentran en los diccionarios y que son categorizados como NA es del 21%, donde se encuentran términos que en su mayoría no tienen buena ortografía, así como abreviaciones que no se reconocen en el lenguaje formal y emojis que no son considerados como palabras. Analizando las cantidades porcentuales de términos negativos y positivos, se encuentra una razón aproximada de 2 a 1, es decir, que por cada 2 palabras negativas que el cliente manifiesta por este canal se encuentra 1 positiva.

El resultado de la clasificación de los tuits del diccionario Bing aplicado a los tuits de los usuarios se presenta en la Tabla 5, la cual se basa en el TP para determinar si el tuit es positivo o negativo.

Tabla 5. Categorización Tuits Diccionario Bing.

Entidad	**Negativos**	**Positivos**	**PN_Tweet_Valor**	**Cantidad de Tuits Categorizados**
Acueducto	960	821	0.85	1781
Enel	904	722	0.79	1626
Vanti	686	347	0.50	1033

Fuente: Elaboración propia.

Vanti contiene un índice de comparación menor entre la cantidad de tuits positivos y negativos (por cada 50 tuits positivos existen 100 negativos). Para las otras dos compañías el índice es muy similar, en donde para Enel cada 79 tuits positivos hay 100 negativos y para el Acueducto, por cada 85 positivos hay 100 negativos. Este indicador debe ser mayor o igual a uno para demostrar que los clientes tuitean más mensajes positivos que negativos.

Realizar el análisis anterior aplicado al número de usuarios que escriben a las cuentas es importante, porque permite determinar si los perfiles que se comunican con la compañía están o no satisfechos con el servicio, lo cual se presenta en la Tabla 6 Categorización Usuarios Diccionarios Bing (incluyendo el ratio de usuarios positivos y negativos (PN_User_Valor)).

Tabla 6. Categorización Usuarios Diccionario Bing.

Entidad	Positivos	Positivos / Negativos	Negativos	PN_User_Valor	Cantidad de Usuarios
Acueducto	167	423	393	1.05	983
Enel	427	182	322	0.82	931
Vanti	362	87	188	0.61	637

Fuente: Elaboración propia.

El Acueducto cumple con el criterio previamente descrito, lo que significa que por cada usuario con categorización de un tuit negativo hay un usuario que ha escrito un tuit positivo. Por su parte el corpus de Enel, permite evidenciar que la cantidad de perfiles que se comunican por este canal es muy similar al Acueducto, diferenciándose únicamente en 50 usuarios. Finalmente, Vanti posee el PN_User_Valor menor de todos, indicando que los usuarios que utilizan este medio se han comunicado y expresan un sentimiento negativo (por cada 2 usuarios negativos hay uno positivo).

4.2.2. Diccionario Afinn

El resultado de la aplicación de este diccionario se puede evidenciar en las Tablas 7 y 8 respectivamente, a nivel general se encuentra que todas las empresas presentan un NPS asociado a los tuits y a los usuarios negativo.

Tabla 7. NPS por Tuit Diccionario Afinn.

Entidad	Promotor	Neutro	Detractor	NPS_Tuit
Acueducto	13	24	3327	-0.9851
Enel	14	24	3086	-0.9833
Vanti	7	7	1517	-0.9862

Fuente: Elaboración propia.

Los tuits están asociados a la categoría detractor, lo que indica que son mensajes que realmente no son muy positivos pudiendo ser la gran mayoría quejas, situación que se puede corroborar al ver los resultados de la Tabla 8.

Tabla 8. NPS por Usuario Diccionario Afinn.

Entidad	NPS_Usuario
Acueducto	-0.9738
Enel	-0.9632
Vanti	-0.9766

Fuente: Elaboración propia.

Todo lo anterior se puede justificar dada la homologación realizada con la escala del NPS. Lo primero que se realiza, es que a cada término de cada tuit se le asigna un valor del diccionario Afinn, después se cuentan las frecuencias de repetición de cada palabra para multiplicarlas por su valor asignado y se realiza una suma de lo anterior para encontrar el valor cuantitativo de cada tuit.

Para tener una escala homologada a la del NPS, se toma el valor más bajo y alto encontrado anteriormente, con los cuales se calcula un rango para obtener los siguientes intervalos:

- El primero, parte del valor mínimo encontrado y le suma el 60% del valor del rango, así, todos los tuits que obtuvieron un valor menor a este se calificaron como detractores.
- El segundo, parte del valor mínimo y le suma el 80% del valor del rango, así, todos los tuits que obtuvieron un valor mayor a este se catalogaron como promotores.
- El tercero, corresponde a los neutros, y hace referencia a los tuits con valores asignados mayores al primer intervalo, pero menores al segundo.

El resultado se puede explicar dado que el intervalo para clasificar un comentario o un usuario como positivo es reducido (únicamente el 20% de los tuits o usuarios que estén por encima del rango de diferencia pueden entrar en la categoría de promotores). Por otro lado, se evidencia una gran cantidad de tuits para todas las 3 empresas que denotan un sentimiento negativo (lo que es corroborado por el análisis del diccionario Bing) y esto permite intuir que el porcentaje de detractores es muy alto, acorde también con los resultados analizados previamente.

4.2.3. Diccionario NRC

La Figura 2 muestra las gráficas con los resultados de la asignación de sentimientos del diccionario NRC a cada uno de los corpus categorizados en el siguiente orden: *Anger (A), Trust (T), Surprise (SP), Sadness (SD), Joy (J), Fear (F), Anticipation (ANT), Disgust (D)*.

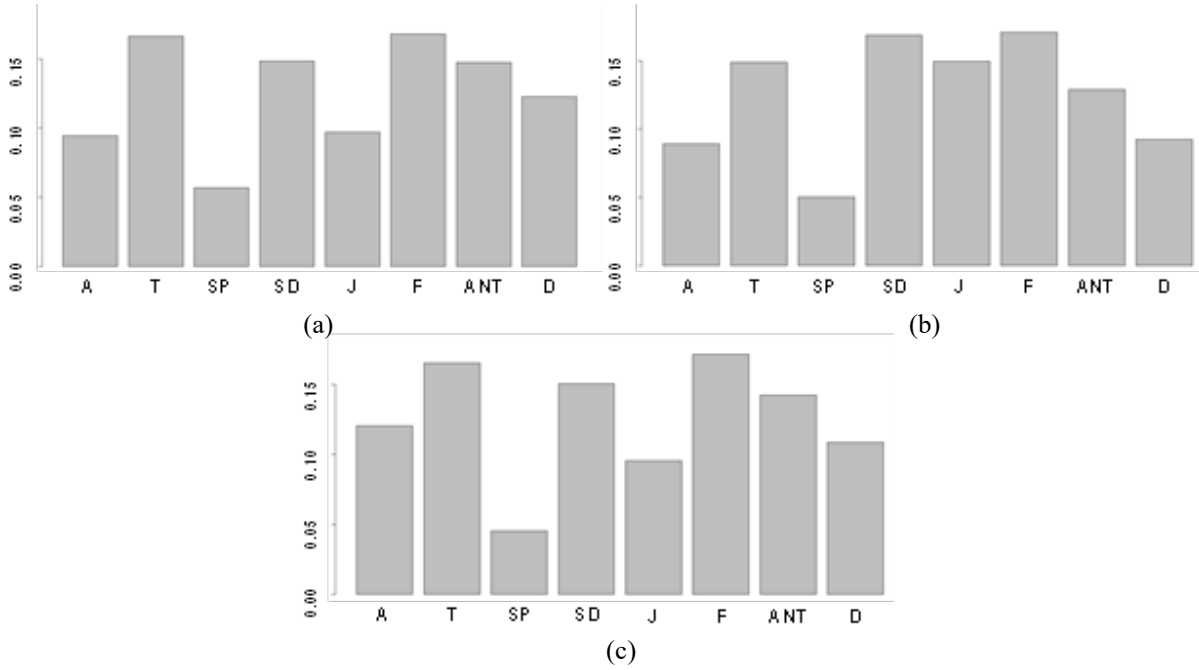

Figura 2. Resultados NRC: Acueducto (a), Enel (b) y Vanti (c).

Para el acueducto, se encuentra que las categorías *trust, fear* y *sadness* tienen una participación de por lo menos el 15%. Para los sentimientos *anger* y *disgust*, que corresponden a más del 30% de las palabras usadas por los usuarios para esta empresa, estas se asocian a un sentimiento negativo.

Para Enel, se puede evidenciar que los dos sentimientos que mayor participación tienen son *sadness* y *fear*, ambos con más del 15%. Por otro lado, *joy* puntúa con menos del 15% seguido por *trust*. Se mantiene la proporción en cuanto a *surprise* (siendo el de menor porcentaje de asignación) junto con *Anticipation* y *disgust*, siendo las categorías con un porcentaje respectivo de cerca del 13% y 10% detrás de las dos primeras correspondientemente.

Finalmente, para Vanti se puede ver que los sentimientos *Fear* y *Trust* son los que mayor participación presentan ambos con más del 15%. Una vez más, se puede explicar la aparición de la categoría *Trust* dado que las palabras se pueden calificar en más de un sentimiento. Se evidencia que *joy* presenta un porcentaje de asignación más bajo que *anticipation*, *disgust* y *anger*; donde en esta última categoría, se infiere que los clientes se encuentran bastante molestos dado su alto porcentaje de asignación. Finalmente, *surprise* presenta aun una participación más baja en comparación con los demás corpus.

4.3. Visualización

El principal objetivo de esta etapa corresponde a la identificación de ideas y conceptos sobre los principales temas que los clientes están tuiteando, mediante el uso de gráficos de frecuencia y de nubes de palabras para el corpus de cada empresa.

4.3.1 Graficas de frecuencia de palabras

Se emplearon las funciones "*count", "filter", "mutate" y "ggplot"* de R [18], teniendo en cuenta términos con una frecuencia mayor a 75 (se considera este número relevante dado que los corpus tomados fueron de 2 meses para cada compañía, por lo cual, son palabras que aparecen en promedio más de dos veces por día). La Figura 3 presenta un ejemplo de las graficas obtenidas, correspondiente a la empresa Vanti.

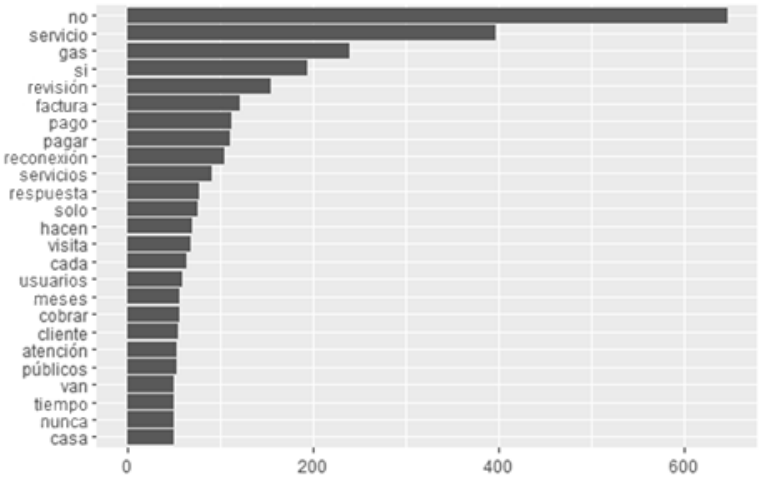

Figura 3. Grafica de frecuencia: Corpus Vanti.

Para el acueducto, se encuentra que es la que mayor cantidad de tuits posee en su corpus entre las 3 empresas; además de tener un número superior de palabras que se repiten a nivel general con una frecuencia de 1250. Al analizar los resultados del gráfico, se considera que los términos más relevantes son los siguientes: **Servicio, Barrio, Chapinero, localidad, sector, respuesta, corte, solución, Usme, sur, daño, tiempo, problema, información, factura, mentiras, avisaron, obras, ayer**, conociendo el contexto de la organización, dan entender algunos temas claves que los usuarios pueden estar indicando y que se presentan a través de este canal. Por otro lado, existen palabras que transmiten un malestar del cliente, y que, por tanto, indican una mala percepción del servicio, para este corpus son las siguientes: **Vergüenza, mentira, cara**.

Para Enel, se tiene que es el segundo corpus que más tuits posee entre las 3 empresas; teniendo una frecuencia máxima de 1000 palabras y que no presenta una variabilidad alta respecto a los términos usados en los mensajes. Dentro de los hallazgos, se resalta que la palabra **Pésimo** presenta la frecuencia más alta entre los términos asociados a comunicar malestar por parte de los usuarios. Otras de las palabras con alta frecuencia encontradas fueron **respuesta** y **solución**, que dada su importancia en el contexto del análisis; se puede interpretar que los clientes no tienen respuesta y solución a sus solicitudes, y esto puede ocasionar la aparición de términos como **llevamos**, **seguimos**, **nuevamente**, **sigue** entre otros.

Un hallazgo importante para Enel es que no aparecen nombres de los barrios de Bogotá, pero si **zonas** o **municipios** de **Cundinamarca**; como por ejemplo **La Calera** y **vereda**, y esto es indicio de que los clientes en

zonas rurales utilizan también este canal y realizan sus reportes por este medio.

Finalmente, Vanti es la compañía que menor cantidad de tuits tiene al poseer una frecuencia máxima de 600 palabras y dado el criterio con el cual se elabora la gráfica; se evidencia que la variabilidad es la más baja de las 3 empresas analizadas. Dentro de los hallazgos, se encuentra algunas palabras de alta relevancia como **Servicio, revisión, factura, pago, reconexión, respuesta, visita, usuarios, cobrar, nunca, tiempo**; y esto permite evidenciar que los términos usados en los mensajes no indican malestar; aunque se debe considerar la baja frecuencia de repetición. Con las anteriores palabras, se puede inferir que los clientes están contactando por este medio para reportar algún tipo de novedad relacionada con el pago, con la facturación y el proceso del cobro de servicio; además de palabras relacionadas con visitas a predios y esto permite inferir en que este trámite para los clientes no está contando con la relevancia que debería darle la empresa.

4.3.2. Nubes de palabras

Esta visualización complementa las gráficas anteriores, ya que posibilitan el análisis de las palabras a mayor profundidad; logrando identificar fácilmente nuevos términos que por su frecuencia no son evidentes. Se utiliza la función *wordcloud* del software R [18]. Los resultados se observan en la Figura 4.

Figura 4. Nube de palabras: Acueducto (a), Enel (b) y Vanti (c).

Para Enel, aparecen términos como **mal** sumándose a las palabras que demuestran inconformidad con los usuarios; donde se hace evidente que los clientes de los municipios aledaños a Bogotá hacen uso del canal para reportar sus incidencias, en especial los de **La Calera** y **Suba** para Bogotá. Los temas de mayor interés y que más se tuitean, están asociados a palabras **como postes, cortes** y **alumbrado**. Por otro lado, las palabras **necesitamos** y ayuda, permiten inferir que los clientes están requiriendo de solución o respuesta a sus inquietudes o inconvenientes de forma prioritaria.

Para el Acueducto, la gráfica permite complementar algunas palabras identificadas en el análisis de frecuencias, especialmente en temas relacionados con localidades y sitios en específico desde donde los clientes escriben. Por ejemplo, la localidad de **Kennedy** aparece junto con **Usme** y **Chapinero**, además de que los términos como **barrios, sector y localidad** indican la alta frecuencia con la que pueden aparecer estos términos en los tuits. Se evidencian términos relacionados sobre las posibles fallas más frecuentes, destacándose palabras como **tubo, tapa, alcantarilla** por lo cual se puede inferir un tema crítico para los usuarios que esta empresa debe considerar.

Finalmente, al analizar Vanti se pueden visualizar nuevas palabras como **ladrones**, lo que indica malestar a nivel general de los usuarios y que junto a términos identificados en el gráfico de frecuencia cómo **pésimo, peor, mal, cortaron, problema** y **error** se pueden etiquetar bajo la palabra **queja;** lo que indica que su atención es deficiente y utilizan este canal para comunicarlo hacia la empresa. Por otro lado, las palabras **cobrando**, **pagar, pago, factura** (que presentan mayor frecuencia), **cobran y cobro** demuestra que los clientes tienen alguna novedad con el proceso de cobro que se les hace de sus servicios.

5. 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. La información que se puede obtener a través de estos canales puede ser explotada si se aplican los métodos adecuados que permitan aprovechar el máximo valor y transformarlos en activos estratégicos, por lo cual, se convierte en un insumo de alta relevancia para las áreas comerciales, áreas de marketing y de *Business Intelligence*.

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, 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.

Cabe resaltar que, los resultados que se pretenden obtener están influenciados por algunos aspectos: la limpieza de datos, transformación de los datos, estructuración de los datos, modelos elegidos para cuantificar y cualificar los textos, entre otros. De estas variables se resaltan los problemas de ortografía, las letras repetidas al final de cada vocablo, abreviaciones y palabras unidas.

Realizar la limpieza de los datos en diferentes fases, permitió refinarlos para contar únicamente con los textos de interés y así considerar netamente los comentarios de los usuarios objeto de la investigación, lo que se demuestra en el primer nivel de limpieza al eliminar mensajes de cuentas corporativas y gubernamentales reduciendo el tamaño de los corpus; sin embargo, esto es necesario dado que estos tuits, pueden alterar los resultados del análisis de sentimientos y la evaluación de la percepción del servicio.

Es importante tener en cuenta la limitante de Twitter respecto a la cantidad de caracteres que puede tener un mensaje (permite escribir tuits de máximo 280 caracteres), por lo que es importante consolidar los hilos de usuario en un solo texto estructurado para tener el contexto general de la información que se quiere transmitir hacia las empresas.

De acuerdo con lo evidenciado en la literatura revisada; 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 o programas que permiten realizar el análisis de datos textuales, no consideran esto por su desarrollo en países con lengua inglesa. Incluir dentro de los diccionarios palabras latinas y colombianas con las cuales los usuarios se expresan hacia las empresas, ayudó a aumentar los puntajes de asignación por palabra para el análisis de sentimientos; mejorando considerablemente la precisión de los resultados contrastados con los diccionarios tradicionales que no consideran esta variable idiomática.

El análisis de datos de cada uno de los corpus de las tres empresas demostró que el top 5 de la mayor cantidad de palabras que se usan en los textos analizados, están asociadas a **No, Servicio, Barrio, Revisión** e incluye el **servicio** asociado para cada empresa (Luz, agua, gas). Esto permite inferir que, por este canal, los usuarios están hablando sobre las problemáticas que presentan con sus servicios públicos y que indican la ubicación para orientar a las personas que atienden estas solicitudes. Además, se resaltan otros términos en común de los tres corpus que transmiten molestia por parte de los clientes relacionados con fallas, con los cobros y con la facturación del servicio.

El uso de los diccionarios ayuda a comprender el comportamiento de este conjunto de datos dado las características de cada uno, mediante la cuantificación y cualificación de la percepción, a través de la homologación de los diferentes indicadores. Dentro de los resultados obtenidos con BING, las tres empresas presentan una inclinación hacia una polaridad negativa tanto en las palabras con las que expresa cada usuario, como en los tuits de cada uno; esto permite inferir que Twitter se está convirtiendo en un canal de atención de reclamaciones y crea una red de influencias entre usuarios que giran en torno a los mismos temas de interés. Por su parte, los resultados obtenidos mediante el diccionario Afinn, demuestran que están polarizados a que los usuarios y comentarios, corresponden a detractores de las empresas, lo que muestra insatisfacción en términos generales con los servicios prestados. Sin embargo, como se menciona en los trabajos futuros, el alcance y límites del indicador deben replantearse para evitar la polarización identificada. Con el diccionario NRC y teniendo en cuenta las asignaciones de los corpus a cada uno de los 7 sentimientos; los usuarios transmiten mensajes asociados a miedo, tristeza y confianza. Sentimientos que, como en los casos anteriores, corresponden a una insatisfacción general con el servicio.

Finalmente, a partir de los resultados obtenidos producto del análisis de una base representativa de información; se puede evidenciar que los mensajes que transmiten los usuarios a través de Twitter para estas tres empresas son negativos; y no solo comunican problemas o falencias del servicio sino también en aspectos relacionados con daños y facturación que son servicios de soporte que afectan directamente el servicio público

que principalmente se presta. Teniendo en cuenta lo anterior, se evidencia que la percepción de servicio por parte de los usuarios no es la mejor, en la ventana de tiempo analizada, se evidencian no solo gran cantidad de usuarios catalogados como detractores o negativos, sino también sentimientos y vocablos que de muestran molestia y poco aprecio por las compañías. Los ratios obtenidos (PN_TV y PN_UV) secundan lo anteriormente mencionado y son un indicador más de la situación evidenciada.

6. Trabajos Futuros

A continuación, se presentan algunos aportes y dificultades relacionados con la minería de texto en las redes sociales y 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:** Para cuantificar y cualificar los sentimientos y proceder con un análisis estadístico, descriptivo y cuantitativo; es necesario contar con diccionarios de términos robustos y que consideren palabras propias de cada país o región del país con el fin de dar mayor precisión a los textos en general y a cada una de las palabras que lo componen.
- **Asignación:** Como se explicó previamente, existen una cantidad considerable de términos catalogados como "N/A" y que no cuenta con una asignación cuantitativa ni cualitativa, dado los errores de ortografía, palabras incompletas entre otros. Futuras investigaciones se pueden orientar en la búsqueda modelos no supervisados aplicados a los corpus, que identifiquen estos escenarios y realicen las correcciones necesarias, para que, al momento de interactuar con los diccionarios, las polaridades y clasificaciones se enriquezcan y generen una mayor precisión de los resultados.
- **Emojis:** Considerar los emojis en el análisis de sentimientos a través de la minería de texto en redes sociales, ayudará a tener un mejor contexto sobre los mensajes que quiere transmitir cada usuario a cada empresa por este medio. Explorar alternativas con los emojis, reducirán la cantidad de caracteres sin puntuar en un análisis estándar de sentimientos y podrá aumentar la polaridad y los ratios de cada mensaje para un análisis más detallado.
- **Homologación:** Los resultados y aplicaciones de la minería de texto, son bastante utilizados en el estudio de la experiencia del cliente, por lo que, homologar la escala NPS a través de los resultados de la aplicación del diccionario Afinn, puede ayudar a identificar los usuarios promotores, neutros y detractores. Sin embargo, se debe explorar la manera de refinar la homologación con dicha escala, con el fin de obtener una categorización más precisa y no generar una polarización hacia uno de los sentimientos.
- **Fuentes de información:** Considerar otras fuentes de información suministradas por cada una de las empresas, pueden incrementar el tamaño del corpus de manera considerable, 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.
- **Encuesta CIER:** Actualmente, existe un informe de la Comisión de Integración Eléctrica Regional (CIER), el cual posee una evaluación del servicio al cliente dividido en 5 grandes pilares que, a su vez, poseen una serie de atributos para calificar y obtener una puntuación general del estado de la compañía. Con lo anterior, se debe explorar los resultados de trabajos futuros que apliquen a investigaciones relacionadas con minería de texto en empresas de servicios públicos energéticos, ejecutando modelos de clusterización y clasificación para contrastarlo con la encuesta CIER, y así tratar de obtener una calificación utilizando Twitter como fuente de información.
- **Secuencias Temporales:** Dentro de los datos que se pueden obtener de Twitter, existe la posibilidad de hacer un estudio de series temporales sobre el comportamiento de las publicaciones. Esto permite generar nuevo conocimiento sobre el comportamiento del canal, al analizar la evolución de los sentimientos temporalmente e identificar las tácticas que están mejorando la satisfacción del usuario, los sentimientos expresados en cada mensaje y dolores que estén afectando al cliente repetitivamente.
- **Alertas tempranas:** La clasificación de usuarios en positivos/negativos y promotores/detractores, puede ser estudiada con más detalle para lograr la implementación de un modelo en tiempo real que se ajuste a la naturaleza de los datos tomados de redes sociales, el cual, se enfoque en la generación de alertas tempranas en los procesos de negocio de cada compañía, para identificar posibles fallas y ofrecer soluciones agiles y confiables.

7. Referencias

[1] Montes Gomez, M. (2001). *Minería de Texto: Un nuevo reto computacional. Minería de Texto: Un nuevo reto computacional*. https://ccc.inaoep.mx/~mmontesg/publicaciones/2001/MineriaTexto-md01.pdf

[2] Eito Brun, R., Senso, J. A. (2004). Minería Textual. *El profesional de la información*, *13* (1), 11-27. http://profesionaldelainformacion.com/contenidos/2004/enero/2.pdf

[3] Bae, Y., Lee, H. (2012). Sentiment Analysis of Twitter Audiences: Measuring the Positive or Negative Influence of Popular Twitterers. *Journal of the american society for information science and technology*, *63* (12), 2522-2535. https://doi.org/10.1002/asi.22768

[4] 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. https://ejournal.um.edu.my/index.php/MJCS/article/view/6797

[5] TIBCO Software Inc. (2022). *¿Qué son los datos estructurados?* https://www.tibco.com/es/reference-center/what-is-structured-data

[6] Rivadeneira Zambrano, F. J., Vélez Flores, B. F., Pinargote Mendoza, W. J., Rivadeneira Zambrano, R. A., Carvajal Rivadeneira, S. M. (2020). Aplicación de minería de texto para el análisis de sentimientos del servicio de telefonía móvil en el ecuador. *Holos*, *7*, 1-16. http://dx.doi.org/10.15628/holos.2020.7994

[7] Powell González, J. E., Carrillo Ruiz, M., Somodevilla García, M. J. (2021). Agrupamiento de poemas de autores suicidas y no suicidas usando K-means y enjambre de partículas. *Revista de Investigación en Tecnologías de Información (RITI), 9* (18), 14-23. https://doi.org/10.36825/RITI.09.18.002

[8] Salaberry, N. (2020). Análisis de contenido en Twitter y el aislamiento social obligatorio. *Revista de investigación en modelos matemáticos aplicados a la gestión y la economía*, *7* (1), 1-15. http://www.economicas.uba.ar/wp-content/uploads/2016/04/Salaberry-Natalia.pdf

[9] Twitter. (2022). *Información sobre las API de Twitter.* https://help.twitter.com/es/rules-and-policies/twitter-api

[10] Ngaboyamahina, M., Yi, S. (2019). *The Impact of Sentiment Analysis on social media to Assess Customer Satisfaction: Case of Rwanda*. IEEE 4th International Conference on Big Data Analytics, Suzhou, China. https://doi.org/10.1109/ICBDA.2019.8713212

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

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

[13] Niño Martínez, N., Vaca, C., Ríos, B. P, Rey, L. J. (2020). Minería de textos y análisis de redes sociales en twitter. En M. A. Yandar Lobón, J. M. Moreno Ospina (Comp.) *La industria 4.0 desde la perspectiva Organizacional* (pp. 85-105). Artes y letras S.A.S. http://dx.doi.org/10.47212/industria4.0-6

[14] Chen, W.-K., Riantama, D., Chen, L.-S. (2021). Using a Text Mining Approach to Hear Voices of Customers from social media toward the Fast-Food Restaurant Industry. *Sustainability*, *13* (1), 268-285. http://doi.org/10.3390/su13010268

[15] Bernabe-Loraca, B., González-Velázquez, R., Carrillo-Canán, A., Granillo-Martinez, E. (2022). Sentiment Analysis and Multiple Means Comparison. *Computación y Sistemas, 26* (1), 91-100. https://doi.org/10.13053/cys-26-1-4155

[16] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Communications of the ACM*, *39* (11), 27-34. https://doi.org/10.1145/240455.240464

[17] Valcárcel Asencios, V. (2004). Data mining y el descubrimiento del conocimiento. *Industrial Data, 7* (2), 83-86. https://www.redalyc.org/articulo.oa?id=81670213

[18] R Core Team. (2021). *The R Project for Statistical Computing*. https://www.R-project.org/

[19] Jockers, M. (2020). *Package 'syuzhet'*. https://cran.r- project.org/web/packages/syuzhet/syuzhet.pdf.