Aplicación de la metodología CRISP-DM en la detección de necesidades ciudadanas en base al análisis de tweets

Edison U. SANANGO
Escuela Politécnica Nacional
Quito, Ecuador

María A. HALLO
Escuela Politécnica Nacional
Quito, Ecuador

RESUMEN: Los tweets han aumentado debido al uso de las redes sociales, en ellos se extrae tópicos y analiza sentimientos. En este artículo se detalla la aplicación de la metodología CRISP-DM en la construcción de un sistema de detección de necesidades de atención municipal en Quito, Ecuador en base al análisis de tweets. Se abordan las principales etapas de la metodología tanto en la extracción de tópicos como en su clasificación en necesidades de atención municipal y en la detección de sentimientos. En la extracción de tópicos el algoritmo utilizado fue Algoritmo de muestreo de Gibbs para el modelo de mezcla multinomial de Dirichlet (GSDMM) lo cual permitió detectar once tópicos de interés. En el análisis de sentimientos se entrenó un modelo de Regresión Logística para obtener la polaridad de cada tópico y se obtuvo una predominancia de sentimiento negativos.

Palabras Clave: CRISP-DM; extracción de tópicos; análisis de sentimientos; GSDMM; tweets.

Application of the CRISP-DM methodology in the detection of citizen needs based on the analysis of tweets

ABSTRACT: *Tweets have increased due to the use of social networks; topics are extracted from them, and feelings are analyzed. This article details the application of the CRISP-DM methodology in the construction of a needs detection system in Quito based on the analysis of tweets. The main stages of the methodology are addressed both in the extraction of topics and their sentimental polarity to classify them as needs. In the topic extraction, the algorithm used was the Gibbs Sampling Algorithm for the Dirichlet Multinomial Mixture Model (GSDMM), which resulted in eleven topics. In the sentiment analysis, a Logistic Regression model was trained to obtain the polarity of each topic and a predominance of negative sentiment was obtained.*

Keywords: *Topic modeling, sentiment analysis, CRISP-DM, GSDMM, tweets*

INTRODUCCIÓN

Con la globalización y el crecimiento de la tecnología, todos los sectores que conforman a la sociedad deben evolucionar [1]. Un aspecto fundamental en el cual se ha evolucionado gigantescamente son los datos, estos son generados de distintas fuentes como patrones de sueño, hábitos de vida, comportamiento de conducción, etc. [2]. Por ende, una institución que no innove ya sea, en optimización de procedimientos, nuevas formas de generar ingresos, reducción de costos, optimizar servicios, etc. no podrá evolucionar [3]. Por otra parte, es importante considerar el entorno en que la institución se encuentra. No todos los sectores tienen las mismas necesidades. En Ecuador no existe una cultura marcada que ponga su enfoque en el análisis de datos y la toma de decisiones en base a esto. Con el paso del tiempo son más sectores que optan por usar las tecnologías de la información como herramienta para mejorar sus servicios y operaciones.

Twitter como red social global es un medidor del impacto sentimental generado en personas debido a determinados eventos [4]. Al hacer uso de Twitter junto a la selección de un lugar o tema en concreto, se obtienen nuevas formas de entender a la sociedad y a su comportamiento. Un ejemplo es la medición de actitudes de personas en áreas urbanas en Estados Unidos [5]. Con esto se cortejó ciudades pequeñas y no estables, con ciudades grandes y estables.

A raíz de este análisis, se determinó que a pesar de las diferencias entre dichas ciudades las actitudes de las personas no cambian abruptamente posteriormente esto derivó en la mejora de políticas públicas que ayuden a mejorar la calidad de vida de las personas. En el desarrollo de ciudades inteligentes *(smart cities*) se utiliza el análisis de sentimientos para conocer la percepción de las personas con respecto a las acciones implementadas en determinada ciudad [6]. Así también, el identificar grupos determinados por raza, condiciones sociales, sector de residencia, etc., ayuda a los gobiernos a aproximarse a ellos de la mejor manera. Esto se puede evidenciar en el trabajo realizado por Murty et al. [7].

Entre otras investigaciones también se puede encontrar que el tono (sentimiento) puede influir en la participación de los ciudadanos en acciones planteadas por entidades gubernamentales. Sin embargo, se evidenció que no es el único factor que influye en su participación [8]. Esto planteó rutas a seguir en el estudio del comportamiento de las personas frente a lo que en estas épocas ya es la forma de comunicación más directa y efectiva, las redes sociales. La gestión municipal de Quito no es la excepción y hay estrategias que se pueden implementar para el correcto aprovechamiento de los datos que generan para el beneficio de la ciudadanía. Uno de los problemas del Municipio de Quito radica en que no se tiene un sistema conocido que detecte las necesidades de los ciudadanos de forma automatizada.

Este estudio busca descubrir las necesidades de la ciudadanía relacionadas con la gestión municipal en Quito basado en el análisis de tweets, esto permitirá al Municipio de Quito conocer las necesidades de las personas y actuar oportunamente. En este proyecto se consideran el análisis de sentimientos y la extracción de tópicos. El análisis de sentimientos involucra determinar la posición evaluativa que un determinado texto posee; puede ser positivo, negativo o neutro [9]. Por otra parte, la extracción de tópicos es el acto de diferenciar los términos de

ISBN: 978-1-950492-76-3
ISSN: 2771-6333

https://doi.org/10.54808/CICIC2024.01.150

un documento basados en su importancia en dicho documento o aquel que describa mejor la idea general [10].

La metodología que este proyecto sigue es CRISP-DM. Proceso Estándar de la Industria Cruzada para la Minería de Datos *(Cross Industry Standard Process for Data Mining*).

MARCO TEÓRICO

El proyecto aborda diferentes conceptos uno de ellos es el análisis de sentimientos que se relaciona con la evaluación mediante la aplicación del procesamiento de lenguaje natural [11]. El análisis sentimental se lleva a cabo mediante técnicas de aprendizaje de máquina como la regresión logística. La regresión logística es un algoritmo de aprendizaje supervisado dentro del mundo del aprendizaje de máquina usado para problemas de clasificación [12]. El resultado es medido con el área bajo la curva ROC. ROC es un gráfico (curva) que representa la proporción de verdaderos positivos frente a los falsos positivos. Esta medida es usada cuando se evalúa el desempeño de un clasificador [13].

Por otra parte, la extracción de tópicos consiste en extraer tópicos que definan un corpus. GSDMM (*Gibbs sampling algorithm for the Dirichlet Multinomial Mixture model*) es una técnica que permite extraer tópicos y puede trabajar con textos cortos, debido a las redes sociales es más común los textos con una longitud relativamente pequeña [14]. GSDMM asume que cada documento pertenece a un solo tópico, ya que es poco probable que un pequeño texto involucre a más de un tópico [15]. GSDMM puede ser evaluado mediante la coherencia de sus tópicos, al extraer tópicos no existe una manera específica de probar su efectividad. Sin embargo, la efectividad de cada tópico también queda a criterio del autor. Igualmente, se define como el promedio de similitudes de palabras en pares formadas por las palabras principales de un tema determinado [16].

METODOLOGÍA

La metodología CRISP-DM plantea una metodología estructurada, aunque flexible al mismo tiempo [17]. Por ende, la metodología que se va a adaptar en este caso suprime pasos no necesarios para el propósito de este proyecto. Además, se modifican otros pasos para adaptarlos al objetivo final.

Las etapas de CRISP-DM que se consideraron en este proyecto son: comprensión del proyecto, comprensión de los datos, preparación de los datos, modelamiento y evaluación. El ciclo de vida que comprende este proyecto se describe en la figura 1, aquí se presenta los pasos usados en el proceso desde la recolección de tweets, hasta la visualización de resultados que se compone de un tablero de mandos (*dashboard*) interactivo web.

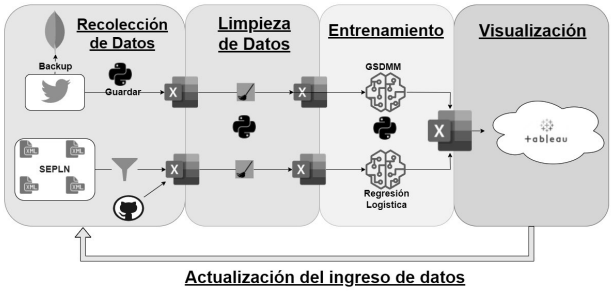

Fig. 1. Ciclo de vida del proyecto

Comprensión del proyecto

En la primera fase, comprensión del proyecto, se busca entender los requerimientos necesarios desde una perspectiva social (empresarial). Esto genera que dichos requerimientos se conviertan en un proyecto de minería de datos. La valoración de la situación fue la siguiente:

- Se previó que los datos para este proyecto son tweets concernientes a las instituciones del Municipio de Quito recolectados de Twitter.
- Los mayores riesgos que envuelven a este proyecto son: el tiempo, la cantidad de información recolectada, datos incomprensibles o con ruido que no permitan una implementación exitosa del proyecto y cuestiones políticas, sociales, económica, etc. y que influyan en el sentimiento de los ciudadanos.

Para los anteriores riesgos mencionados se tienen los siguientes planes de contingencias: CRISP-DM al ser una metodología flexible, en caso de tener problemas en el tiempo de entrega, solo se contemplarían las etapas estrictamente necesarias. Si los datos son incomprensibles se aplican técnicas que permitan realizar un mejor filtrado, limpieza y procesamiento de los datos. De influir factores externos, se pretende realizar una segmentación de los datos en determinado rango de tiempo. Con esto se puede focalizar el análisis en un espacio limitado. Al ser un proyecto de minería de datos se platearon objetivos que conciernen a esta área los cuales se presentan a continuación:

- Recolectar tweets relacionados con instituciones que pertenecen al municipio de Quito mediante el uso del lenguaje de programación Python con técnicas adecuadas como raspado web *(web scraping)*.
- Implementar el filtrado, limpieza e integración de los datos recolectados y con datos de repositorios públicos.
- Construir un modelo que extraiga tópicos derivados de los tweets recolectados mediante técnicas de aprendizaje de máquina (*machine learning)*.

Comprensión de los datos

En la segunda fase, comprensión de los datos, se recolectan datos e identifican problemas en estos. Se busca descubrir señales de posibles conjuntos de datos que ayuden a plantear estrategias o respuestas a los objetivos planteados. En primer lugar, se recolectaron dos conjuntos de datos, el primero corresponde a los tweets de las cuentas de personas (ciudadanía) e interviene en el proceso de extracción de tópicos. El segundo se generó a partir de la unificación de dieciocho subconjuntos de datos e interviene en el análisis de sentimientos. Los métodos usados para la recolección del primer conjunto de datos fueron Tweepy, una librería de Python. Esta librería realiza raspado web de una cuenta de twitter.

Por otra parte, los dieciocho subconjuntos de datos unificados para conformar el segundo conjunto de datos se extrajeron de bases de datos públicas como la Sociedad Española del Procesado del Lenguaje Natural (SEPLN) [18], y un repositorio en Github de acceso público brindado por el usuario charlesmalafosse. Después de obtener los datos, se realizó un análisis para explorar los datos, se encontró que en el primer conjunto de datos las cuentas más mencionadas por la ciudadanía son @LoroHomero y @MunicipioQuito. En el segundo conjunto de datos la polaridad sentimental de los registros tuvo diferentes valores como 'NONE', 'N', 'NEU', 'P', 'positive', 'POSITIVE', 'NEGATIVE', 'NEUTRAL'. Estos

valores representan únicamente tres sentimientos: positivo, neutral y negativo.

Preparación de los datos

En esta etapa se generan los datos tal cual como van a ser usados en la fase de modelamiento. Las tareas dentro de esta fase pueden ser implementadas en un orden diferente al propuesto [17]. El primer paso que se realizó fue la selección de los datos necesarios para el proyecto, se generó a partir de la integración de diferentes conjuntos de datos. Precisamente, un paso de la metodología es la integración de los datos, este paso entregó lo siguiente:

- ***Id_tweet*:** Es el identificador único del tweet, usado para eliminar tweets repetidos.
- ***Text*:** Texto utilizado para realizar el análisis posterior.
- ***Sentiment*:** Es la polaridad del sentimiento en el tweet.

El siguiente paso dentro de la preparación de datos fue la limpieza de datos, en este apartado se realizaron las siguientes actividades en los diferentes archivos creados en pasos anteriores: eliminar filas repetidas, eliminar valores nulos, eliminar retweets, convertir a minúsculas el texto, eliminar tweets cuentas propias, convertir sentimientos a únicamente positivo, neutral y negativo. El siguiente paso fue construir los datos, en este punto se realizaron las siguientes actividades: remover urls, usernames & hashtags, remover caracteres especiales, comprobar tweet en español, tokenizar tweets, eliminar tokens no alfabéticos, lematizar, crear bigrams, juntar bigrams, eliminar registros con menos de tres tokens, balancear datos, ordenar datos aleatoriamente. Después de estos procedimientos el archivo data_collected_from_twitter.csv usado para la extracción de tópicos tuvo 72554 registros y seis columnas. En el archivo sentiment_data.csv usado para el análisis de sentimientos contiene 30000 registros y ocho columnas.

Modelamiento

En la etapa de modelamiento se selecciona el modelo correcto para cumplir con los objetivos establecidos. Primero, se seleccionaron las técnicas de modelamiento. Las dos primeras para la extracción de tópicos, Latent Dirichlet Allocation (LDA) y Gibbs sampling algorithm for the Dirichlet Multinomial Mixture model (GSDMM). En el análisis de sentimientos se optó por la Regresión Logística. El siguiente paso fue generar el diseño de pruebas, la prueba utilizada para la extracción de tópicos fue la coherencia de los tópicos. Los tópicos al ser derivados de la interpretación del autor son susceptibles a resultados variantes ya que están sujetos a la opinión subjetiva de cada persona [19].

Por ende, también se consideró la facilidad de inferencia de los tópicos. Por otra parte, en el análisis de sentimientos al ser un problema de clasificación con aprendizaje supervisado, se separó en conjuntos de datos de entrenamiento y prueba en una relación 90% y 10% respectivamente. Como último paso dentro de la etapa de modelamiento, se construyeron los modelos. En la extracción de tópicos se establecieron dos condiciones tanto para LDA y GSDMM en sus dos opciones (unigram y bigrams), la cantidad de tópicos (siete) y la cantidad de iteraciones (diez). Tanto LDA como GSDMM toman como entrada un corpus, un diccionario y el texto que se utilizó para obtener la medida de coherencia que se consideró en la etapa de evaluación de resultados. Este proceso se hizo dos veces, unigram y bigrams. En cambio, el segundo abordaje que se consideró para la extracción de tópicos fue GSDMM que, junto a los tópicos generados por el modelo, se obtuvo la coherencia de este. Al igual que con LDA este proceso se hizo dos veces (unigram y bigrams). Por otra parte, la construcción del modelo para el análisis de sentimientos comenzó con la lectura de los 30000 registros, estos se dividen un conjunto de datos para entrenamiento y otro para pruebas, esta división fue aleatoria. Posteriormente, se obtuvo una matriz de características Frecuencia de término – Frecuencia de documento inversa TF-IDF (Term Frequency – Inverse Document Frequency), esta matriz permite 'pesar' una palabra según su importancia en un documento (tweet) dentro de un corpus. El siguiente paso fue crear el modelo de Regresión Logística que es usado para crear y entrenar el clasificador. Para determinar si el clasificador generado es bueno se evaluó con el área bajo la curva ROC (AUC ROC).

Evaluación

La etapa de evaluación busca asegurar que el modelo generado ha seguido todos los pasos necesarios para cumplir con su objetivo. Se construyen diferentes modelos, se selecciona el que tiene un mejor rendimiento. El primer paso fue obtener los resultados de la extracción de tópicos. Como se explicó anteriormente los criterios de evaluación son la coherencia del tópico y la facilidad para deducir los tópicos (criterio del autor). El modelo ganador de LDA fue la opción entrenada con bigrams. En cambio, el modelo ganador de GSDMM fue el modelo unigram, se realizó una comparación entre los modelos ganadores de LDA y GSDMM.

TABLA I. PALABRAS SIGNIFICATIVAS POR CLÚSTER DEL MODELO GSDMM UNIGRAM

N°	Palabras significativas	Tópico
0	'ciudad', 'persona', 'sector', 'espacio', 'trabajo', 'apoyo', 'proyecto', 'seguridad', 'atención', 'centro'	Seguridad en la ciudad
1	'bus', 'transporte', 'gente', 'persona', 'control', 'contagio', 'medida', 'unidad', 'restricción', 'servicio'	Falta de control sanitario
2	'espacio', 'agenda', 'museo', 'actividad', 'parte', 'cultura', 'evento', 'ciudad', 'sábado', 'centro'	Actividades culturales
3	'alcalde', 'ciudad', 'obra', 'gente', 'municipio', 'concejal', 'metro', 'corrupción', 'trabajo', 'prueba'	Corrupción en el Municipio
4	'ruido', 'gas', 'problema', 'camión', 'pueblo', 'app', 'bar', 'decibel', 'volumen', 'música'	Contaminación auditiva
5	'agua', 'servicio', 'respuesta', 'mes', 'trámite', 'gracia', 'atención', 'número', 'información', 'turno'	Ayuda en servicios municipales
6	'calle', 'sector', 'vía', 'parque', 'ayuda', 'ciudad', 'basura', 'persona', 'barrio', 'agua'	Atención en necesidades de los barrios

En este caso, una persona hispanohablante puede inferir los tópicos generados en los resultados del modelo GSDMM con mayor facilidad que en el modelo LDA. Es así como GSDMM presentó mayor facilidad de inferencia de los tópicos generados por su modelo. En la tabla I se presentan las palabras más significativas agrupadas y el tópico que representa todas las palabras con GSDMM. Para el análisis de sentimientos, la medida que se eligió para determinar que el modelo fue efectivo, es tener la medida área bajo la curva (AUC) mayor a

0.75. Después del entrenamiento, el modelo obtuvo un AUC ROC de 0.892476 lo cual supera el objetivo planteado como se presenta en la figura 2.

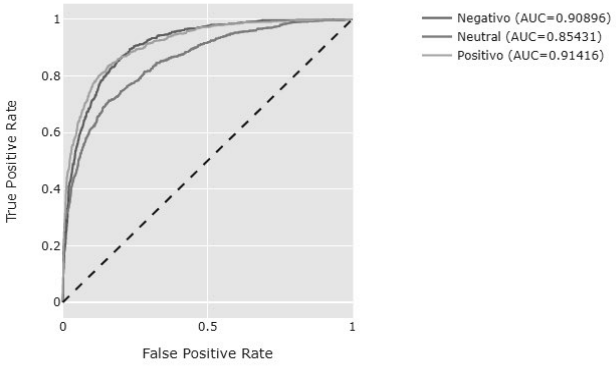

Fig. 2. Gráfica AUC ROC del modelo de regresión logística generado

Desarrollo del sistema de visualización

La arquitectura del sistema es la siguiente:

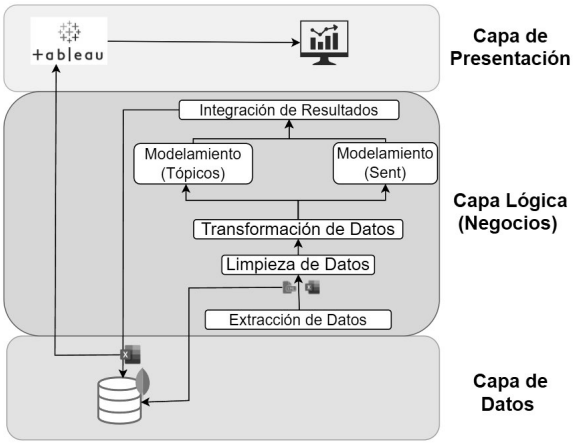

Fig. 3. – Arquitectura del sistema

Los componentes principales dentro de cada capa son:

Capa de Datos: En esta capa se encuentran los datos respaldados como copia de seguridad y los archivos obtenidos después de la recolección de datos, estos archivos inicialmente en formato .csv.

Capa Lógica: Dentro de esta capa constan los componentes extracción de datos, en el cual los datos son recolectados de Twitter, la limpieza de datos, la transformación de datos para la siguiente etapa. En el componente de modelamiento se realiza la generación de los modelos para la extracción de tópicos y el análisis sentimental. Finalmente, la integración de los resultados y la creación de los datos finales para la capa de presentación.

Capa de Presentación: En esta capa se encuentra la visualización generada dentro de Tableau Public, el tablero de mandos (*dashboard*) toma como entrada los datos generados en la capa lógica y entrega como resultado las gráficas agrupadas dentro de una sola vista.

RESULTADOS Y DISCUSIÓN

TABLA II. PALABRAS SIGNIFICATIVAS POR CLÚSTER DEL MODELO GSDMM UNIGRAM.

Tópico
Seguridad en la ciudad
Falta de control sanitario
Actividades culturales
Corrupción en el Municipio
Contaminación auditiva
Ayuda en servicios municipales
Atención en necesidades de los barrios

Los tópicos obtenidos del modelo *unigram* GSDMM se presentan en la tabla II. A partir de los resultados presentados en la tabla II se obtuvieron subtópicos del tópico 'Atención en necesidades de los barrios' los cuales se presentan en la tabla III junto con la cantidad de tweets catalogados como positivos, negativos y neutrales. La distribución de sentimientos por tópico se presenta en la figura 4, se presentan los 5 tópicos más relevantes.

TABLA III. DISTRIBUCIÓN DE SENTIMIENTOS POR TÓPICO CON MAYOR GRANULARIDAD.

Tópico	**Positivo**	**Neutral**	**Negativo**	**Total**
Falta de control sanitario	1079	3565	7136	11780
Actividades culturales	816	3171	509	4496
Construcción del metro	528	1588	2632	4748
Corrupción del hijo del alcalde	1019	1897	5414	8330
Remoción del alcalde	374	1541	2064	3979
Contaminación auditiva	397	1660	1646	3703
Ayuda en servicios municipales	1061	2628	4457	8146
Mal estado de las calles	492	1749	2312	4553
Venta informal	533	1352	2671	4556
Problemas de basura en parques y agua	550	1435	2824	4809
Seguridad en la ciudad	3282	3461	6711	13454
Total sentimiento	**10131**	**24047**	**38376**	**72554**

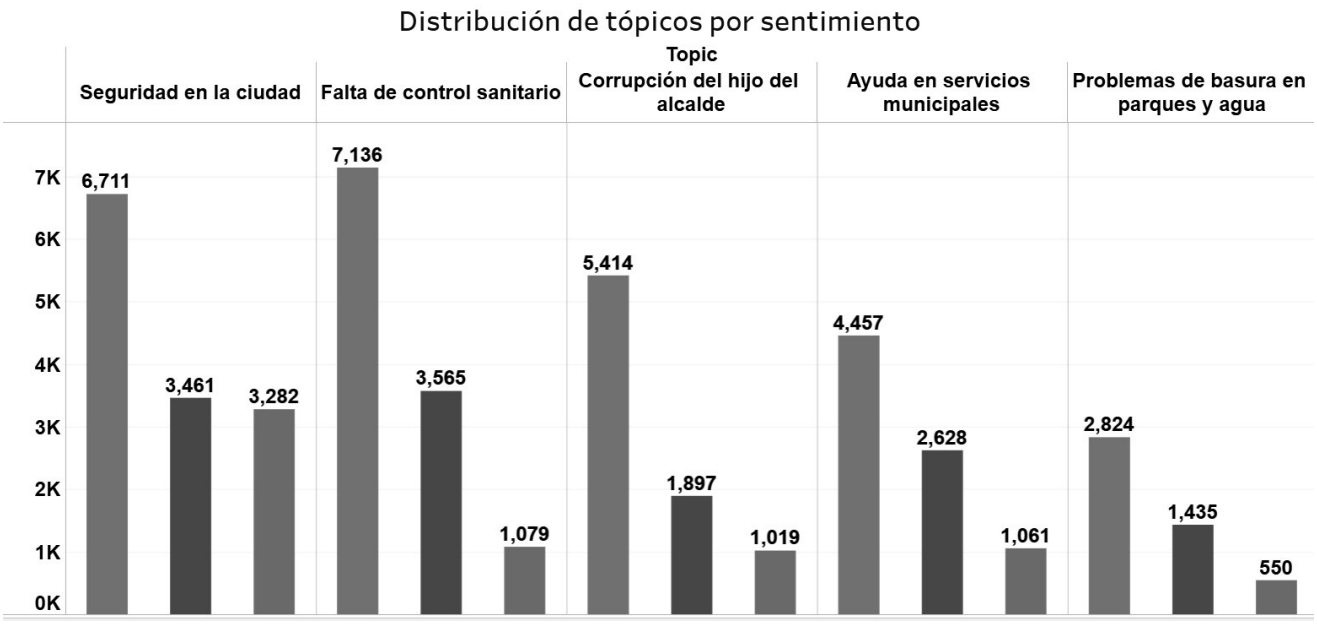

Fig. 4. Distribución de sentimientos por tópico

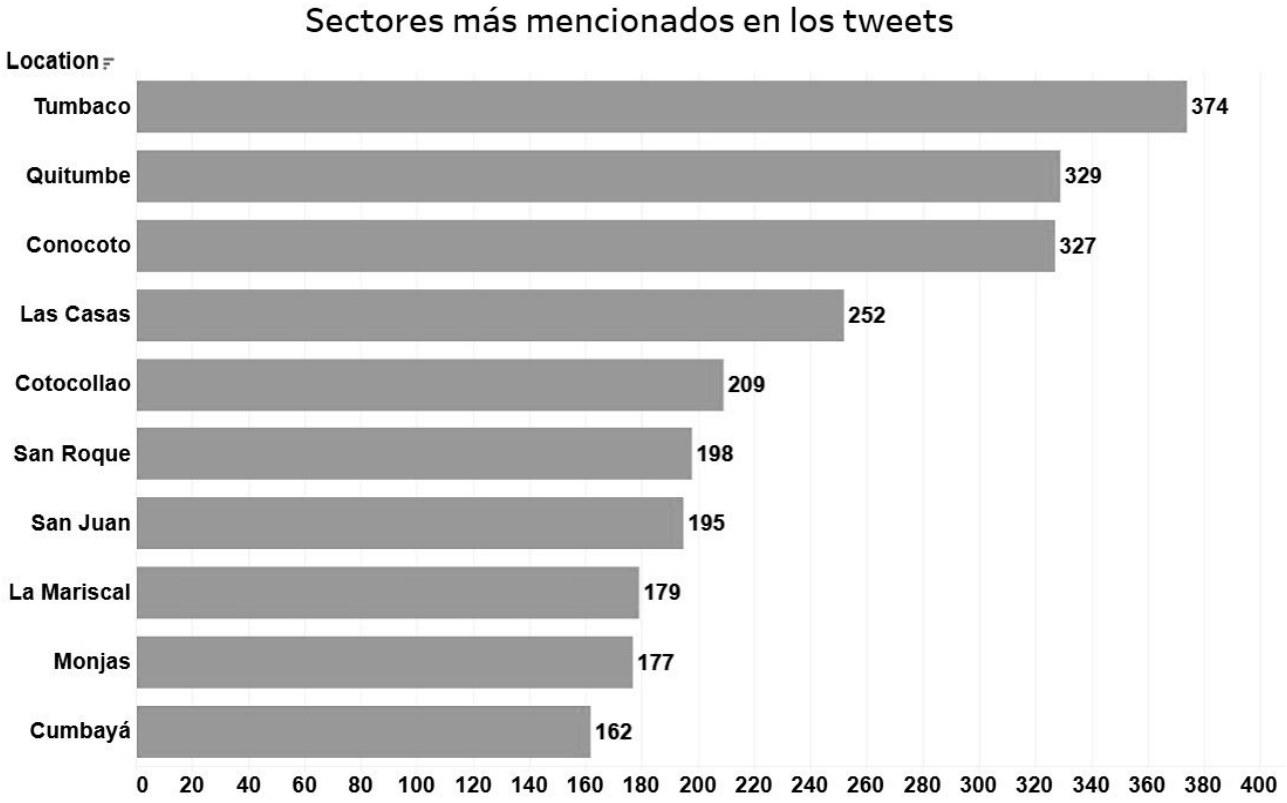

Fig. 5. Sectores más mencionados en los tweets

En la figura 5 se presenta los 25 sectores menciones en los tweets. Los sectores que más destacan son Tumbaco, Quitumbe, Conocoto, Las Casas, Cotocollao y San Roque. Después de identificar los sectores más mencionados en general, se clasificaron los sectores distribuidos por tópicos. Es decir, se obtuvo los tres sectores más mencionados en cada uno de los once tópicos encontrados anteriormente, esto se presenta en la figura 6.

Fig. 6. Distribución de tópicos por sectores más mencionados

Finalmente, el tablero de mandos creado (dashboard) resumido se presenta en la figura 7 al final del documento. Con base a los resultados presentados y al considerar que los tópicos que presentan una mayor opinión negativa sobre positiva y neutral son considerados como necesidades, se infiere que las principales necesidades de la ciudadanía quiteña son:

- Se necesita detener y clarificar los hechos de corrupción que se conocen en el Municipio de Quito. Los puntos principales involucrados en este tema son la remoción del alcalde, la construcción del metro y los problemas que involucran al hijo del alcalde en temas de corrupción.
- Los barrios y sector populares de Quito necesitan atención en: el mal estado de las calles, la venta informal, los problemas de basura en parques y servicios de agua y luz.
- Se necesita mayor control en temas de seguridad de la ciudadanía, aunque no es una competencia directa del Municipio, se necesitan acciones para brindar un ambiente más confiable y seguro.
- La contaminación auditiva necesita tener un control, principalmente, en la entrega del gas doméstico, algunos proveedores hacen uso de música a alto volumen lo cual genera molestias en la ciudadanía.
- Los sectores que más atención necesitan en materia de Seguridad son Quitumbe, San Roque y Monjas, por otra parte, en los sectores Tumbaco, Las Casas y Cumbayá
- requieren solución a problemas relacionados con basura en los parques y servicio de agua potable.

Cabe mencionar que los datos recolectados están en un rango de fechas desde el 14 de mayo de 2019 hasta el 1 de julio de 2021.

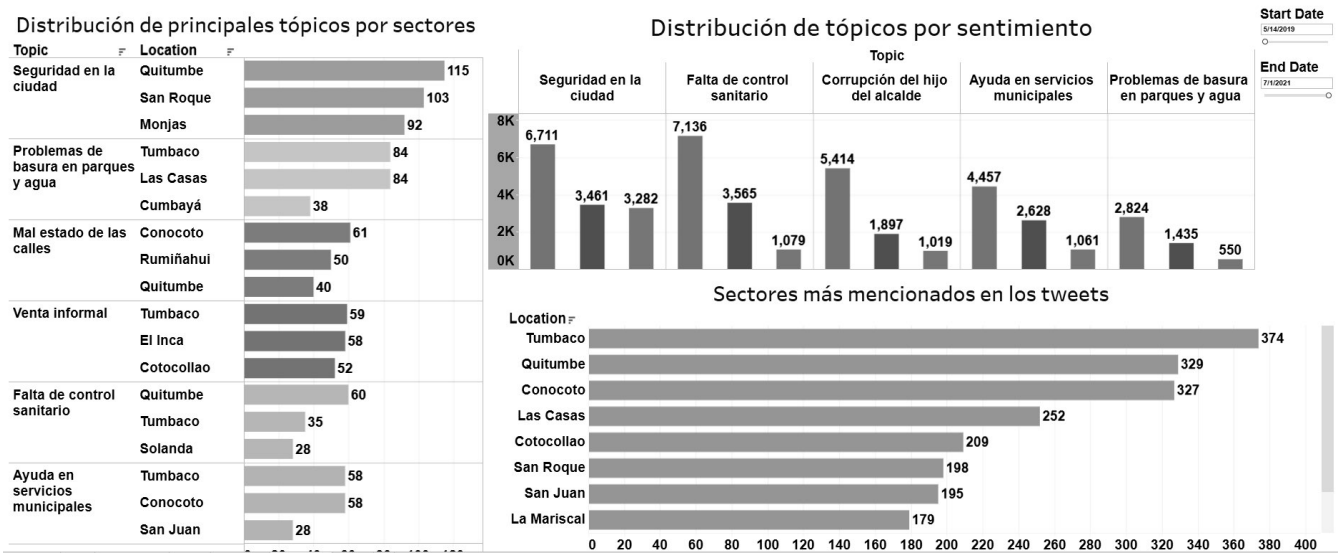

Figura 7. Tablero de mandos creado (Dashboard)

CONCLUSIONES

Se consultó en la literatura trabajos relacionados tanto con la extracción de tópicos como con el análisis sentimental en los cuales se detallaron trabajos por separado y no se constató trabajos que ensamblen estos dos procesos dentro de un mismo resultado aplicado al entorno ecuatoriano. El análisis y clasificación de tweets filtraron el ruido que se pudo generar a partir de datos basura o repetidos. Como punto a considerar para trabajos futuros es el análisis de emoticones ya que pueden brindar otra alternativa para realizar el análisis sentimental. Igualmente, otro aspecto a considerar es la ubicación geográfica de los tweets, ya que el análisis de ubicación de este proyecto considera sectores mencionados en tweets, pero no el lugar desde donde los tweets fueron generados. Esto puede influir en la distribución de cada tópico a un nivel granular de sectores, barrios o zonas previamente delimitadas anteriormente.

En los resultados obtenidos, bajo el análisis de este proyecto, los tópicos negativos de mayor influencia son: falta de sanitario, seguridad en la ciudad, corrupción del hijo del alcalde, ayuda en servicios municipales y problemas de basura en parques y agua son vistos como necesidades a ser resueltas por el ente municipal. Los sectores de Quito que más necesitan atención son Tumbaco, Quitumbe, Conocoto, Las Casas, Cotocollao y San Roque ya que fueron los sectores más mencionados y con mayor polaridad negativa. En cuanto a la parte técnica GSDMM se desenvuelve mejor en textos cortos como se detalla en la literatura, sin embargo, un aspecto a considerar fuertemente es el tiempo, LDA es más rápido que GSDMM.

REFERENCIAS

[1] A. Saunders, "La era de la Perplejidad: Repensar el mundo que conocíamos", Mar. 31, 2022. [En línea]. Disponible: https://www.bbvaopenmind.com/wp-content/uploads/2018/01/BBVA-OpenMind-La-era-de-la-perplejidad-repensar-el-mundo-que- conociamos.pdf

[2] A. Saunders, "La era de la Perplejidad: Repensar el mundo que conocíamos", Mar. 31, 2022. [En línea]. Disponible: https://www.bbvaopenmind.com/wp-content/uploads/2018/01/BBVA-OpenMind-La-era-de-la-perplejidad-repensar-el-mundo-que-conociamos.pdf

[3] J. Grus, Data Science from scratch: First principles with python, Sebastopol, CA, USA: O'Reilly Media, 2015.

[4] K. Umachandran, I. Jurčić, V. Corte y D. Ferdinand-James, "Industry 4.0.: The new industrial revolution" en Big Data Analytics for Smart and Connected Cities, 2017, pp. 138-156

[5] S. Alotaibi, R. Mehmood and I. Katib, "Sentiment Analysis of Arabic Tweets in Smart Cities: A Review of Saudi Dialect," 2019 Fourth International Conference on Fog and Mobile Edge Computing (FMEC), 2019, pp. 330-335, doi: 10.1109/FMEC.2019.8795331

[6] J.B. Hollander y H. Renski, Measuring Urban Attitudes Using Twitter: An Exploratory Study, Lincoln Institute of Land Policy, Medford, MA, USA, 2015

[7] E. H. Alkhammash, J. Jussila, M. D. Lytras and A. Visvizi, "Annotation of Smart Cities Twitter Micro-Contents for Enhanced Citizen's Engagement," in IEEE Access, vol. 7, pp.116267-116276, 2019, doi: 10.1109/ACCESS.2019.2935186.

[8] D. Murthy, A. Gross, A. Pensavalle, "Urban Social Media Demographics: An Exploration of Twitter Use in Major American Cities", Journal of Computer-Mediated Communication, vol. 21, no. 1, pp. 33-49, Ene. 2016, doi: 10.1111/jcc4.12144

[9] S. M. Zavattaro, P. E. French y S. D. Mohanty, "A sentiment analysis of U.S. local government tweets: The connection between tone and citizen involvement", Government Information Quarterly, vol. 32, no. 3, pp. 333-341, doi: 10.1016/j.giq.2015.03.003

[10] W. Pedrycz and S.-M. Chen, Sentiment Analysis and Ontology

[11] Engineering An Environment of Computational Intelligence, 1st ed. 2016. Cham: Springer International Publishing : Imprint: Springer, 2016.

[12] Z. Wang et al., TwiInsight: Discovering Topics and Sentiments from Social Media Datasets, 2017.

[13] S. Kiritchenko, X. Zhu y S. M. Mohammad, "Sentiment Analysis of Short Informat Texts", Journal of Artificial Intelligence Research, vol.50, pp. 723-762. Ago. 2014. doi: 10.1613/jair.4272.

[14] A. Subasi, Practical machine learning for data analysis using Python. Cambridge, MA: Elsevier Academic Press, 2020.

[15] D. Hand and R. Till, "A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems", Machine Learning, vol. 45, pp.171-186, 2004.

[16] J. Mazarura and A. de Waal, "A comparison of the performance of latent Dirichlet allocation and the Dirichlet multinomial mixture model on short text," 2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech), 2016, pp. 1-6, doi: 10.1109/RoboMech.2016.7813155

[17] J. Yin y J. Wang, "A dirichlet multinomial mixture model-based approach for short text clustering", Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014.

[18] F. Rosner, A. Hinneburg, M. Roder, M. Nettling, y A. Both, "Evaluating topic coherence measures", ArXiv, 2014.

[19] IBM SPSS Modeler CRISP-DM Guide, IBM Corporation.

[20] Sociedad Española para el Procesamiento del Lenguaje Natural, "TASS: Workshop on Semantic Analysis at SEPLN", http://tass.sepln.or/tass_data/download.php?auth= Ft0aSs2sV4peVv2eor9.

[21] S. Hansen, M. McMahon, y A. Prat, "Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach", The Quarterly Journal of Economics, vol. 133, no. 2, pp. 801-870, May. 2018. doi: 10.1093/qje/qjx0