ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA DE SISTEMAS

Desarrollo de un sistema de detección de necesidades de atención municipal en base al análisis de tweets

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN SISTEMAS INFORMÁTICOS Y DE COMPUTACIÓN

Edison Ubaldo Sanango Simbaña

edison14_@hotmail.com

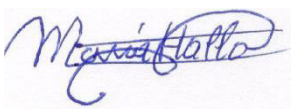
DIRECTORA: Hallo María PhD.

maria.hallo@epn.edu.ec

Quito, 23 de febrero de 2022

AVAL

Certifico que el presente trabajo fue desarrollado por el estudiante Edison Ubaldo Sanango Simbaña, bajo mi supervisión.



_____

MARÍA HALLO PhD.

DIRECTORA DEL TRABAJO DE TITULACIÓN

DECLARACIÓN DE AUDITORÍA

Yo EDISON UBALDO SANANGO SIMBAÑA, declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentada para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración cedo mis derechos propiedad intelectual correspondientes a este trabajo, a la Escuela Politécnica Nacional, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.

_____

EDISON UBALDO SANANGO SIMBAÑA

AGRADECIMIENTO

A Mercedes y Ubaldo, por ser los primeros maestros y amigos que he tenido en mi vida. A ustedes les debo todo.

A Dora, que, sin importar la distancia, con su apoyo he logrado ser mejor persona.

A todos, por lo bueno y lo malo, me ha servido para crecer como profesional y persona.

ÍNDICE DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

RESUMEN

Los tweets ganaron espacio los últimos años debido al uso de las redes sociales, en ellos se puede analizar diferentes aspectos como sentimientos y extraer tópicos de ellos. En este proyecto se detalla la construcción de un sistema de detección de necesidades concernientes al Municipio de Quito en base al análisis de tweets.

Los temas que aborda este proyecto son la extracción de tópicos para descubrir los temas más importantes que se hablan en Quito y según su polaridad sentimental clasificarlos como necesidades, y el análisis de sentimientos para orientar la polaridad (positiva, negativa o neutral) de la ciudadanía en cada tópico encontrado junto con la ubicación de sectores relevantes y más mencionados por la ciudadanía. En la extracción de tópicos se consideraron técnicas de limpieza y procesamiento de datos como eliminar signos de puntuación, convertir a minúsculas, eliminar las palabras no útiles (*stopwords*) del idioma español, entre otras. Asimismo, el algoritmo utilizado fue Algoritmo de muestreo de Gibbs para el modelo de mezcla multinomial de Dirichlet GSDMM (*Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture model)* que tuvo un mejor desempeño que el algoritmo Asignación latente de Dirichlet LDA (*Latent Dirichlet Allocation)*.

Después de aplicar GSDMM se generaron siete tópicos de los cuales se analizó la polaridad sentimental. En la parte del análisis de sentimientos se entrenó un modelo de Regresión Logística para obtener la polaridad (positivo, negativo y neutro) de cada tópico. Además, se generaron subtópicos de algunos tópicos debido a su poca granularidad de detalle.

Palabras clave: Extracción de tópicos, análisis de sentimientos, CRIPS-DM, GSDMM, tweets.

ABSTRACT

Tweets gained fame in recent years due to the use of social networks, they allow analyze different aspects such as feelings and extract topics from them. This project details the development of a needs detection system regarding the Municipality of Quito based on the analysis of tweets.

The topics that are involved in this project are the extraction of topics to discover the most important aspects which are spoken in Quito and according to their sentimental polarity classify them as needs and the analysis of feelings to classify the polarity (positive, negative or neutral) of citizenship attitude in each topic found along with the location of important and most mentioned places by citizens. In the extraction of topics, cleaning and data processing techniques were considered, such as eliminating punctuation marks, converting to lowercase, removing stopwords from the Spanish language, among others. Likewise, the algorithm used was Gibs Sampling algorithm for the Dirichlet Multinomial Mixture model (GSDMM), which performed better than the Latent Dirichlet Allocation (LDA) algorithm.

After applying GSDMM, seven topics were obtained, of which the sentimental polarity was analyzed. In the sentiment analysis section, a Logistic Regression model was trained to obtain the polarity (positive, negative and neutral) of each topic. In addition, subtopics of some topics were generated due to their low granularity of detail.

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

1. INTRODUCCIÓN

1.1. Descripción del problema

Con la fuerte globalización y el crecimiento de las tecnologías, todos los sectores que conforman a la sociedad deben evolucionar constantemente (Saunders, 2017). 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. (Grus, 2015). Los datos comprenden un proceso desde su recolección hasta la toma de decisiones apoyadas en la información que estos generan (Cano, 2007). Igualmente, existen herramientas determinadas que están destinadas a realizar todo este proceso. 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 (Umachandran, Jurčić, Corte, & Ferdinand-James, 2019).

Por otra parte, es importante considerar el entorno en que la institución se encuentra. No todos los sectores tienen las mismas necesidades, por lo cual no hay una regla general para las instituciones públicas, privadas, ONG, etc. en el uso de determinada tecnología (ITahora, Resultados Tendencias y Prioridades IT 2021, 2021). Más concretamente, en Ecuador no existe una cultura fuertemente marcada que ponga su enfoque en el análisis de datos y la toma de decisiones en base a esto. Aunque en 2020, el Analítica y Datos Masivos (*Big Data & Analytics*) fue la tecnología emergente primordial en Ecuador (ITahora, Tendencias Tecnológicas de mayor impacto en Ecuador para el 2020, 2020).

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. Esto no hace referencia al almacenamiento de datos, acción que la mayoría de las instituciones realizan sin importar el tamaño de esta. En cambio, es el obtener información valiosa a partir del aprovechamiento de los datos en bruto generados por la institución directa o indirectamente, por ejemplo, tweets y su posterior tratado para el apoyo en la toma de decisiones que velan por la mejora de la institución. Hasta el momento, no se ha reflejado la implementación de una estrategia que tome ventaja de los datos que se generan para mejorar los servicios de esta institución.

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. Así mismo, al no tener este sistema no puede apoyarse en sus datos para una mejor toma de decisiones que, indudablemente, mejorarán sus servicios y operaciones.

1.2. Propuesta

Este proyecto 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 con acciones claves.

Diferentes aspectos son considerados dentro del desarrollo de este proyecto, los principales son el análisis de sentimientos y la extracción de tópicos. El análisis de sentimientos o generalmente conocido como minería de opiniones (*opinion mining*) (Luo, Chen, Xu, & Zhou, Trust-based Collective View Prediction, 2013) ha sido un campo extensamente abordado, aunque su uso junto a la clasificación de los sentimientos no es muy común. El análisis de sentimientos involucra determinar la posición evaluativa que un determinado texto posee; puede ser positivo, negativo o neutro (Pedrycz & Chen, 2016). Igualmente, se define como la capacidad de identificar y clasificar las opiniones subjetivas en diferentes documentos o textos (Luo, Chen, Xu, & Zhou, Trust-based Collective View Prediction, 2013). Esto tiene la capacidad de abarcar cualquier tipo de documento, ya sea, grandes textos o pequeños como SMS o tweets. Esto obedece a una necesidad imperiosa por conocer el pensar y sentimiento de las personas interesadas. Habitualmente, el interés es generado por una necesidad o deseo de mejora, por ende, se refleja en los sentimientos que, a su vez, en la época actual, se reflejan por redes sociales entre ellas Twitter (Pedrycz & Chen, 2016).

Por otra parte, la extracción de tópicos también conocida como detección de tópicos, extracción de palabras claves o extracción de frases clave es el acto de diferenciar los términos de un documento basados en su importancia en dicho documento o aquel que describa mejor la idea general (Wang, Bai, Chowdhury, Xu, & Seow, 2018). Se basa en las repeticiones dentro de un documento y luego las palabras top (más repetidas o importantes) son seleccionadas para definir el tópico (Liu, Huang, Zheng, & Sun, 2010).

La metodología que este proyecto sigue es CRISP-DM. Proceso Estándar de la Industria Cruzada para la Minería de Datos CRISP-DM *(Cross Industry Standard Process for Data Mining*) es una metodología que establece procedimientos específicos y moldeables para trabajos de minería de datos, ejecutando así procesos de forma sistemática sin actividades vanas o inservibles (Moine, Haedo, & Gordillo, 2010). Se caracteriza por tener un nivel de abstracción que abarca lo macro hasta lo micro (IBM, 2011). Al ser dividido, en primera instancia, en distintas fases, para cada una de estas fases se determina tanto las tareas como los entregables (Marbán, Mariscal, & Segovia, 2009). Las fases pueden ser implementadas de forma no secuencial, es decir, su implementación no es rígida.

Es importante en cualquier campo que esta metodología sea aplicada, que se estudie o mapee su contexto para realizar solo lo necesario. Esto puede ser tanto un proyecto presente o un proyecto a futuro. La diferencia entre estos dos mapeos es que el presente es usado una sola vez. En cambio, el mapeo a futuro considera las experiencias de distintos proyectos anteriores. Las seis fases que constan dentro de esta metodología son: Comprensión del negocio, comprensión de los datos, preparación de los datos, modelamiento, evaluación y despliegue (IBM, 2011). La fase de despliegue no fue considerada en este proyecto debido a la naturaleza de este, esto se debe a que la metodología permite moldear los pasos según la necesidad de cada proyecto.

Todas las fases mencionadas proporcionan entregables o salidas que, en muchos casos, se convierten en entradas para las tareas de las siguientes fases. Ante esto, las principales salidas son: los objetivos del negocio, el plan del proyecto, beneficios y costos, informes sobre los datos y su calidad. Igualmente, ya en el proceso de implementación, la configuración de parámetros de los modelos y los modelos son importantes (IBM, 2011).

1.3. Objetivo General

El objetivo general de este proyecto es construir un sistema de detección de necesidades de la ciudadanía relacionado con la gestión municipal de un caso de estudio mediante el análisis de tweets.

1.4. Objetivos Específicos

Dentro de este proyecto se consideran algunos proyectos específicos como son los siguientes:

- Revisar trabajos relacionados.
- Ejecutar un proceso secuencial de recolección, análisis y clasificación de tweets.
- Desarrollar un sistema de detección de necesidades.
- Realizar pruebas con un caso de estudio.

1.5. Trabajos Relacionados

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. Ciudades y países no han sido la excepción y, de igual manera, optan por usar la tecnología a su favor.

Asimismo, dichas entidades junto con sus ciudadanos también son objetos de estudio (Alkhammash, Jussila, Lytras, & Visvizi, 2019). Twitter como red social global es un medidor del impacto sentimental generado en personas debido a determinados eventos (Alotaibi, Mehmood, & Katib, 2019).

Dentro de diferentes sitios web como *Google Scholar*, *Research Gate* y *ScienceDirect* entre otros, se presentan algunos artículos y trabajos que hablan del uso de Twitter y la percepción sentimental sobre diferentes temas que se hablan en los tweets. Estos artículos fueron encontrados mediante la búsqueda de términos específicos dentro de los sitios web mencionados. Los términos principales fueron: análisis de sentimientos, redes sociales, Twitter, extracción de tópicos y ciudades, tanto en español como en inglés. El conocimiento adquirido de estos artículos influyó en la determinación del alcance y la estrategia a seguir dentro del proyecto, así como tener la capacidad de contemplar y solventar posibles contratiempos y bloqueos dentro del proyecto que ya se presentaron en los artículos estudiados.

Haciendo 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 (Hollander & Renski, 2015). 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.

Igualmente, dentro de las 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 (Alkhammash, Jussila, Lytras, & Visvizi, 2019). A esto se suma las soluciones viales que son implementadas en las ciudades inteligentes como sensores en las carreteras y sistemas de transporte inteligente lo cual ayuda a reducir la contaminación con CO_2. Todo esto mediante el uso de análisis de datos masivos en tiempo real, esto brinda seguridad pública y reforzamiento de sistemas aplicantes de leyes (Ahmed, Radenski, Bouhorma, & Ahmed, 2016). 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 (Murthy, Gross, & Pensavalle, 2016).

Entre otras investigaciones también se puede encontrar que el tono (sentimiento) puede influir a los ciudadanos a participar activamente en acciones planteadas por entidades

gubernamentales. Sin embargo, se evidenció que no es el único factor que influye en su participación (Zavattaro, French, & Mohanty, 2015). 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.

Los trabajos descritos anteriormente brindan una perspectiva de sentimiento y su aplicación al entorno estudiado, el desarrollo de este proyecto se centra en el entorno ecuatoriano específicamente en la ciudad de Quito. La manera en que estos trabajos se adaptan a la aplicación de este proyecto es en la manera de clasificar los sentimientos (positivo, neutral, negativo), determinar qué aspectos del texto de un tweet se pueden usar para el análisis de sentimientos y cuáles para la extracción de tópicos. Asimismo, identificar de qué lugares se puede obtener tweets que estén relacionados directamente con el objetivo del proyecto.

1.6. Marco teórico

Es necesario comprender qué temas se aplican en este proyecto. Así mismo, considerar diferentes aproximaciones o abordajes es igualmente necesario. Las herramientas, técnicas y procedimientos usados deben acoplarse dentro de la metodología y los pasos posteriores de esta. Por ende, se decide considerar los siguientes aspectos para realizar una selección más informada posteriormente.

1.6.1. Análisis de Sentimientos

El análisis de sentimientos *(opinion mining*) se relaciona con la evaluación mediante la aplicación del procesamiento de lenguaje natural, entre otras técnicas, del sentimiento percibido en un texto mediante la clasificación de opiniones en distintas polaridades, las más comunes son positivo, negativo y neutral. (Kiritchenko, Zhu, & Mohammad, 2014). Este sentimiento o polaridad es derivado de la opinión o posición del autor respecto a un tema o tópico. Un texto o documento se clasifica en hechos (objetivo) y opiniones (subjetivo) (Luo, Chen, Xu, & Zhou, Sentiment Analysis, 2013). Los hechos son clasificados como neutrales ya que, por lo general, pertenecen a datos que no expresan una polaridad clara. Por otra parte, las opiniones reflejan el sentimiento o actitud del autor.

1.6.1.1. Derivación (*stemming*) y lematización (l*emmatization)*

La derivación (s*temming)* y la lematización *(lemmatization*) son procesos que ayudan a eliminar las posibles variaciones gramaticales de una palabra y convertirlos a su forma base. En idiomas como el español debido a sus conjugaciones, ciertas palabras hacen referencia a lo mismo, pero con una o dos letras extras. A pesar de tener el mismo objetivo, la derivación

elimina los sufijos al final de una palabra. Por otra parte, la lematización hace uso de un 'diccionario' para realizar la misma acción, es decir, determina si dos palabras tienen una misma raíz (Jurafsky & Martin, 2020).

1.6.1.2. Tokenización

La tokenización *(tokenization*) es el proceso de segmentar un texto extenso en palabras o segmentos de palabras. Se hace uso de un tokenizador (un caracter específico), generalmente es el espacio en blanco, sin embargo, muchas veces se necesitan algoritmos más complejos debido a la complejidad de los segmentos deseados (Jurafsky & Martin, 2020). En otras palabras, dado un documento (una unidad de texto) cortarlo en partes llamadas tokens que son una secuencia de caracteres. En muchas ocasiones este paso es inevitable para hacer procesamiento de texto (Anta, Chiroque, Morere, & Santos, 2013).

1.6.1.3. Etiquetado de las partes de la oración (*part of speech*)

Dentro del procesamiento de texto es necesario determinar si una palabra es, entre otros, un verbo o sustantivo, esto es fundamental para conocer posibles palabras continuas y una correcta estructura sintáctica (Jurafsky & Martin, 2020). Este proceso depende, inherentemente, a cada idioma, por ejemplo, no se aplican las mismas reglas del inglés en el árabe (Ludeling & Kyto, 2008) y consiste en asignar una parte de una oración (*part of speech)* a cada palabra de un documento mediante técnicas ya establecidas como el Modelo Oculto de Markov (*Hidden Markov Model*) o Campo Aleatorio Condicional (*Conditional Random Field*), (Jurafsky & Martin, 2020).

1.6.1.4. Léxico

Un factor que interviene en el análisis de sentimientos es el léxico. Hay palabras naturalmente definidas como buenas y malas que generalmente se encuentran definidas en un 'diccionario' (Nadkarni, Ohno-Machado, & Chapman, 2021). Según este diccionario se puede clasificar según sus sentimientos documentos que contengan estas palabras.

1.6.1.5. Regresión Logística

La regresión logística es un algoritmo de aprendizaje supervisado dentro del mundo del aprendizaje de máquina (*machine learning*) usado para problemas de clasificación (Subasi, 2020). Este modelo necesita que los valores objetivos sean categóricos y no continuos. Es decir, generalmente existen dos posibles valores como respuesta, si/no, verdad/falso, etc. (Belyadi & Haghighat, 2021). Por otra parte, la regresión logística múltiple es usada en la

clasificación de problemas de tres o más posibles respuestas, por ejemplo, determinar entre cinco posibles respuestas cual es la correcta (Edgar & Manz, 2017). La regresión logística básicamente usa una función logística cuyo resultado es una respuesta binaria (variable dependiente) según las características con las que se entrena el modelo (variables independientes) (Belyadi & Haghighat, 2021). La fórmula de la regresión logística se presenta a continuación.

$$p = \frac{1}{1 + e^{-(a+bX)}}$$

1.6.1.6. El área bajo la curva ROC (ROC AUC)

ROC es un gráfico (curva) que representa la proporción de verdaderos positivos frente a los falsos positivos. El área bajo la curva ROC es una medida que resume la información contenida en la curva (McClish, 1989). Esta medida es ampliamente usada cuando se evalúa el desempeño de un clasificador (Hand & Till, 2001). El gráfico ROC representa las compensaciones entre beneficios (verdaderos positivos) y costos (falsos positivos). En la figura 1 la línea diagonal representa una probabilidad aleatoria o el 'azar' del modelo evaluado, para que un modelo sea bueno, debe aproximarse a la esquina superior izquierda y estar sobre la diagonal. En cambio, si se encuentran cerca de la esquina inferior derecha y por debajo de la diagonal representa un mal modelo (Fawcett, 2006). El área bajo la curva puede ser calculada con la fórmula del área de pequeños trapezoides formados a partir de la curva. Asimismo, esta área es importante para determinar el resultado de la evaluación de clasificadores. La figura 1 muestra una comparación entre dos curvas que representan dos clasificadores, el área roja es más grande que el área azul, esto significa que la curva roja creada a partir del clasificador es mejor que el clasificador que creó el área azul.

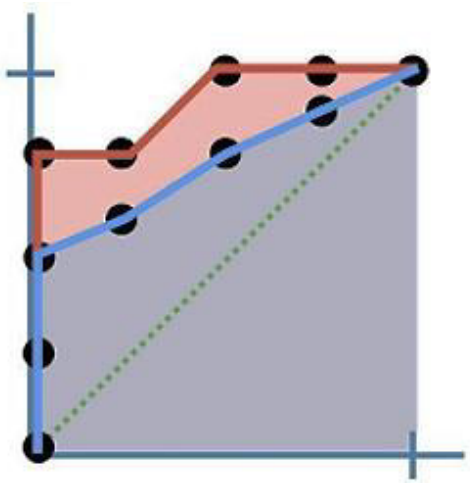


Figura 1. Gráfica AUC ROC

1.6.1.7. Nivel de análisis de sentimientos

El análisis de sentimientos se realiza en textos cuyas características brindan la posibilidad de inferir la opinión o actitud del autor sobre un determinado tema (Kiritchenko, Zhu, & Mohammad, 2014). Sin embargo, el centro neurálgico dentro de este proceso es el texto. Las consideraciones pueden variar dependiendo del texto. Un ejemplo es la longitud del texto, este puede ser grande como un libro con diferentes capítulos, o una sola oración. Este último ha crecido en los últimos años debido a las redes sociales y su impacto en la sociedad (Patil & Yalagi, 2016). El análisis de sentimientos se puede realizar a diferentes niveles como se detalla a continuación.

1.6.1.7.1. Documento

Este nivel clasifica todo un documento respecto a un tema, por lo general, servicios o productos. Todo el contenido del documento es resumido en una sola opinión (Katrekar).

1.6.1.7.2. Oraciones

En este caso existen dos procesos: determinar la subjetividad de la oración y la clasificación del sentimiento. Las oraciones subjetivas son clasificadas entre positivas y negativas. En cambio, las oraciones objetivas, generalmente, brindan hechos y/o datos que no expresan una actitud o posición del autor (Patil & Yalagi, 2016).

1.6.1.7.3. Características / Aspectos

En este proceso se busca extraer características de un texto respecto a un tema y después realizar el análisis sentimental (Katrekar).

1.6.2. Extracción de Tópicos

1.6.2.1. Asignación latente de Dirichlet (*Latent Dirichlet Allocation LDA)*

LDA es un modelo estadístico generativo que permite explicar la similitud de conjuntos de datos mediante grupos no observados. LDA fue presentado por David Blei, Andrew Ng y Michael Jordan en 2003 en su artículo (Blei, Ng, & Jordan, 2003) y es la técnica más popular para la extracción y modelamiento de tópicos. LDA toma una cantidad de tópicos preestablecida, cada uno de estos tópicos representa un conjunto de palabras aún no definidas. El objetivo de LDA es relacionar todos los documentos con los tópicos, de esta manera las palabras de los documentos son tomadas por los tópicos con lo cual se crea un clúster de palabras para cada tópico (Chen & Wang). LDA considera que cada documento

está conformado por una distribución de tópicos y cada tópico está compuesto de una distribución de palabras (Blei, Ng, & Jordan, 2003).

1.6.2.2. Algoritmo de muestreo de Gibbs para el modelo de mezcla multinomial de Dirichlet (*Gibbs sampling algorithm for the Dirichlet Multinomial Mixture model GSDMM)*

GSDMM es una alternativa a LDA para trabajar con textos cortos, debido a las nuevas redes sociales es más común los textos con una longitud relativamente pequeña (Mazarura & Waal, 2016). 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 (Yin & Wang, 2014). Las palabras que contiene un documento se generan de un solo tópico y no de una distribución de tópicos como sucede en LDA.

GSDMM emplea un proceso llamado Proceso de grupo de películas MGP (*Movie Group Process*) que según los autores de GSDMM (Yin & Wang, 2014) MGP "imagina los documentos como estudiantes en un curso de discusión de películas y las palabras de un documento como las películas que el estudiante ha visto. Esto se traduce en un mejor desempeño de GSDMM con respecto a LDA al momento de identificar tópicos en los textos (Yin & Wang, 2014). El problema de la agrupación de textos cortos resulta ser agrupar a los estudiantes en grupos para que los estudiantes del mismo grupo compartan intereses similares (listas de películas similares), mientras que los estudiantes de diferentes grupos compartirán intereses diferentes".

Asimismo, GSDMM emplea los parámetros α (*alfa)* y β (*beta)* que según sus autores representan dos conceptos importantes al momento de entrenar el modelo. Primero, *alfa* está relacionado con la probabilidad inicial de que un documento elija un grupo. Este valor no puede ser cero ya que significa que un documento nunca elegirá un grupo en caso de que el grupo esté vacío. Si alfa aumenta la probabilidad también aumenta. Por otra parte, *beta* se relaciona con la probabilidad de que un documento elija un grupo, aunque no tenga palabras que lo relacionen con los grupos. Este valor no puede ser cero ya que significa que un documento no es asignado a ningún grupo (tópico) (Yin & Wang, 2014).

1.6.2.3. Coherencia de tópicos

Al extraer tópicos no existe una manera específica de probar su efectividad, tradicionalmente, se emplea la medida de la coherencia, esta se entendía como la capacidad de los tópicos de ser "humanamente entendidos" (Mifrah & Benlahmar, 2020). Sin embargo, recientemente la

coherencia se interpreta como la medida que permite distinguir si un tópico es bueno o malo. Igualmente, se define como el promedio de similitudes de palabras en pares formadas por las palabras principales de un tema determinado (Rosner, Hinneburg, Roder, Nettling, & Both) . Esta medida permite categorizar los tópicos de manera que se aproximen más al juicio humano.

1.6.3. SCRUM

El marco de trabajo Scrum indica que el desarrollo de un proyecto se debe llevar a cabo mediante sprints. Un sprint es una unidad de tiempo que se establece para cumplir objetivos planteados en ese sprint, el resultado de un sprint es un incremento progresivo del proyecto y por ende, la continua finalización de las historias de usuario especificadas en el backlog del producto (*product backlog*). Dentro de un sprint existen diferentes eventos que se realizan para ayudar a cumplir los objetivos del Sprint como se visualiza en la figura 5. El primero es la planificación de sprint (*sprint planning*) en donde se determina cuáles actividades se realizarán en el sprint. El scrum diario (*daily scrum*) es un evento diario en el cual se detallan los progresos y bloqueos que se tuvieron. En la etapa final del sprint se realiza la revisión del sprint (*sprint review*) en la cual se inspecciona los resultados acordes al objetivo del sprint. Finalmente, en la retrospectiva del sprint (*sprint retrospective*) se identifican los cambios para mejorar en el siguiente sprint, con este evento se finaliza el sprint (Schwaber & Sutherland, 2020).

2. METODOLOGÍA

Para el desarrollo de este proyecto se consideró tres metodologías que son generalmente usadas durante los proyectos de ciencia de datos (KDnuggets, 2014). Estas son *Cross Industry Standard Process for Data Mining* (CRISP-DM), *Sample, Explore, Modify, Model, Assess* (SEMMA) y *Knowledge Discovery in Databases* (KDD). Para llegar a esta conclusión se analizaron los siguientes factores influyentes: adaptabilidad al proyecto, flexibilidad en sus pasos, dominio de la herramienta, documentación, popularidad. Estos factores fueron evaluados en una escala del 1 al 5 cuyo valor dependió exclusivamente del autor de este proyecto.

Tabla 1. Evaluación de Metodologías para proyectos de ciencia de datos

Factor	CRISP-DM	SEMMA	KDD
Adaptabilidad al proyecto	**4**	**1**	**2**
Flexibilidad en sus pasos	**5**	**3**	**2**
Dominio de la metodología	**2**	**2**	**2**
Documentación	**4**	**2**	**1**
Popularidad	**4**	**1**	**2**
Total	**19**	**9**	**9**

La metodología que obtuvo mejor resultado fue CRISP-DM. En los 5 factores predominantes para esta evaluación se dedujo lo siguiente:

La adaptabilidad al proyecto se evidenció mayormente en la metodología CRISP-DM, esta comprende una mejor aproximación y entendimiento del proyecto. KDD y SEMMA no comprenden esta 'fase' al mismo nivel de CRISP-DM por lo cual son menos recomendadas para este trabajo. En cuanto a la flexibilidad, CRISP-DM es un proceso no rígido y cíclico en sus fases, esto da la facultad de modificar, repetir u omitir ciertas fases dentro de todo el proceso. Por otra parte, no se tiene un dominio completo en ninguna metodología, solamente la suficiente experiencia en ellas para considerarlas como posibles guías para trazar las directrices en este proyecto. CRISP-DM tiene la metodología más completa, aunque desactualizada ya que su última versión fue en 2011 por parte de IBM. Finalmente, CRISP-DM también domina en popularidad, según la encuesta realizada por (KDnuggets, 2014) esta metodología mantiene una amplia ventaja sobre las otras dos metodologías, como se puede ver en la figura 2.

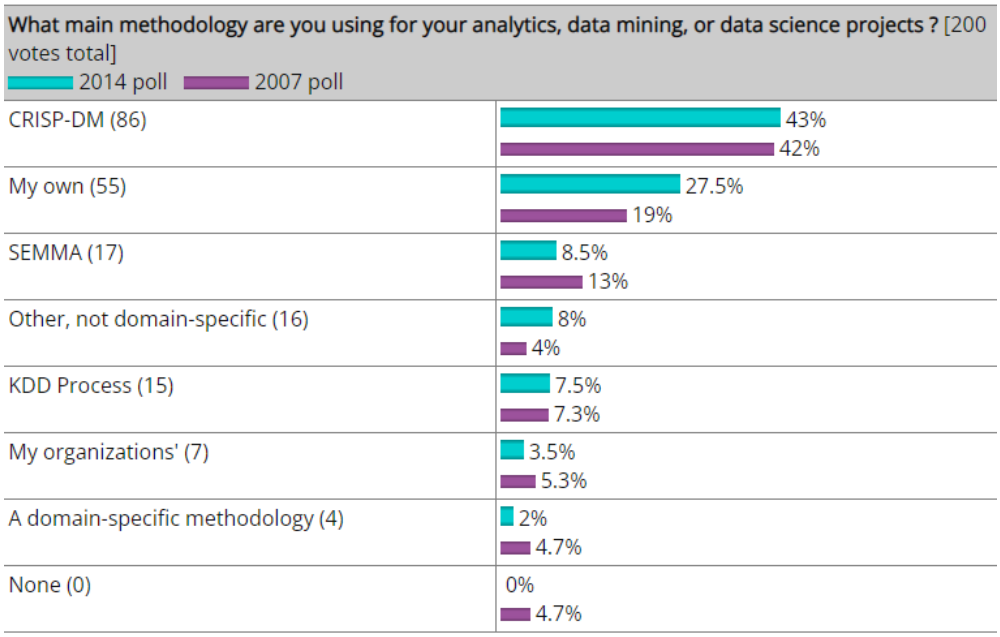


Figura 2. Principales metodologías para proyectos de ciencias de datos

Es importante destacar que se usan metodologías 'creadas' por el mismo desarrollador del proyecto ocupando así la segunda posición en esta encuesta. Es así como CRISP-DM plantea una metodología estructurada, aunque flexible al mismo tiempo (IBM, 2011). 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. El ciclo de vida que comprende este proyecto se describe en la figura 3, 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.

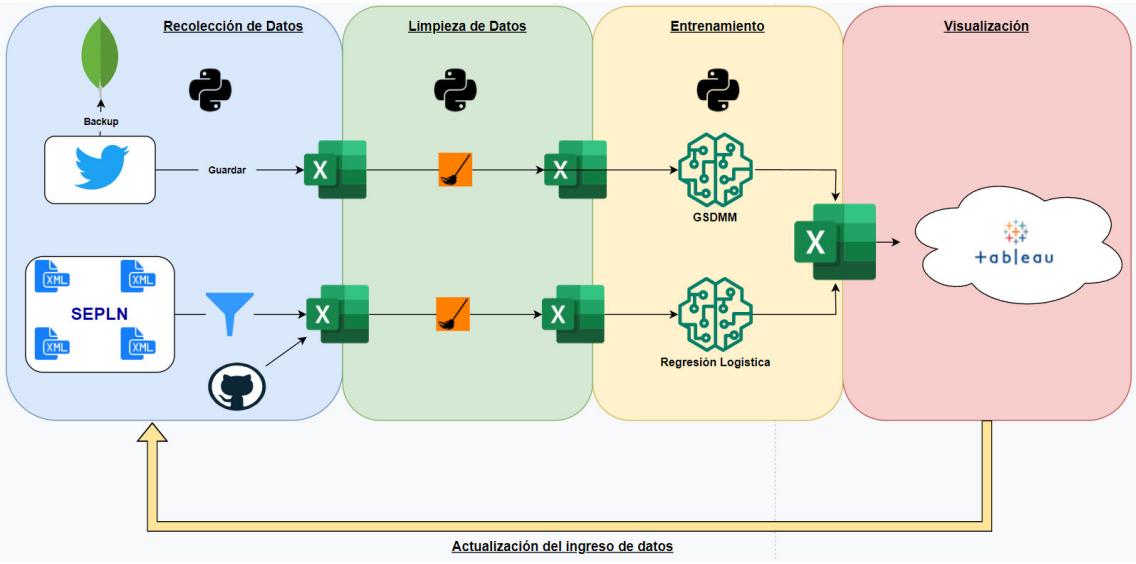


Figura 3. Ciclo de vida del proyecto.

El sistema se compone de cuatro etapas principales: recolección de datos, limpieza de datos, entrenamiento y visualización.

Dentro de la recolección de datos se operó con los datos iniciales, en esta sección se realizó el respaldo de los tweets recolectados como método de prevención contra la pérdida de datos y, por ende, pérdida de tiempo. Esta copia de seguridad se realizó en MongoDB Atlas el cual es el servicio de almacenamiento en la nube que ofrece MongoDB. Por otra parte, los datos obtenidos de la SEPLN son de libre acceso por lo cual no fue necesario realizar una copia de seguridad. Con los archivos obtenidos en la recolección de datos se realizó la limpieza de datos en la cual los datos fueron procesados con lo cual solo los registros necesarios permanecieron en los archivos.

Los registros resultantes en los archivos entraron a la fase de entrenamiento tanto para el análisis sentimental como la extracción de tópicos por separado y la combinación de los resultados de los dos modelos entrenados generaron un archivo único final. La etapa de visualización utilizó el archivo generado por los modelos y mediante el uso de Tableau Public se generó el tablero de mandos (*dashboard*) interactivo. De ser necesario una actualización de datos (nuevos tweets recolectados) el ciclo de vida se repite totalmente, es decir, desde la recolección de datos hasta la visualización.

Al tener una actualización de datos es muy probable que los tópicos generados en el ciclo anterior ya no sean los mismos que se generarían con los nuevos datos, esto se debe a que los tópicos dependen exclusivamente de los tweets y el contenido de cada tweet varía según el tiempo y los eventos que sucedan en la vida cotidiana. Es importante mencionar que la recolección de datos, limpieza de datos y entrenamiento fueron realizados en Python como lenguaje de programación, en cambio, la visualización de los resultados se trabajó en Tableau Public lo cual es un aplicativo web cuyo resultado se publica en un alojamiento (h*osting*) web gratuito y de libre acceso. Para llegar al resultado final presentado en la figura 3 se siguieron los siguientes pasos, estos pasos constan en la metodología CRISP-DM y presentan adaptaciones a las necesidades propias del proyecto.

2.1. 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. En esta fase, hay cuatro tareas: determinar los objetivos del proyecto (negocio), valorar la situación, determinar objetivos de minería de datos y producir el plan del proyecto (IBM, 2011).

2.1.1. Determinar objetivos del proyecto

Antes de la elaboración de este proyecto, no existía una herramienta de conocimiento público que permita deducir las necesidades, intereses o temas de preocupación de la ciudadanía. Generalmente, los problemas no son conocidos por las autoridades, en caso de tener conocimiento de estos, la solución no se brinda de manera rápida. Por ende, para que el proyecto tenga un objetivo claro, los objetivos que se plantearon anteriormente, es decir el objetivo general y específicos de este proyecto se toman nuevamente en esta etapa. El criterio de éxito que se tiene en este proyecto es poder identificar temas concernientes a la gestión municipal de Quito y conocer la percepción sentimental de las personas en cuanto a ese tema.

2.1.2. Valorar la situación

La situación antes de iniciar con este proyecto tuvo los siguientes detalles:

- Se previó que los datos disponibles para este proyecto son tweets concernientes a las instituciones del Municipio de Quito y otras muy relacionadas a esta que pueden ser recolectados de Twitter. Estas instituciones son nombradas en el anexo 1 ubicada en los Anexos. Es importante mencionar que esta tabla fue generada el 22 y 23 de junio de 2021, en caso de posibles cuentas añadidas o removidas de la lista 'Siguiendo' de la cuenta @MunicpioQuito en Twitter. Además, se consideró que para obtener una mejor realidad de las necesidades de la ciudadanía era necesario establecer una fecha de partida. Por ende, se eligió la fecha en que el señor Jorge Yunda asumió la alcaldía de la ciudad, es decir, el 14 de mayo de 2019.
- Asimismo, se aseguró que estas entidades tengan como ente superior a la Municipalidad Quiteña lo cual se puede constatar en la página web oficial del Municipio de Quito. Sin embargo, hay instituciones que no tienen una cuenta en Twitter. Por otra parte, existieron instituciones que tienen una cuenta en Twitter, pero que, al momento de realizar la recolección de datos, esta ha sido suspendida, robada o abandonada como la Secretaría de Inclusión (@InclusionQuito), Patronato San José (@PatronatoSJ), Salud Quito (@SaludQuito), entre otras. Por ende, aunque no se pudo obtener los tweets de estas cuentas, si fue posible obtener las menciones que ha recibido.
- 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., que influyan de manera repentina y brusca en el sentimiento de los ciudadanos y que se refleje en sus tweets.

Para los anteriores riesgos mencionados se tienen los siguientes planes de contingencias:

- Tiempo: CRISP-DM al ser una metodología flexible, en caso de tener problemas en el tiempo de entrega, se omitirán los pasos menos necesarios dentro de este proyecto y solo se contemplarían las etapas estrictamente necesarias.
- Datos incomprensibles: Aplicar otras técnicas que permitan realizar un mejor filtrado, limpieza y procesamiento de los datos en cuestión. Asimismo, el ruido dentro de los datos también se contempla.
- Factores externos: 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 o ampliarlo para contemplar diferentes tendencias en el tiempo.

2.1.3. Determinar objetivos de minería de datos

Una vez que se determinaron los objetivos del proyecto, se transforman a objetivos dentro del campo de la minería de datos. Es decir, plantear objetivos a un nivel estrictamente técnico que generen pautas claras para continuar con los siguientes pasos de la metodología que plantea CRISP-DM. Con lo cual se llegó a determinar los siguientes objetivos:

- 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 previamente, asimismo, con datos utilizados de repositorios públicos.
- Construir un modelo que extraiga tópicos derivados de los tweets recolectados mediante técnicas de aprendizaje de máquina (*machine learning).*

2.2. Comprensión de los datos

En la segunda fase, comprensión de los datos, se recolectan datos e identifican problemas en estos, ya sea por problemas de calidad, integridad, etc. Igualmente, se busca descubrir señales de posibles conjuntos de datos que ayuden a plantear estrategias o respuestas a los objetivos planteados. Se presentarán cuatro tareas: recolectar datos iniciales, describir los

datos, explorar los datos y verificar la calidad de los datos (Marbán, Mariscal, & Segovia, 2009).

2.2.1. Recolectar los datos iniciales

En esta etapa se obtuvieron dos conjuntos de datos:

El primer conjunto de datos corresponde a los tweets de las cuentas de personas (ciudadanía) e interviene en el proceso de extracción de tópicos. El segundo conjunto de datos se generó a partir de la unificación de dieciocho subconjuntos de datos e interviene en el análisis de sentimientos. Estos dos conjuntos de datos se guardaron en archivos .csv.

Los métodos usados para la recolección de datos fueron Tweepy, una librería de Python. Esta librería realiza raspado web (*web scraping*) de una cuenta con ciertos parámetros que se ajustan a discreción del programador. Se optó por este método por la limitante que tiene la API oficial de Twitter en su cantidad de tweets que se pueden obtener de manera gratuita. De cada cuenta mencionada en el anexo 1, se hizo una recopilación de tweets, con esto se obtuvo los comentarios de la ciudadanía. Se estableció un límite de 5000 tweets por cada cuenta ya que la disparidad entre tweets generados entre las cuentas utilizadas fue grande. 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) (SEPLN, 2017), y un repositorio en Github de acceso público brindado por el usuario *charlesmalafosse* (Malafosse, 2019).

Después de recolectar los datos de Twitter se hizo el ingreso de estos en el almacenamiento en la nube de MongoDB como método de respaldo. Para esto igualmente, se usó Python y la librería Pymongo que permite la conexión y carga de datos desde Python a Mongo DB.

2.2.2. Describir los datos

Los dos conjuntos de datos que se obtuvieron en el paso anterior tienen un formato .csv. El primer archivo 'data_collected_from_twitter.csv' consta de seis columnas "*Id_tweet*", "*Created_at*", "*Text*", "*Author*", "*Account_Mentioned*" y "*_id*" con 246590 registros. La columna "*Id_tweet*" es el identificador único de cada tweet, "*Created_at*" es la fecha en la que el tweet se generó, "*Text*" es el texto del tweet, "*Author*" es la cuenta que generó el tweet, "*Account_Mentioned*" son todas las cuentas que fueron mencionadas en el tweet y "*_id*" es el identificador del registro en el almacenamiento en la nube. Estos registros satisficieron los

requerimientos planteados inicialmente, ya que se obtiene el texto del tweet y la/las cuentas involucradas.

Por otra parte, el segundo conjunto de datos 'sentiment_data.csv' consta de 3 columnas "*Id_tweet*", "*Text*" y "*Sentiment*". La columna "*Id_tweet*" es el identificador único del tweet, "*Text*" es el texto del tweet y "*Sentiment*" es la polaridad del sentimiento del tweet. Estos dos archivos fueron limpiados y procesados en pasos posteriores.

2.2.3. Explorar los datos

En una revisión a los datos del archivo 'data_collected_from_twitter.csv' se observó lo siguiente:

- Las cuentas más mencionadas por la ciudadanía son @LoroHomero y @MunicipioQuito, entes principales de todas las instituciones municipales de la ciudad de Quito. Igualmente, esta última actúa como vocera de otras cuentas ya que genera tweets que involucra a otras relacionadas al Municipio de Quito.
- Las cuentas @RSQuito, @OMSCQuito y @QuitoGlobal tienen menos de 1000 menciones cada una.

En el archivo 'sentiment_data.csv' se observó lo siguiente

- La polaridad sentimental de la columna "*Sentiment*" tuvo diferentes valores como 'NONE', 'N', 'NEU', 'P', 'positive', 'POSITIVE', 'NEGATIVE', 'NEUTRAL'. Estos valores representan únicamente tres sentimientos: positivo, neutral y negativo. Por otra parte, existen valores que deben ser omitidos ya que representan ruido en los datos como '@iunida', 'rajoy', '@conrubalcaba', '@marianorajoy', '@ppopular', '#sumatealcambio', 'psoe', '@upyd'.

2.2.4. Verificar la calidad de los datos

Los datos recolectados cumplieron con los requerimientos para cumplir con el propósito de este proyecto. Se consideró la limpieza de datos como etapa importante dentro de este proceso. Se puede realizar una transformación de caracteres como emojis o tildes en el texto que comprende el campo de la columna '*Text'* en los dos archivos. Por ejemplo, en el texto "🔴.#URGENTE" después de la limpieza de datos el texto resultante debe ser ".#URGENTE".

2.3. Preparación de los datos

En la tercera fase, la preparación de los datos, existen seis tareas: seleccionar los datos, limpiar los datos, construir los datos, integrar los datos y dar formato a los datos. En esta fase 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, siempre y cuando el conjunto de datos final abarque las tareas necesarias para entregar un resultado confiable (IBM, 2011).

2.3.1. Seleccionar los datos

Después de la consideración anterior, los archivos generados anteriormente quedaron de la siguiente manera:

Campos seleccionados para el posterior análisis en el archivo data_collected_from_twitter.csv:

- ***Id_tweet*:** Es el identificador único del tweet, usado para eliminar tweets repetidos.
- ***Created_at*:** Es la fecha en la que el tweet fue generado.
- ***Text*:** Texto utilizado para realizar el análisis posterior. Es el campo más importante y en el cual se centra este proyecto.
- ***Author*:** Cuenta de Twitter la cual generó el tweet. Ver si se utiliza, tal vez para ver los tweets que son de instituciones propias del municipio.
- ***Account_Mentioned*:** Cuentas a las cuales se menciona en el tweet.

Campos seleccionados para el posterior análisis en el archivo sentiment_data.csv. Este archivo se generó a partir de la integración de diferentes conjuntos de datos explicados en el siguiente paso:

- ***Id_tweet*:** Es el identificador único del tweet, usado para eliminar tweets repetidos.
- ***Text*:** Texto utilizado para realizar el análisis posterior. Es el campo más importante y en el cual se centra este proyecto.
- ***Sentiment*:** Es la polaridad del sentimiento en el tweet.

2.3.2. Integrar los datos

El archivo sentiment_data.csv se conformó a partir de la integración de dieciocho subconjuntos de datos. Los dieciocho subconjuntos de datos en su totalidad fueron obtenidos

de bases de datos públicas, de estos dieciséis fueron archivos .xml es así como se utilizó la librería ElementTree de Python para poder leer, extraer y unificar los datos necesarios. Los dos subconjuntos de datos en formato .csv restantes se leyeron y extrajo su contenido. Después de la integración constaron 42165 registros en el archivo llamado 'sentiment_data.csv'.

2.3.3. Limpiar los datos

Dentro de esta etapa se aplicaron las siguientes acciones a los dos archivos para aumentar la calidad de los datos previamente recolectados:

- **Eliminar filas repetidas:** Se eliminaron los registros repetidos mediante la comparación de sus IDs. Este dato fue recolectado en los tweets y representa el identificador único del tweet, por lo cual no puede ser repetido.
- **Eliminar valores nulos:** Se eliminaron los registros cuyo campo '*Text'* no contenga ningún dato, es decir, un campo vacío.
- **Eliminar retweets:** No se consideraron los *retweets* debido a la repetición del texto. Al identificar "RT @" al inicio del tweet se borró dicho tweet.
- **Control de mayúsculas y minúsculas:** Se controló los caracteres en mayúsculas del campo '*Text'*, de tal manera todo el texto se convirtió a minúsculas. Por ejemplo, "La voluntad popular" se transformó a "la voluntad popular".

En el archivo 'data_collected_from_twitter.csv' se añadió un paso:

- **Eliminar tweets cuentas propias:** Se eliminaron los tweets que mencionan a la propia cuenta. Por ejemplo, la cuenta @MunicipioQuito generó el siguiente tweet "@jai_mito15 Buenos días, los vehículos eléctricos e híbridos están exentos de la medida de restricción. Saludos @MunicipioQuito.", en este ejemplo se observa que la cuenta que genera el tweet también existe en el texto. Esto representa ruido en la detección de tópicos ya que, una institución puede mencionar y expresar sentimientos o tópicos que afectan al resultado final.

En el archivo 'sentiment_data.csv', se añadió un paso:

- **Convertir sentimientos:** Este paso transformó todas las polaridades presentes a únicamente tres (positivo, neutral, negativo).

Después de realizar estos controles, el archivo data_collected_from_twitter.csv pasó de 246590 a 157293 registros y el archivo sentiment_data.csv originalmente con 42165, tuvo 37293 registros.

2.3.4. Construir los datos

Los siguientes pasos aplican tanto para la extracción de tópicos y el análisis de sentimientos:

- **Remover urls, usernames & hashtags:** Se eliminaron todas las direcciones web, nombres de usuarios y los *hashtags* presentes en cada tweet. Por ejemplo, el tweet *"@TumbacoI @MunicipioQuito No se escucha bien. Yo por mi parte hice un resumen de Todas las Preguntas y Dudas acerca del tema de #Botániqo. Tal vez a alguien le interese en #Cumbayá, #Tumbaco. https://t.co/efoFrh4YMq"* después de aplicar este procedimiento cambió a *"No se escucha bien. Yo por mi parte hice un resumen de Todas las Preguntas y Dudas acerca del tema de . Tal vez a alguien le interese en , . ".* Esto se hizo para tener un mejor entrenamiento tanto en la extracción de tópicos y análisis de sentimientos, evitando palabras con aporte nulo al resultado final.
- **Remover caracteres:** Los caracteres *.,;:$%&#()*+<>=¿?!¡@/"^_`~\[\]\{\}/\-\\]"* fueron eliminados del texto de cada tweet con el objetivo de dejar únicamente caracteres alfabéticos. También se eliminó el salto de línea *\n.* Letras repetitivas también fueron eliminadas, por ejemplo, en el texto "noooo" después de este proceso el texto fue "no".
- **Comprobar tweet en español:** Se comprobó que el texto de cada tweet está escrito en español. Se utilizaron dos librerías para este propósito *langdetect* y *textblob*. Cada librería provee un método que detecta el idioma de un texto indicado. Es así como se hizo un filtrado de cada registro según el idioma del texto del tweet. En caso de que una de estas dos librerías detecte que el texto ingresado, es decir cada tweet, está escrito en español, se considera dicho tweet caso contrario no. Por ejemplo, el texto *"Este juego me gusta"* es considerado en español y el texto "*I don't like this game*" no se considera en español.
- **Tokenizar tweets:** Después del filtrado por el idioma del texto, se realizó la tokenización del texto. Esto consistió en segmentar el texto en palabras mediante la identificación de un carácter, en este caso el espacio " ". Además, se eliminaron las palabras no útiles (*stopwords*) del idioma español. El proceso consistió en el ingreso de un texto y retornar una lista de palabras. Por ejemplo, el texto de entrada fue "Ecuador es un gran país" y la salida obtenida fue "['Ecuador', 'gran', 'país']".
- **Eliminar tokens no alfabéticos:** Con la finalidad de obtener mayor precisión en el modelo a entrenar, se optó por eliminar los caracteres no alfabéticos. Anteriormente, se

eliminaron signos de puntuación, en este paso se eliminaron los caracteres numéricos. Estos pueden ser tokens que tienen solo caracteres numéricos o una combinación entre alfabéticos y numéricos. Por ejemplo, el token “7am”, “0988765432” y “1000mg” se eliminaron, y tokens como “ciudad”, “alcalde” y “casas” no se eliminaron.

- **Lematizar:** Se hizo la lematización de los tokens que superaron el anterior filtro. Esto consistió en eliminar las variaciones gramaticales de la palabra o token y convertirlos a su forma base. Adicionalmente, se eliminaron todas las partes de la oración (*parts of speech)* menos los sustantivos que ayudaron a obtener los tópicos.
- **Crear bigrams:** Mediante la librería NLTK de Python, se generó *bigrams* a partir de los tokens generados anteriormente. Por ejemplo, a partir de “['Ecuador', 'gran', 'país']” se generó “[('Ecuador', 'gran'), ('gran', 'país')]”.
- **Juntar bigrams:** Para generar una entrada limpia y entendible en el modelo a entrenar, se generó “palabras” que derivaron de la unión de los *bigrams* creados anteriormente. Por ejemplo, de la entrada “[('Ecuador', 'gran'), ('gran', 'país')]” se obtuvo la salida “['Ecuador_gran', 'gran_país]”. Con esto, se pretende simular una palabra, aunque contemple el caracter “_”.

Pasos adicionales para la extracción de tópicos:

- **Eliminar registros con menos de tres tokens:** En caso de que un tweet, después de los pasos anteriores resultó con tres o menos tokens, se eliminó debido a que no se tienen suficientes *tokens* para poder realizar el proceso de extracción de tópicos. Esto pudo derivar en ruido en los datos.

Pasos adicionales para el análisis de sentimientos:

- **Balancear datos:** Para la generación del modelo de entrenamiento fue necesario que la cantidad de datos de sentimientos positivos, negativos y neutrales sean igual. Para esto se estableció 10000 registros de cada sentimiento, en total 30000.

Después de todos los pasos mencionados, los archivos tuvieron la siguiente cantidad de registros:

- El archivo tweets_cleaned obtenido a partir del procesamiento del archivo data_collected_from_twitter.csv usado para la extracción de tópicos resultó con 72554 registros y seis columnas las cuales se muestran en la tabla 2.

Tabla 2. Columnas del archivo tweets_cleaned

Columna	Función
Id_tweet	Identificador único del tweet.
Created_at	Fecha y hora de la creación del tweet.
Text	Texto sin procesar del tweet.
Text_Tokenized	*Tokens* generados a partir de la columna '*Text'* aplicado el procesamiento.
Texto_Limpio_Noun	*Tokens* filtrados a partir de la columna 'Text_Tokenized' solo sustantivos.
Texto_Limpio_Noun_Bigram	Bigrams generados a partir de la columna 'Texto_Limpio_Noun'.

- El archivo sentiment_data_cleaned obtenido a partir del procesamiento del archivo sentiment_data.csv usado para el análisis de sentimientos resultó con 30000 registros y ocho columnas las cuales se muestran en la tabla 3.

Tabla 3. Columnas del archivo sentiment_data_cleaned

Columna	Función
Id_tweet	Identificador único del tweet.
Text	Texto sin procesar del tweet.
Sentiment	Polaridad del sentimiento, positivo, negativo o neutral.
Text_Token	*Tokens* generados a partir de la columna '*Text'*.
Texto_Limpio	*Tokens* generados aplicando la lematización a partir de la columna 'Text_Token'.
Texto_Limpio_Bigrams	Bigrams generados a partir de la columna 'Texto_Limpio'.
Texto_Limpio_joined	*Tokens* unidos generados a partir de la columna 'Texto_Limpio'.
Text_Token_joined	*Tokens* unidos generados a partir de la columna 'Text_Token'.

2.3.5. Dar formato a los datos

En este punto se hicieron las siguientes acciones para cada conjunto de datos:

- **Control del orden de registros:** Con el objetivo de evitar crear el modelo con datos cronológicamente ordenados, lo que pudo causar un entrenamiento impreciso, se ordenó aleatoriamente los datos.

2.4. Modelamiento

La siguiente fase es el modelamiento, en esta fase se selecciona el modelo correcto para cumplir con los objetivos establecidos. Existen diferentes abordajes y, por ende, diferentes modelos que pueden ser aplicados. Sin embargo, los datos generados deben cumplir las necesidades de dichos modelos por lo cual, los modelos deben acoplarse a estos. Dentro de esta fase, se presentarán cuatro tareas: seleccionar las técnicas de modelamiento, generar diseños de pruebas, construir el modelo y evaluar el modelo (IBM, 2011).

2.4.1. Selección de la técnica de modelamiento

Se seleccionaron tres diferentes técnicas en el modelamiento. Para la extracción de tópicos se desarrollaron dos técnicas, cada una de ellas generó dos modelos (*unigrams* y *bigrams*). La primera fue LDA y la segunda GSDMM. LDA y GSDMM son algoritmos usados dentro del Procesamiento del Lenguaje Natural NLP (*Natural Language Processing*) (Jelodar, Wang, Yuan, & Feng, 2017). Los datos de entrada en este proyecto (tweets) difieren en el tamaño de su texto a textos en libros o revistas. Los tweets al tener un máximo de 280 caracteres deben ser analizados de diferente manera. LDA es el método más usado para la extracción de tópicos de textos con grandes volúmenes de palabras.

Por otra parte, acorde a la literatura, GSDMM se desenvuelve mejor con textos de menor tamaño (Yin & Wang, 2014). Por ende, se desarrollaron los dos modelos para, posterior a las pruebas, determinar cuál es el modelo que mejor desempeño tiene con los datos utilizados en este proyecto. Es importante mencionar que se necesita especificar la cantidad de tópicos con los que se quiere entrenar ambos modelos.

Dentro de la extracción de tópicos, las partes de la oración (*parts of speech*) que tienen mayor influencia en este proceso son los sustantivos (Kibble, 2013). Cada parte de la oración (*part of speech*) da una característica única a una palabra dentro de un texto, por ejemplo, los sustantivos se refieren a gente, personas y cosas, por otra parte, los verbos describen acciones. Por ende, para la extracción de tópicos se decidió considerar solamente los sustantivos. Por otra parte, en el análisis de sentimientos se optó por la Regresión Logística, como asunción en este caso, la entrada al modelo consideró todo el texto, es decir, el tweet completo sin ningún procesamiento. Esta decisión se tomó con base a la comparación y evaluación del texto sin procesamiento con el texto ya procesado.

2.4.2. Generar diseño de pruebas

Dentro de este proyecto existen dos macroprocesos: extracción de tópicos y análisis de sentimientos. La prueba utilizada para la extracción de tópicos fue la coherencia generada a partir de los modelos creados. 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 (Hansen, McMahon, & Prat, 2014). Por ende, también se consideró la facilidad de inferencia de los tópicos como un factor para determinar el modelo que mejor se ajusta a las necesidades de este proyecto.

Este último factor al no tener una base científica que respalde su proceso es relativo al igual que la selección de palabras que define a un tópico. 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. La medida de efectividad fue el área bajo la curva (*area under the curve*) con un resultado no menor de 0.75.

2.4.3. Construir el modelo

Por parte de la extracción de tópicos, se establecieron dos condiciones que aplicaron tanto para LDA como para GSDMM en sus dos opciones (*unigram* y *bigrams*). La primera fue la cantidad de tópicos que se estableció en siete. Por otra parte, se realizaron diez iteraciones para realizar un promedio entre las ejecuciones de cada modelo.

Para construir el modelo LDA, se creó una función que genera y entrena el modelo. Esta función toma 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. El corpus es un conjunto de documentos utilizado en el entrenamiento del modelo, el diccionario está conformado de todas las palabras que conforman el *corpus* del entramiento, en este caso, los tweets recolectados. Como salida se retorna el modelo entrenado, la coherencia de ese modelo y la cantidad de tópicos con la que se entrenó tal modelo. Este proceso se hizo dos veces, el primero se hizo para generar el modelo que usó *unigram* y el segundo con *bigrams*. El conjunto de datos de entrada estuvo compuesto, entre otras columnas, por el texto separado en *tokens* en *unigram* y *bigrams*.

En cambio, el segundo abordaje que se consideró para la extracción de tópicos fue GSDMM debido a su mejor desempeño en textos cortos (Yin & Wang, 2014). Se hizo uso de la librería *MovieGroupProcess* en la cual los parámetros que se modificaron fueron la cantidad de

tópicos, los valores *alfa* (0.1), *beta* (0.1) y el número de iteraciones (30). Posteriormente este modelo fue entrenado con los tweets. Se generó un corpus y diccionario y, 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 una con *unigram* y la segunda con *bigrams*. Por otra parte, la construcción del modelo para el análisis de sentimientos comenzó con la lectura de los 30000 registros del archivo sentiment_data_cleaned que fue generado en el paso 2.3.4 (Construcción de los datos) de este documento. Este archivo contiene los tweets obtenidos de repositorios públicos procesados para el entrenamiento y prueba del modelo a entrenar.

Después de haber leído los datos, estos se dividen un conjunto de datos para entrenamiento y otro para pruebas, esta división fue aleatoria y se realizó con la función *train_test_split* de la librería sklearn.*model_selection.* La columna que contiene las etiquetas (variable dependiente) es *Sentiment*. En cambio, los datos que actuaron como características para el entrenamiento (variables independientes) se obtuvieron de la columna *Text*, *Texto_Limpio_joined* o *Text_Token_joined.* Las tres columnas difieren en su contenido como se explicó en el punto 2.3.4. Se tuvo las tres opciones para evaluar cual tiene mejor desempeño y elegir el mejor. La proporción de la división de los registros fue 90% para entrenamiento y 10% para pruebas, esto se traduce en 27000 y 3000 registros respectivamente.

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*. Este proceso se llevó a cabo con *TfidfVectorizer* de la librería Scikit Learn, se estableció que deben existir como mínimo 3 repeticiones de un término para que pueda ser considerado dentro del vocabulario. Igualmente, se estableció que se pueden generar *unigram* y *bigrams*. Posteriormente, aprendió el vocabulario desde el conjunto de datos de entrenamiento mediante la función 'fit'. A continuación, se transformó los datos de entrenamiento (variables independientes) a la matriz de términos de documentos. Esta matriz usa el vocabulario y las frecuencias de los documentos aprendidas con la función 'fit'. Después de realizar los pasos anteriores, se tuvo listo el input que recibiría el modelo, este input tuvo una forma (27000,32388) es decir, 27000 registros y 32388 características cuyo valor es un valor numérico.

El siguiente paso fue crear el modelo de Regresión Logística que es usado para crear y entrenar el clasificador. Se utilizó la función LogisticRegression de la librería sklearn, en la cual se establecieron 30000 iteraciones como máximo y se seleccionó multinomial en el

parámetro multi_class, este último da la capacidad de realizar una distribución de probabilidad ya que existen más de dos clases (positivo, negativo y neutral). A continuación, se entrenó el modelo con la matriz de términos de documentos obtenida con anterioridad a partir del conjunto de datos de entrenamiento y sus respectivas etiquetas. Finalmente, se realizó la predicción con el conjunto de datos de prueba y sus etiquetas, esto resultó en un array con los resultados cuya forma es (3000,1).

Para determinar si el clasificador generado es bueno se evaluó con el área bajo la curva ROC (AUC ROC). En esta parte, se utilizaron las predicciones en el paso anterior y las etiquetas separadas como datos de prueba. Para esto se empleó la función roc_auc_score de la librería sklearn, esta función contempló un parámetro llamado multi_class cuyo valor se determinó en 'ovr'. Al establecer 'ovr' en este parámetro, se indicó que se haga una comparación de una contra el resto (one vs rest), esto trata el caso multiclase de la misma manera que el caso de múltiples etiquetas (learn, 2007-2021).

2.5. Evaluación

La fase de evaluación busca asegurar que el modelo generado ha seguido todos los pasos necesarios para cumplir con su objetivo. Igualmente, si se construyen diferentes modelos se seleccionará el que tiene un mejor rendimiento. Aquí se realizan las siguientes tareas: evaluar los resultados, revisar el proceso y determinar los siguientes pasos (IBM, 2011).

2.5.1. Evaluar los resultados

Dentro de la evaluación de resultados se siguió una secuencia de pasos para determinar el mejor modelo para la extracción de tópicos y, por parte del análisis de sentimientos, saber si cumplió o no con el objetivo de tener 0,75 o más en la medida de efectividad.

Para la extracción de tópicos se siguieron los siguientes pasos:

1. Obtener los resultados de los modelos LDA

El modelo ejecutado con *unigram* con siete tópicos tuvo el resultado detallado en la tabla 4.

La coherencia promedio del modelo LDA con *unigram* fue 0.254486, el modelo con mayor coherencia fue el de la novena repetición con 0.285891. Por otra parte, el modelo ejecutado con *bigrams* tuvo el resultado detallado en la tabla 5.

La coherencia promedio del modelo LDA con *bigrams* fue 0.680919 y el modelo con mayor coherencia fue el de la cuarta repetición con 0.704175. Por tal razón, entre los dos modelos ganadores, el modelo *bigrams* (0.704175) presentó una superioridad notoria al modelo *unigram* (0.285891) por lo cual, el modelo de la cuarta repetición de *bigrams* resultó como modelo ganador en LDA.

Tabla 4. Ejecución modelo LDA unigram

Número de Repetición	Coherencia
1	0.224829
2	0.248049
3	0.259157
4	0.254377
5	0.216439
6	0.277330
7	0.254309
8	0.261668
9	0.285891
10	0.262819

Tabla 5. Ejecución modelo LDA bigrams

Número de Repetición	Coherencia
1	0.681681
2	0.660066
3	0.691663
4	0.704175
5	0.687070
6	0.676483
7	0.660308
8	0.690107
9	0.681269
10	0.676370

2. Obtener los resultados de los modelos GSDMM

El modelo ejecutado con *unigram* tuvo los resultados detallados en la tabla 6. La coherencia promedio del modelo GSDMM con *unigram* fue 0.422002 y el modelo con mayor coherencia fue el de la sexta repetición con 0.478228. Por otra parte, el modelo ejecutado con *bigrams* tuvo los resultados detallados en la tabla 7.

Tabla 6. Ejecución modelo GSDMM unigram

Cantidad de Tópicos	Coherencia
1	0.388155
2	0.429214
3	0.372865
4	0.344683
5	0.461296
6	0.478228
7	0.442007
8	0.446964
9	0.451765
10	0.404852

Tabla 7. Ejecución modelo GSDMM bigrams

Cantidad de Tópicos	Coherencia
1	0.495778
2	0.466464
3	0.521279
4	0.467625
5	0.491858
6	0.489293
7	0.481592
8	0.455034
9	0.508997
10	0.486343

La coherencia promedio del modelo GSDMM con *bigrams* fue 0.486426 y el modelo con mayor coherencia fue el de la tercera repetición con 0.521279.

No existió una notoria diferencia entre las coherencias de los dos modelos, por ende, la decisión se tomó con base a la facilidad de deducir los tópicos derivados de las palabras significativas de cada grupo obtenido en el entrenamiento. En la tabla 8 se presentan las palabras significativas por clúster en el modelo ganador de *unigram* de GSDMM.

En la tabla 9 se presentan los pares de palabras significativos por clúster en el modelo ganador de *bigrams* de GSDMM.

En la tabla 8 y 9 se presentaron las 10 palabras y pares de palabras más significativos tanto del modelo GSDMM *unigram* como *bigrams* respectivamente. Después de observar las palabras significativas se determinó los tópicos que definen (engloban) dichas palabras. En el modelo *unigram* se asignó los siete tópicos más rápido que en el modelo *bigrams*, este último tuvo tres grupos a los cuales no fue posible asignarles un tópico adecuado por la variabilidad y poca similitud entre sus palabras significativas, este caso se presenta en el tópico 2, 3 y 5 del modelo *bigrams*.

Tabla 8. Palabras significativas por clúster del modelo GSDMM unigram

N° de tópico	**Palabras significativas**	**Tópico asignado**
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 de las autoridades
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

Tabla 9. Palabras significativas por clúster del modelo GSDMM bigrams

N° de tópico	Palabras significativas	Tópico asignado
0	'fin semana', 'ciudad humanidad', 'declaración ciudad', 'humanidad ceremonia', 'ceremonia través', 'lunes viernes', 'kit alimento', 'persona edad',	Actos humanitarios
1	'música volumen', 'servicio agua', 'rata rata', 'fin semana', 'volumen fiesta', 'agua sector', 'uso espacio', 'escándalo música', 'lunes junio', 'arte	Contaminación auditiva
2	'camino eucalipto', 'número contacto', 'uso mascarilla', 'remoción alcalde', 'festival evento', 'mes tradición', 'tradición festival', 'través agenda',	*
3	'modelo gestión', 'fin semana', 'situación vulnerabilidad', 'uso espacio', 'ayuda persona', 'música quinteto', 'actividad espacio', 'proceso vía',	*
4	'unidad mantenimiento', 'requerimiento planificación', 'planificación unidad', 'reporte requerimiento', 'calle calle', 'mantenimiento área', 'mantenimiento	Falta de gestión municipal
5	'fin semana', 'accidente tránsito', 'ciudad alcalde', 'derechos humanos', 'parte jornada', 'teatro parte', 'alcalde mano', 'persona calle', 'unidad seguridad',	*
6	'camión gas', 'problema pueblo', 'ruido decibel', 'decibel camión', 'gas app', 'app problema', 'acción protección', 'cultura espacio', 'sector calle', 'número	Contaminación auditiva

(*): No se pudo determinar un tópico de manera clara debido a la poca similitud entre sus pares de palabras más significativos.

Después de haber considerado el valor de la coherencia y la facilidad para definir los tópicos según las palabras más significativas, se determinó que el modelo ganador de GSDMM fue el modelo *unigram*. Esta decisión, principalmente, fue tomada ya que a todos los grupos creados en el entrenamiento (siete) se les asignó la definición del tópico, es decir, las palabras que definen el tópico, de manera más sencilla a comparación del modelo *bigrams*. Por ende, el modelo *unigram* de la sexta repetición fue el ganador en GSDMM.

Después de haber obtenido el modelo ganador tanto de LDA como de GSDMM, se obtuvo las palabras más significativas de cada tópico en los dos modelos como se detallan en las

tablas 10 y 8. Para presentar las palabras más significativas más significativas del modelo de GSDMM se considera la tabla 8.

3. Comparar LDA con GSDMM y definición de tópicos.

En el modelo ganador de LDA, es decir, el modelo de la cuarta repetición de *bigrams* tuvo los siguientes pares de palabras en cada tópico.

Tabla 10. Palabras significativas modelo LDA

N° de tópico	Palabras significativas	Tópico asignado
0	'camión gas', 'decibel camión', 'gas app', 'problema pueblo', 'app problema', 'ruido decibel', 'acción protección', 'contenedor basura', 'resultado prueba',	Contaminación auditiva
1	'remoción alcalde', 'uso espacio', 'parte parte', 'calle hueco', 'sector calle', 'compra prueba', 'control espacio', 'dirección número', 'persona proceso',	*
2	'fin semana', 'caso tiempo', 'cultura espacio', 'actividad museo', 'alcalde ciudad', 'semana confinamiento', 'espacio labor', 'grupo atención',	*
3	'prueba pcr', 'tal vez', 'registro propiedad', 'ciudad humanidad', 'calle ciudad', 'ceremonia través', 'declaración ciudad', 'humanidad ceremonia', 'red	Actividades de gestión del municipio
4	'inicio semana', 'recolección basura', 'música volumen', 'número contacto', 'calle sector', 'calle calle', 'obra ciudad', 'alcalde alcalde', 'liderazgo	Atención en necesidades de los barrios
5	'camino eucalipto', 'sector ciudad', 'lunes junio', 'arte magia', 'magia lunes', 'junio show', 'persona edad', 'show conéctate', 'bus corredor', 'modelo operación'	Actividades culturales
6	'proceso vía', 'vía nivel', 'nivel instancia', 'vuelo proceso', 'basura calle', 'ciudad alcalde', 'solución problema', 'mes trámite', 'acera deslav', 'millón	*

(*): No se pudo determinar un tópico de manera clara debido a la poca similitud entre sus pares de palabras más significativas.

Al comparar el valor de las coherencias del modelo ganador de GSDMM (0.478228) con el de LDA (0.704175), se observó que existe una gran diferencia entre estas. En la tabla 10 se

detalla los resultados de la determinación de los tópicos según las palabras obtenidas dentro de cada clúster esto se obtuvo con el modelo LDA. En cambio, en la tabla 8 se observan los mismos resultados, pero con el modelo GSDMM. En este caso, una persona hispanohablante puede inferir los tópicos generados en los resultados presentados por el modelo GSDMM con mayor facilidad que en el modelo LDA. Por otra parte, según la literatura GSDMM considera un texto como perteneciente de un solo tópico y al considerar que un tweet generalmente habla de un solo tema debido a su tamaño (máximo 280 caracteres), GSDMM tiene superioridad en la decisión del proyecto. Es así como, a pesar de que la coherencia en el modelo LDA fue mayor, el modelo GSDMM presentó mayor facilidad de inferencia de los tópicos generados por su modelo. Por tal motivo, el modelo GSDMM fue el ganador debido a que se ajustó mejor a las necesidades de este proyecto y los tópicos obtenidos se presentan en la tabla 8 (en donde están los tópicos de GSDMM).

Por otra parte, para el análisis de sentimientos, como se mencionó en la anterior fase de la metodología, se optó por la regresión logística. El conjunto de datos se dividió tanto para entrenamiento y prueba por lo cual, 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. Es así como después de realizar la construcción, entrenamiento y prueba del modelo con Regresión Logística, los resultados fueron los que se presentan en la tabla 11:

Tabla 11. Resultado AUC ROC

Columna de entrada	AUC ROC
Text	0.892476
Texto_Limpio_joined	0.848056
Text_Token_joined	0.879383

Estos resultados se traducen en un mejor desempeño de la columna *Text*, esta columna contiene los tweets recolectados sin ningún tratamiento o modificación. Por ende, el modelo que fue entrenado con la columna *Text* fue el ganador. En la figura 4 se visualiza las curvas ROC generadas por cada clase (Positivo, Negativo, Neutral) a partir del modelo ganador que fue entrenado con la columna *Text*. El promedio de los valores presentados en la figura 4 es 0.892476. Él área bajo la curva ROC se usa para comparar diferentes modelos de clasificación, en este caso, se usó para evaluar que el modelo elegido tiene un resultado mayor de lo determinado (0.75) lo cual fue cumplido, en este caso las tres clases (Positivo, Negativo, Neutral) cumplieron este criterio .

Figura 4. Gráfica AUC ROC

2.6. Desarrollo del sistema de visualización

La metodología CRISP-DM brinda pautas para realizar proyectos de minería de datos, pero no abarca toda la construcción de sistemas por tal razón, el marco de trabajo que se siguió en el componente de visualización fue Scrum que está dentro de las metodologías ágiles. Debido a la naturaleza de este componente fue necesario utilizar Scrum ya que, presentar resultados rápidos y progresivos era una necesidad imperiosa.

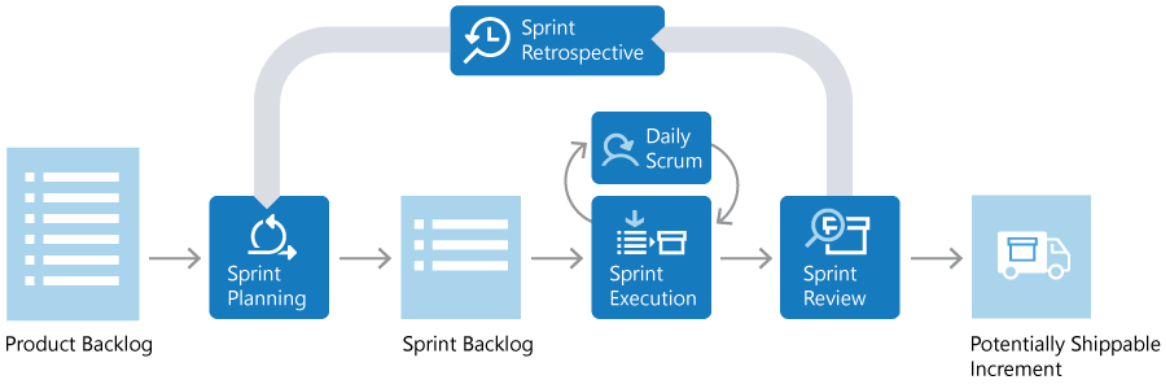


Figura 5. Ciclo de vida de un Sprint (Jacobs & Kaim, 2021)

Scrum determina diferentes aspectos los cuales se presentan a continuación:

- Roles

Los roles de Scrum determinan cargos específicos que se tienen dentro de un proyecto para que este pueda cumplir con los objetivos planteados. En este proyecto el scrum master y el equipo de desarrollo fueron llevados a cabo por el autor del proyecto, en cambio, el dueño del producto (*product owner*) es el director de este proyecto.

- Backlog

El backlog del producto es en donde se especifican todas las características y/o requerimientos que debe tener el proyecto. En este caso los requerimientos fueron:

Tabla 12. Backlog

N° Historia	Historia de Usuario	Prioridad	Puntos de historia
1	Como dueño del producto necesito distinguir los tópicos de mayor importancia obtenidos para priorizarlos en el dashboard.	Alta	5
2	Como dueño del producto necesito conocer la polaridad sentimental clasificada por sectores de Quito para determinar los sectores que necesitan mayor atención.	Alta	8
3	Como dueño del producto necesito diferenciar los tópicos más relevantes por sectores de Quito para presentar los principales problemas divididos por sectores	Media	8
4	Como dueño del producto necesito conocer la distribución de sentimientos acumulada para concluir generalizadamente los resultados del proyecto visualmente.	Baja	3

- Sprint Backlogs

Dentro de la construcción del sistema de visualización se tuvieron diferentes sprints los cuales se detallan en la tabla 13.

Tabla 13. Sprint Backlogs

N° Historia	Historia de Usuario	Sprint	Actividades
1	Como dueño del producto necesito distinguir los tópicos de mayor importancia obtenidos para priorizarlos en el dashboard.	1	Se importó los datos, crear las variables, construir la gráfica y editar el estilo final. En esta gráfica cada tópico consta de tres barras (negativo, positivo, neutral)
2	Como dueño del producto necesito conocer la polaridad sentimental clasificada por sectores de Quito para determinar los sectores que necesitan mayor atención.	2	Crear la gráfica en la que se presenta la polaridad sentimental clasificada por sectores de Quito de manera que se distinga los sectores de mayor importancia.
3	Como dueño del producto necesito diferenciar los tópicos más relevantes por sectores de Quito para presentar los principales problemas divididos por sectores	2, 3	Se realizó la gráfica en la que se diferencian los tópicos más relevantes por sectores de Quito. Esto se realizó entre el sprint 2 y 3.
4	Como dueño del producto necesito conocer la distribución de sentimientos acumulada para concluir generalizadamente los resultados del proyecto visualmente.	3	En esta gráfica se presenta un diagrama de pastel con tres secciones que representan las polaridades sentimentales y los acompaña la cantidad de tweets catalogados en cada sentimiento.

- Ceremonias

Se presentan en la tabla 14 las ceremonias que se hicieron por cada sprint marcadas con un X, las que no se marcan con un guion medio (-).

Tabla 14. Ceremonias realizadas

Ceremonia	Sprint				
		1	**2**	**3**	**4**
	Sprint	X	X	X	X
	Sprint Planning	X	-	X	-
	Daily Scrum	X	X	Intermitente	X
	Sprint Review	-	X	-	X
	Sprint Retrospective	-	X	-	X

- Burndown Chart

El burndown chart de la figura 6 de este proyecto contempla las actividades o tareas pendientes por realizar en cada sprint (línea morada) y las tareas realizadas (línea naranja).

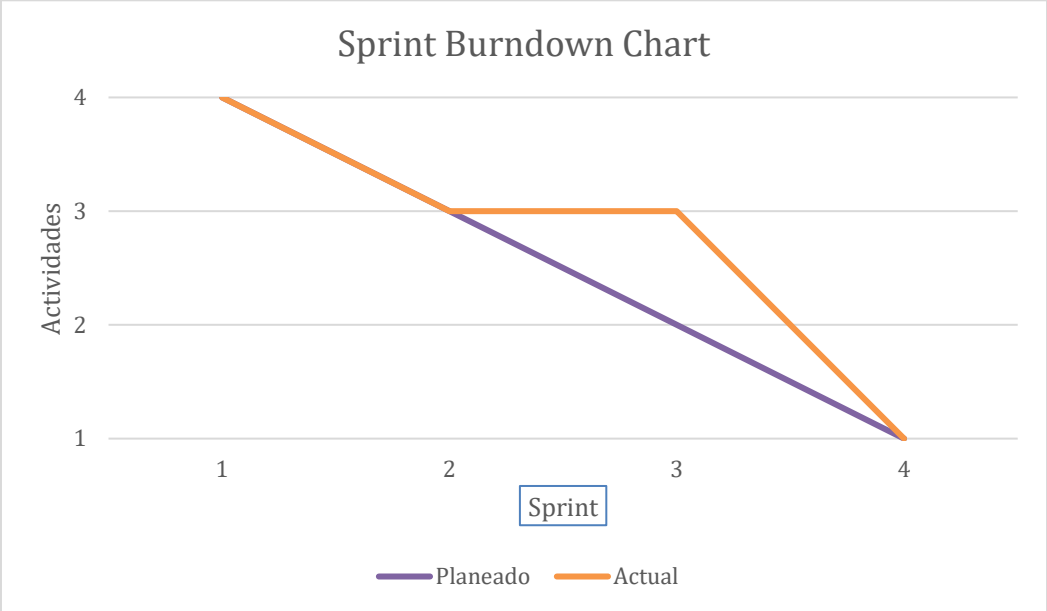


Figura 6. Burndown chart

2.6.1. Arquitectura del sistema

Este proyecto se basó en una arquitectura por capas, las capas que se presentaron fueron tres, la primera fue la capa de Datos, la capa lógica o de negocios y la capa de presentación como se presenta en la figura 7.

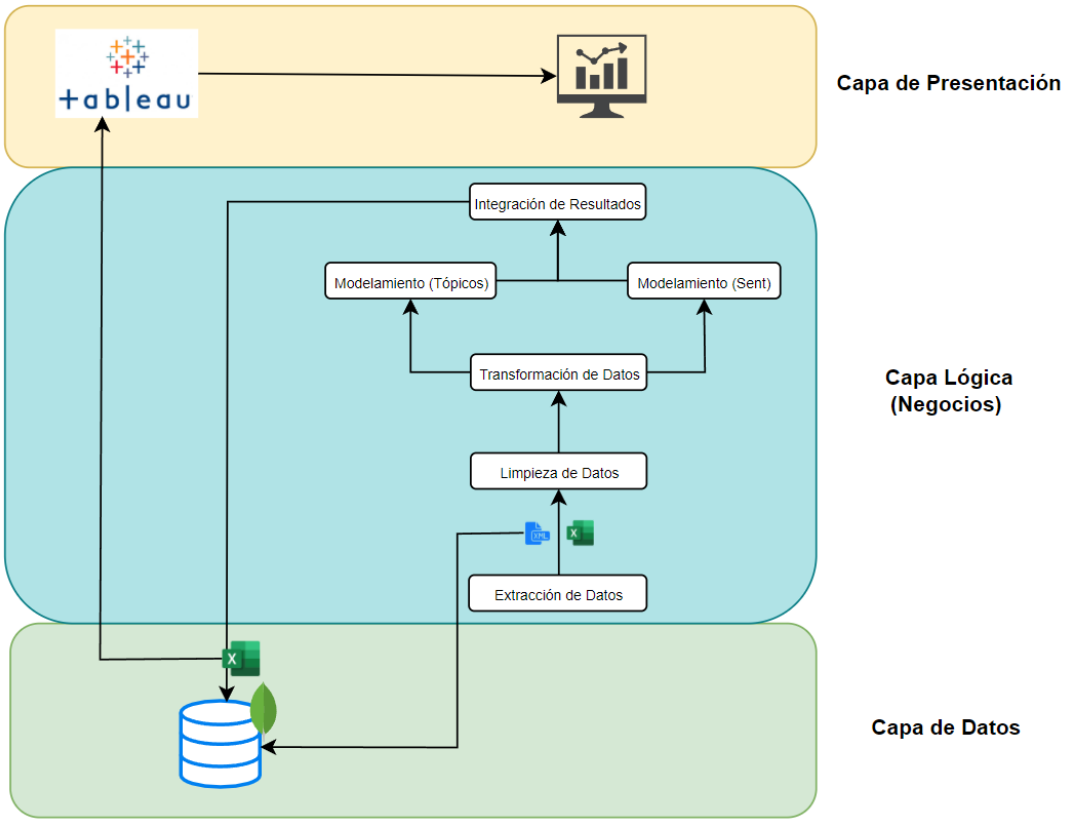


Figura 7. Arquitectura del sistema

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 en la cual se procesan los datos y se elimina el ruido que estos datos pueden tener. Además, se realiza la transformación de datos en la que se adaptan estos para la siguiente etapa. Se continua con el componente de modelamiento en el cual se realiza la generación de los modelos tanto de la extracción de tópicos como del análisis sentimental. Finalmente, se realiza la integración de los resultados de los dos modelos generados

anteriormente y la creación de los datos finales que se presentan posteriormente en 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.

2.6.2. Visualización

Para presentar los resultados se trabajó en la herramienta Tableau Public, esta herramienta ofrece almacenamiento gratuito para los tableros interactivos web, además, son de libre acceso lo cual da la posibilidad de socializar de mejor manera los resultados de este proyecto. Tradicionalmente, los tableros de mando (*dashboards*) han sido creados para mejorar el proceso de toma de decisiones dentro de una empresa o grupos con objetivos específicos (Yigitbasioglu & Velcu, 2012). Estos son la recopilación de diferentes gráficos presentados en un solo cuadro para optimizar las decisiones que se toman con respecto a los resultados presentados. Generalmente, los tableros de mando (*dashboards*) sirven para monitorear en tiempo real cualquier tipo de medida que sea de interés para la empresa, puede ser datos de inventarios, ventas, cantidad de clientes, etc. A continuación, se detalla la descripción de los pasos para generar un tablero de mandos (*dashboard*).

1. Cargar el archivo generado por los modelos entrenados en el aplicativo web.
2. Crear la hoja de cálculo (*worksheet*) en donde se va a generar el gráfico.
3. Seleccionar las variables que se van a manejar para crear el tablero de mando (*dashboard)*, en este caso fueron las columnas '*Topic'* y '*Sentiment'* las cuales con tienen el tópico asignado al tweet y la polaridad del sentimiento del tweet respectivamente.
4. Crear el parámetro '*End Date*' y '*Start Date'* que sirvieron para aplicar un filtro de rango de fecha máxima y mínima respectivamente.
5. Aplicar el filtro de fechas con los parámetros construidos en el paso 3.
6. Aplicar colores a las barras de la polaridad del sentimiento para distinguirlas entre negativo, neutral y positivo.
7. Construir el tablero de mando (*dashboard*) utilizando la hoja de cálculo (*worksheet*) creado en el paso 2 como en la figura 13.
8. Guardar el tablero de mando (*dashboard*) y obtener el enlace de libre acceso el cual puede ser compartido. Esto puede ser visualizado en el siguiente enlace: https://public.tableau.com/app/profile/eio5858/viz/SubtopicsLocations/Dashboard

Los resultados que se comentan en esta sección se presentan en el punto Resultados y Discusión.

3. RESULTADOS Y DISCUSIÓN

Después de realizar los procesos mencionados en los pasos previos, se añadieron dos pasos para generar las visualizaciones. La primera fue integrar los dos resultados, es decir, los resultados de la obtención de los tópicos y los del análisis sentimental. Esto se consiguió mediante la creación de un nuevo archivo que se conformó de ocho columnas:

- **Id_tweet:** Es el identificador único del tweet, usado para eliminar tweets repetidos.
- **Created_at:** Es la fecha en la que el tweet fue generado.
- **Text:** Texto utilizado para realizar el análisis posterior. Es el campo más importante y en el cual se centra este proyecto.
- **Texto_Limpio_Noun:** *Tokens* filtrados a partir de la columna 'Text_Tokenized' solo sustantivos.
- **Topic:** Tópico en el cual fue clasificado el tweet.
- **Topic Score:** Valor que demuestra la certeza que se tiene de que el tweet perteneza al tweet asignado.
- **Sentiment:** Polaridad de sentimiento del tweet.
- **Location:** Sector/barrio mencionado en el tweet.

Para crear la columna 'Location' se recorrió cada tweet perteneciente al archivo con los datos resultantes de la limpieza y transformación de datos. Los sectores/barrios de Quito se agruparon en un conjunto como lo presenta en el anexo 2 adjuntada en los Anexos, cada uno de estos se comprobó si está contenido dentro de cada tweet, en caso de existir en el tweet se ubicó el nombre de ese sector en la columna 'Location' en su respectiva fila (tweet). A continuación, se presentan los resultados encontrados:

Los tópicos obtenidos a partir del modelo *unigram* GSDMM se presentan en la tabla 15. La percepción sentimental de la ciudadanía con respecto a los tópicos obtenidos se presenta en la tabla 16 y se visualiza en la figura 8 en la que se muestra la cantidad de tweets distribuidos por tópicos y sentimientos. Para tener un mejor criterio sobre los resultados, se calculó el porcentaje de cada sentimiento con respecto a su tópico, ya que la cantidad de tweets es variable entre tópicos. Los porcentajes se presentan en la tabla 17.

Tabla 15. Tópicos obtenidos

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

Tabla 16. Distribución de sentimientos por tópico

Tópico	Positivo	Neutral	Negativo	Total tópico
Falta de control de las autoridades	1079	3565	7136	**11780**
Actividades culturales	816	3171	509	**4496**
Corrupción en el Municipio	1921	5026	10110	**17057**
Contaminación auditiva	397	1660	1646	**3703**
Ayuda en servicios municipales	1061	2628	4457	**8146**
Atención en necesidades de los barrios	1575	4536	7807	**13918**
Seguridad en la ciudad	3282	3461	6711	**13454**
Total sentimiento	**10131**	**24047**	**38376**	**72554**

Tabla 17. Distribución de sentimientos por tópico en porcentajes

Tópico	Positivo	Neutral	Negativo	Total tópico
Falta de control de las autoridades	9,15959	30,26315	60,57724	**16,23618**
Actividades culturales	18,14946	70,52935	11,32117	**6,19676**
Corrupción en el Municipio	11,26223	29,46590	59,27185	**23,50938**
Contaminación auditiva	10,72103	44,82851	44,45044	**5,10378**
Ayuda en servicios municipales	13,02479	32,26123	54,71397	**11,22749**
Atención en necesidades de los barrios	11,31628	32,59088	56,09282	**19,18295**
Seguridad en la ciudad	24,39423	25,72469	49,88107	**18,54342**
Total sentimiento	**13,96339**	**33,14358**	**52,89301**	**100**

Los tópicos ‘Falta de control de las autoridades’ y ‘Corrupción en el Municipio’ son los tópicos que presentaron más opinión negativa por parte de la ciudadanía ya que el porcentaje de tweets con percepción negativa de ambos fue del 60% aproximadamente. Entre todos los tópicos, ‘Actividades culturales’ presentó el menor porcentaje de opinión negativa con solo el 11%, asimismo, es el tópico que mayor neutralidad presenta con aproximadamente el 70%. Igualmente, el tópico en el que prevaleció la opinión positiva fue ‘Seguridad en la ciudad’ con un 24% aproximadamente.

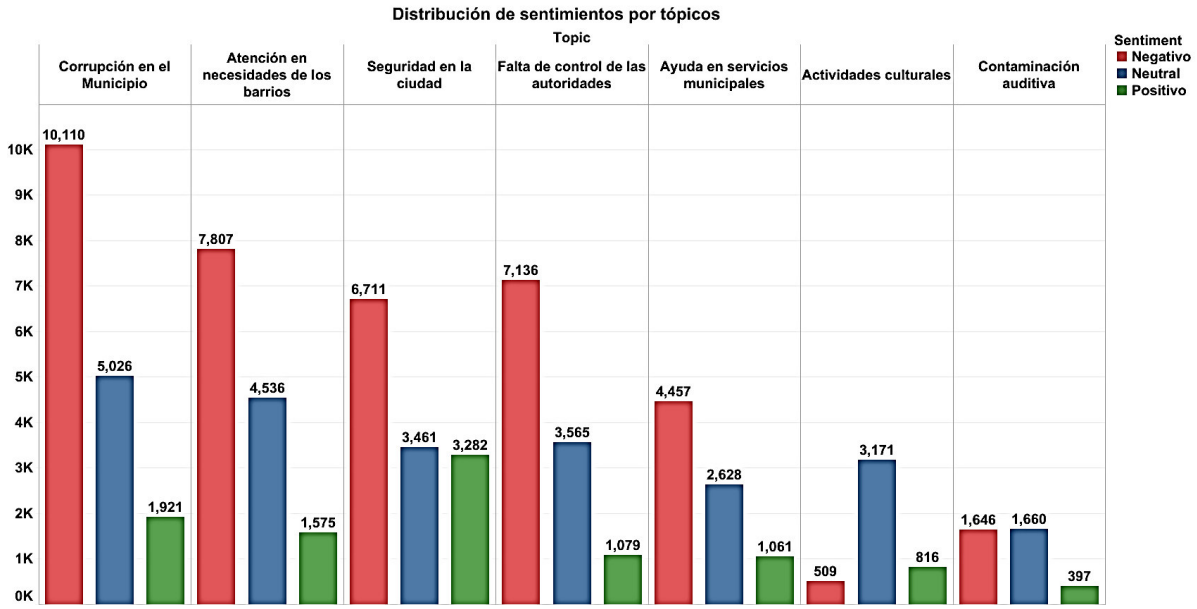


Figura 8. Gráfica de distribución de sentimientos por tópico

Por otra parte, el tópico que tuvo mayor cantidad de tweets fue ‘Corrupción en el Municipio’ con aproximadamente el 23% de todos los tweets recolectados, esto quiere decir 17057 tweets, y distribuido entre los tres sentimientos. Esto difiere del tópico ‘Contaminación auditiva’ y ‘Actividades culturales’ que solamente contaron el 5% y 6% de los tweets totales, es decir, 3703 y 4496 tweets respectivamente.

Debido a que los tópicos ‘Corrupción en el Municipio’ y ‘Atención en necesidades de los barrios’ abarcaron la mayor cantidad de tweets y al mismo tiempo los nombres de los tópicos no tuvieron mucha granularidad se decidió realizar el proceso nuevamente en esos tópicos por separado. Esto significa que se separó en un conjunto de datos los tweets que pertenecieron al tópico ‘Corrupción en el Municipio’ y en otro conjunto de datos los tweets del tópico ‘Atención en necesidades de los barrios’’.

El proceso fue el mismo que se describió anteriormente. Al realizar este proceso se desmenuzó al tópico 'Corrupción en el Municipio' en tres tópicos más específicos que fueron: 'Remoción del alcalde', 'Construcción del metro' y 'Corrupción del hijo del alcalde'.

Por otra parte, el tópico 'Atención en necesidades de los barrios' se dividió para formar los tres nuevos tópicos: 'Mal estado de las calles', 'Venta informal' y 'Problemas de basura en parques y agua'. El proceso para obtener la frase que define al tópico fue la misma que se utilizó en el proceso principal, es decir, ver la agrupación de las palabras más significativas y con ellas formar una frase que englobe al tópico. En la tabla 18 se presenta la nueva distribución sentimientos por tópicos y en la tabla 19 su respectivo porcentaje.

Tabla 18. Distribución de sentimientos por tópico con mayor granularidad

Tópico	Positivo	Neutral	Negativo	Total tópico
Falta de control de las autoridades	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**

En todos los nuevos tópicos encontrados el sentimiento negativo fue el que predominó, en cambio, el sentimiento positivo fue el que menor cantidad de tweets tuvo. Se mantuvo la tendencia ya que, en la mayoría de los tópicos también hubo predominio del sentimiento negativo. En la figura 9 se observa la nueva gráfica formada por los datos presentados en la tabla 18.

Tabla 19. Distribución de sentimientos por tópico con mayor granularidad en porcentajes

Tópico	Positivo	Neutral	Negativo	Total tópico
Falta de control de las autoridades	9,15959	30,26315	60,57724	**16,23618**
Actividades culturales	18,14946	70,52935	11,32117	**6,19676**
Construcción del metro	11,12047	33,44566	55,43386	**6,54409**
Corrupción del hijo del alcalde	12,23289	22,77310	64,99399	**11,48110**
Remoción del alcalde	9,39934	38,72832	51,87232	**5,48419**
Contaminación auditiva	10,72103	44,82851	44,45044	**5,10378**
Ayuda en servicios municipales	13,02479	32,26123	54,71397	**11,22749**
Mal estado de las calles	10,80606	38,41423	50,77970	**6,27532**
Venta informal	11,69885	29,67515	58,62598	**6,27946**
Problemas de basura en parques y agua	11,43688	29,83988	58,72322	**6,62816**
Seguridad en la ciudad	24,39423	25,72469	49,88107	**18,54342**
Total sentimiento	**13,96339**	**33,14358**	**52,89301**	**100**

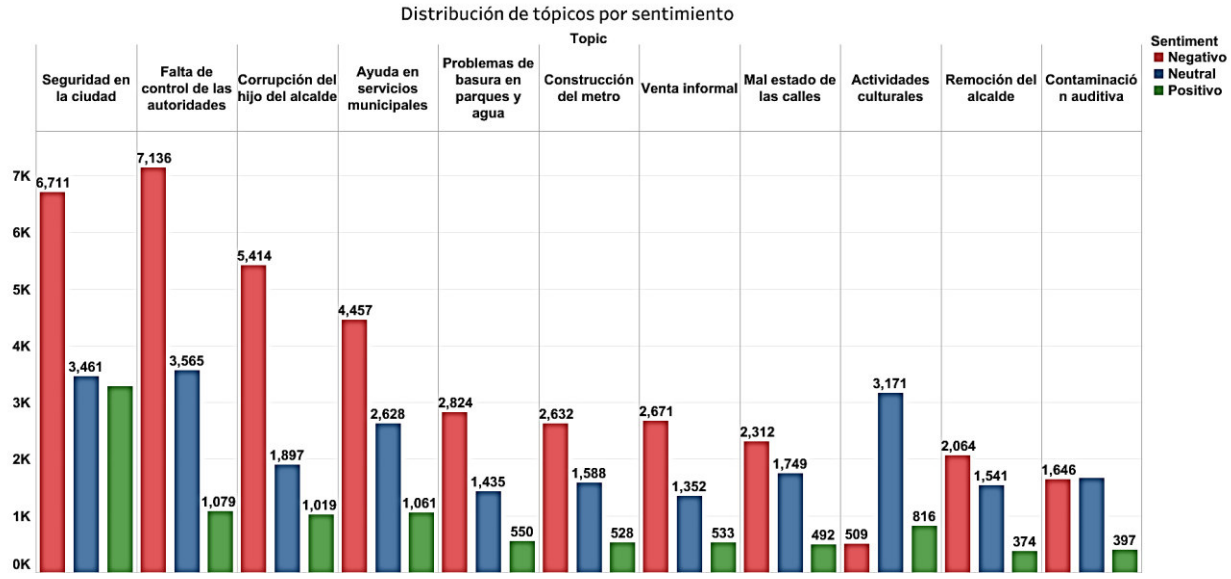


Figura 9. Gráfica de distribución de sentimientos por tópico con mayor granularidad

En consecuencia, del análisis de sentimiento que se llevó a cabo, se observó que existe una mayor prevalencia del sentimiento negativo en la ciudadanía ya que obtuvo el 52% de los tweets totales, esto se traduce en 38376 tweets. Al contrario, el sentimiento positivo solo tuvo el 13,96% de los tweets totales, es decir, 10131 tweets, y la neutralidad se vio reflejada en un 33,14% con 24047 tweets como se refleja en la figura 10.

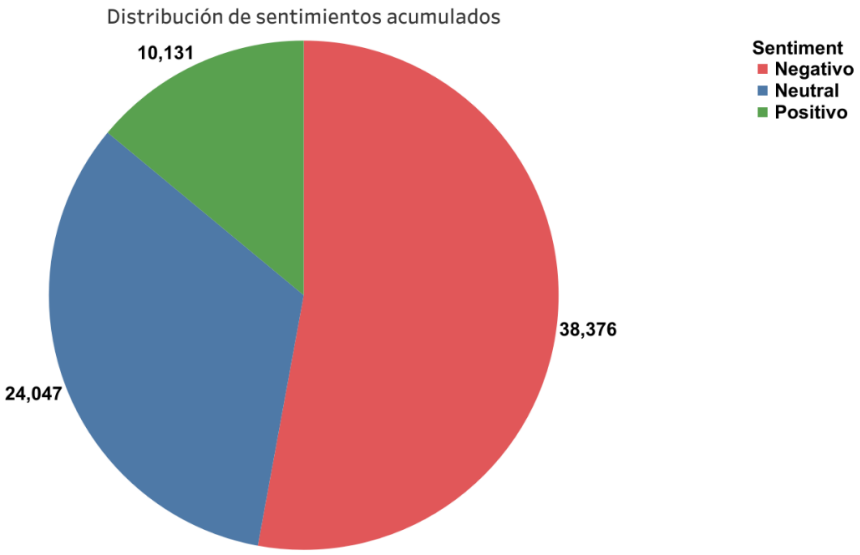


Figura 10. Distribución de sentimientos acumulados

Otro aspecto importante en este proyecto fue localizar los sectores de la ciudad en los que determinados tópicos tienen mayor incidencia. Por ende, se ubicaron los sectores más mencionados en los tweets, este proceso no consideró una distribución por tópicos sino, fue una clasificación únicamente por sectores generalizada. En la figura 11 se presenta los 25 sectores menciones en los tweets. Los sectores que más destacan son Tumbaco, Quitumbe, Conocoto, Las Casas, Cotocollao y San Roque.

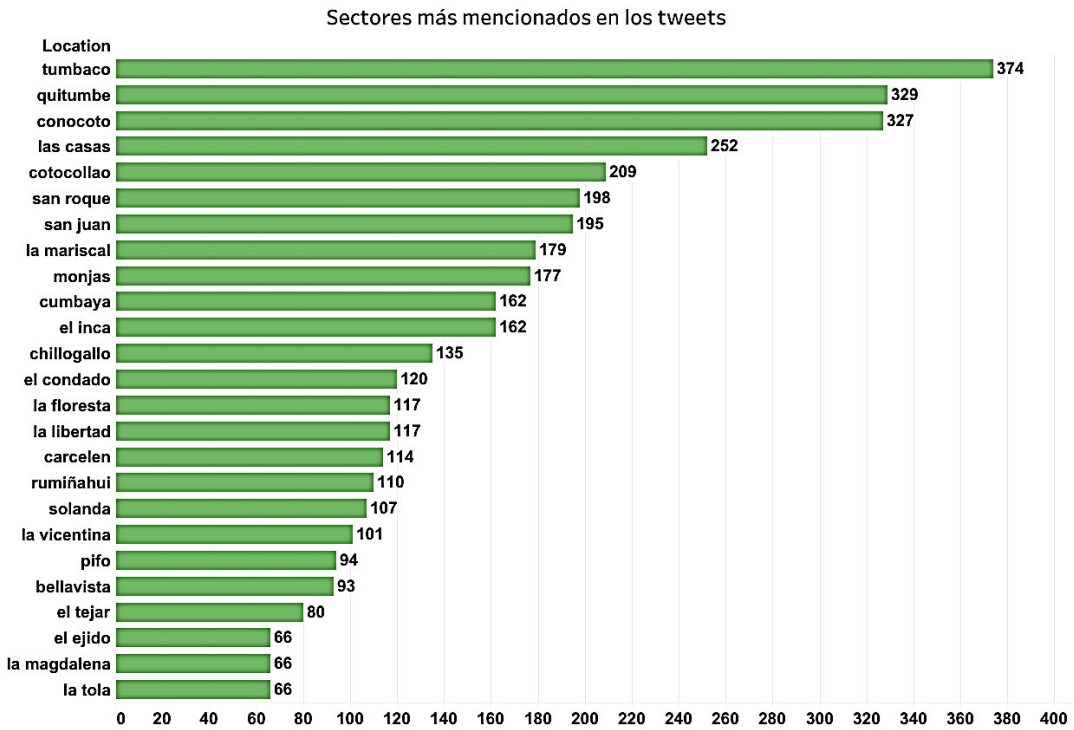


Figura 11. Sectores más mencionados en los tweets

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. En la figura 12 se presentan los resultados, el tópico Seguridad en la ciudad y Problemas de basura en parques y agua son los que presentan más sectores mencionados en sus tweets. Los sectores que más sobresalen son Quitumbe, San Roque, Monjas, Tumbaco y Las Casas.

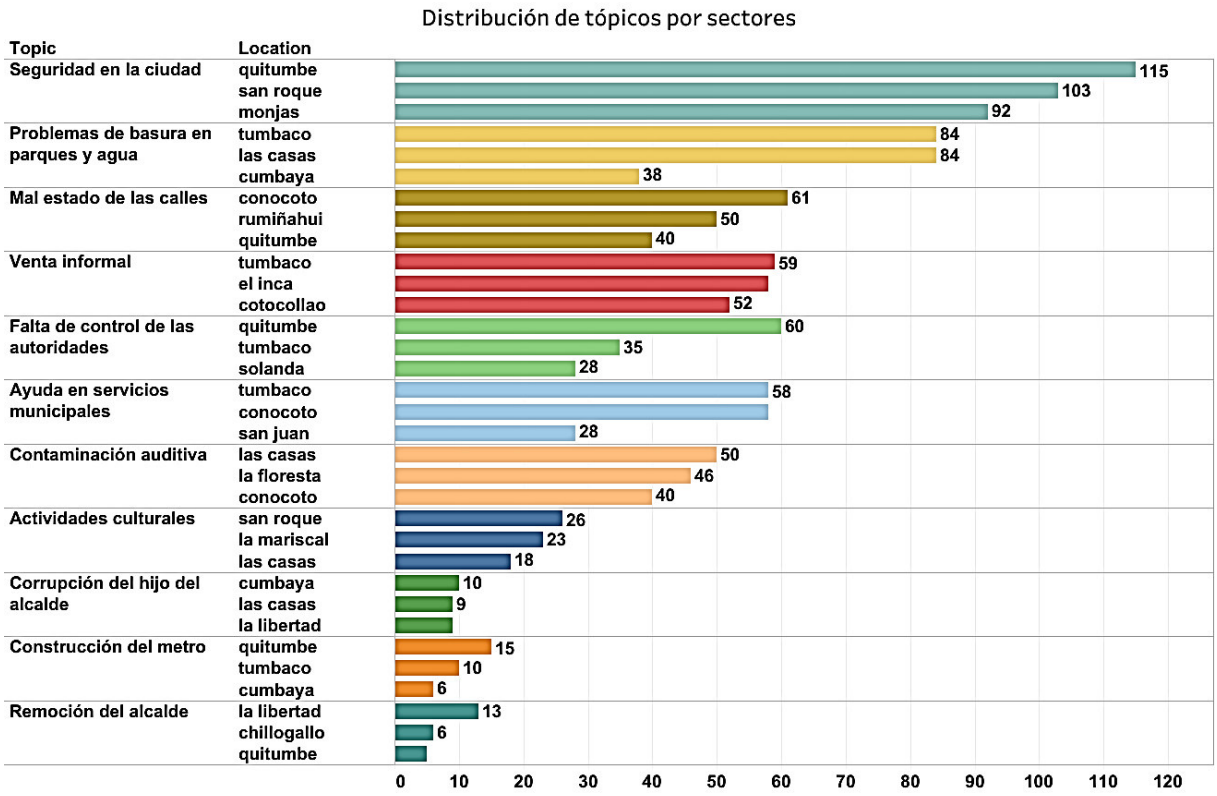


Figura 12. Distribución de tópicos por sectores más mencionados

El tablero de mandos (dashboard) que se creó en pasos anteriores se presenta en la figura 13 con datos más resumidos. En el gráfico Distribución de tópicos por sentimientos solo se presenta el negativo ya que, por lo visto en la figura 9 es el sentimiento predominante entre los tres. Por otra parte, se presenta los 12 primeros sectores más mencionados en los tweets en el gráfico Sectores más mencionados en tweets. Finalmente, en el gráfico Distribución de principales tópicos por sectores se presenta los 7 tópicos y sectores que fueron más mencionados.

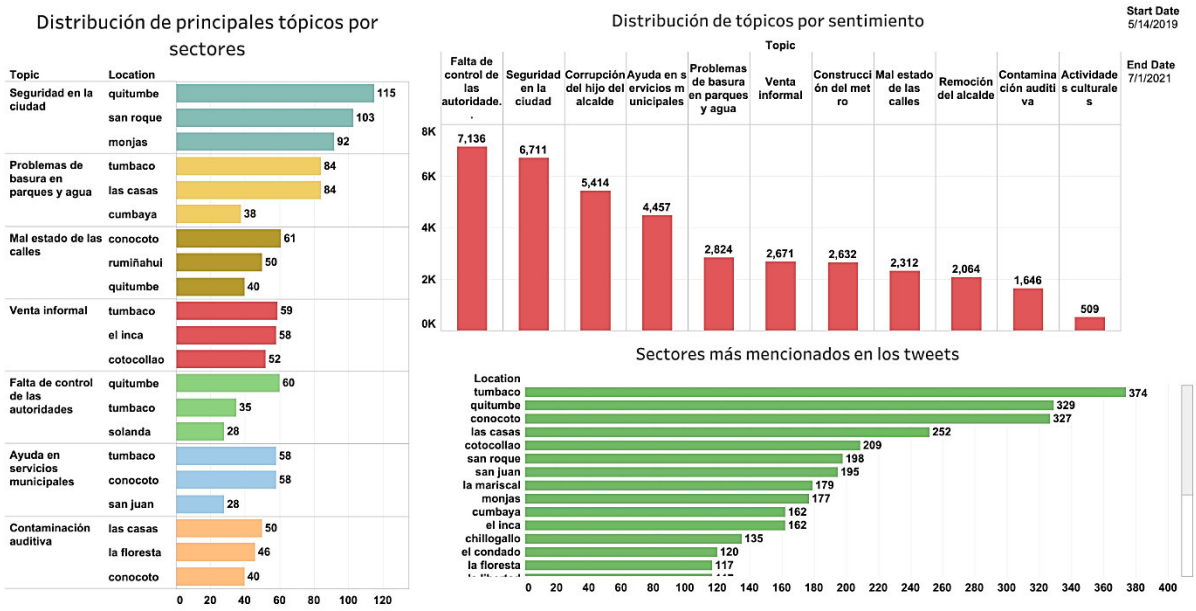


Figura 13. Tablero de mandos creado (Dashboard)

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.
- Tener mejor atención en servicios municipales como trámites y solicitudes.
- 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.

4. 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.

Gracias al análisis, clasificación y tratamiento de los tweets se filtró el ruido generado a partir de datos basura o repetidos, con este proceso se evidenció que la cantidad de tweets totales (72554) comparando con los tweets iniciales (246590) disminuyó en un 70.57% aproximadamente.

Por otra parte, el desarrollo del sistema de detección de necesidades se completó y abarcó tres procesos principales la extracción de tópicos, el análisis de sentimientos y el sistema de visualización, el principal bloqueo en este desarrollo fue la unicidad de los datos, ya que, al recolectar tweets diferentes a los ya recolectados, los tópicos cambiaron. Esto se solucionó con la creación de diferentes modelos entrenados en la etapa de pruebas, con esto se comprobó cada modelo con tweets relacionados a los datos con los que fue entrenado.

En los resultados obtenidos, al generalizar se deduce que la opinión negativa (38,37%) es la que más predomina en la ciudadanía quiteña. Por el contrario, la opinión positiva (10,13%) no tiene mucha incidencia ya que incluso está por debajo de la opinión neutral (24,04%). Por ende, bajo el análisis de este proyecto, los tópicos negativos de mayor influencia, es decir, Falta de control de las autoridades, 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.

5. RECOMENDACIONES

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.

Por otra parte, con base al desarrollo de este proyecto se constató que 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 ya que, se demoró la tercera parte del tiempo que GSDMM empleó para realizar el proceso, aunque en este proyecto la rapidez del algoritmo no fue crucial, para trabajos futuros que involucren una decisión entre estas dos técnicas se debe considerar la velocidad del algoritmo para tomar la decisión.

6. REFERENCIAS BIBLIOGRÁFICAS

Ahmed, K. B., Radenski, A., Bouhorma, M., & Ahmed, M. B. (2016). Sentiment Analysis for Smart Cities: State of the Art and Opportunities. *ICOMP*, *Internet Computing and Internet of Things,* 16, 55-61. http://worldcomp-proceedings.com/proc/p2016/ICM3084.pdf

Alkhammash, E. H., Jussila, J., Lytras, M. D., & Visvizi, A. (2019). Annotation of Smart Cities Twitter Micro-Contents for Enhanced Citizen's Engagement. *IEEE Access*, 7, 116267-116276. https://www.doi.org/10.1109/ACCESS.2019.2935186

Alotaibi, S., Mehmood, R., & Katib, I. (2019, June). Sentiment Analysis of Arabic Tweets in Smart Cities: A Review of Saudi Dialect. In *2019 Fourth International Conference on Fog and Mobile Edge Computing (FMEC)*, (pp. 330-335). IEEE. https://www.doi.org/10.1109/FMEC.2019.8795331

Alpaydin, E. (2014). *Introduction to Machine Learning (3ra ed).* London: The MIT Press. https://dl.matlabyar.com/siavash/ML/Book/Ethem%20Alpaydin-Introduction%20to%20Machine%20Learning-The%20MIT%20Press%20(2014).pdf

Anta, A. F., Chiroque, L. N., Morere, P., & Santos, A. (2013). Sentiment Analysis and Topic Detection of Spanish Tweets: A Comparative Study of NLP Techniques. *Procesamiento del Lenguaje Natural*, 50, 44-52. https://www.redalyc.org/pdf/5157/515751576005.pdf

Belyadi, H., & Haghighat, A. (2021). *Machine Learning Guide for Oil and Gas Using Python*. Gulf Professional Publishing. https://doi.org/10.1016/C2019-0-03617-5

Berrar, D. (2018). Bayes' Theorem and Naive Bayes Classifier. *Encyclopedia of Bioinformatics and Computational Biology*, 1, 403–412. https://doi.org/10.1016/B978-0-12-809633-8.20473-1

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

Cano, J. L. (2007). *Business Intelligence: Competir con Información.* Fundación Banesto. https://itemsweb.esade.edu/biblioteca/archivo/Business_Intelligence_competir_con_informacion.pdf

Chen, S., & Wang, Y. (s.f.). *Latent Dirichlet Allocation.* https://acsweb.ucsd.edu/~yuw176/report/lda.pdf

Edgar, T. W., & Manz, D. O. (2017). Chapter 4 - Exploratory Study. In T. W. Edgar, & D. O. Manz (Eds.), *Research Methods for Cyber Security*, 95-130, Syngress. https://doi.org/10.1016/B978-0-12-805349-2.00004-2

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters,* 27(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

Grus, J. (2015). *Data Science from scratch: First principles with python.* O'Reilly Media. https://www.m-fozouni.ir/wp-content/uploads/2020/08/Joel_Grus_Data_Science_from_Scratch_First_Princ.pdf

Hand, D. J., & Till, R. J. (2001). A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning*, 45, 171-186. https://link.springer.com/content/pdf/10.1023/A:1010920819831.pdf

Hansen, S., McMahon, M., & Prat, A. (2018). Transparency and Deliberation within the FOMC: A Computational Linguistics Approach. *The Quarterly Journal of Economics,* 133(2)*,* 801-870, https://doi.org/10.1093/qje/qjx045

Hollander, J. B., & Renski, H. (2015). Measuring Urban Attitudes Using Twitter: An Exploratory Study. Lincoln Institute of Land Policy. https://www.lincolninst.edu/sites/default/files/pubfiles/3607_2954_Hollander%20WP15JH1.pdf

IBM. (2011). IBM SPSS Modeler CRISP-DM Guide. IBM Corporation. https://inseaddataanalytics.github.io/INSEADAnalytics/CRISP_DM.pdf

ITahora. (2020). *Tendencias tecnológicas de mayor impacto en Ecuador para el 2020.* ITahora. https://www.itahora.com/wp-content/uploads/2020/04/TendenciasTecnologicas2020_EY_ITAhora_FINAL.pdf

ITahora. (2021). *Resultados Tendencias y Prioridades IT 2021.* ITahora. https://itahora.com/wp-content/uploads/2021/02/RESULTADOS-TENDENCIAS-Y-PRIORIDADES-IT-2021-ECUADOR.pdf

Jacobs, M., & Kaim, E. (14 de Mayo de 2021). *What is Scrum*. Microsoft. https://docs.microsoft.com/en-us/devops/plan/what-is-scrum

Jelodar, H., Wang, Y., Yuan, C., & Feng, X. (2017). *Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey.* https://arxiv.org/pdf/1711.04305.pdf

Jurafsky, D., & Martin, J. H. (2020). *Speech and Language Processing (3ra ed).* Stanford: https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

Katrekar, A. (s.f.). *An Introduction to Sentiment Analysis.* GlobalLogic Inc. https://www.globallogic.com/se/wp-content/uploads/2019/12/Introduction-to-Sentiment-Analysis.pdf

KDnuggets. (Octubre, 2014). What main methodology are you using for your analytics, data mining, or data science projects? Poll. https://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html

Kibble, R. (2013). *Introduction to natural language processing.* University of London. https://london.ac.uk/sites/default/files/study-guides/introduction-to-natural-language-processing.pdf

Kiritchenko, S., Zhu, X., & Mohammad, S. (2014). Sentiment Analysis of Short Informat Texts. *Journal of Artificial Intelligence Research, 50*, 723-762. https://doi.org/10.1613/jair.4272

Learn, S. (s.f.). Scikit learn. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html

Liu, Z., Huang, W., Zheng, Y., & Sun, M. (2010, October). Automatic Keyphrase Extraction via Topic Decomposition. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 366-376).

Ludeling, A., & Kyto, M. (2008). Development of tag sets for part-of-speech tagging. *Corpus Linguistics: An International Handbook*, 1, 501–526. https://eprints.whiterose.ac.uk/81781/1/DevelopmentTagSetPOSTagging.pdf

Luo, T., Chen, S., Xu, G., & Zhou, J. (2013). *Trust-based Collective View Prediction*. Springer New York. https://doi.org/10.1007/978-1-4614-7202-5

Malafosse, C. (2019). Github. https://github.com/charlesmalafosse/open-dataset-for-sentiment-analysis

Marbán, O., Mariscal, G., & Segovia, J. (2009). *Data Mining and Knowledge Discovery in Real Life Applications.* IntechOpen. https://doi.org/10.5772/6438

Mazarura, J., & Waal, A. d. (2016). A comparison of the performance of latent Dirichlet allocation and the Dirichlet multinomial mixture model on short text. In *2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*, (pp. 1-6). IEEE.

McClish, D. K. (1989). Analyzing a portion of the ROC curve. *Medical decision making : an international journal of the Society for Medical Decision Making,* 9(3), 190-195. https://doi.org/10.1177/0272989X8900900307

Mifrah, S., & Benlahmar, E. H. (2020). Topic Modeling Coherence: A Comparative Study between LDA and NMF Models using COVID'19 Corpus. *International Journal of Advanced Trends in Computer Science and Engineering*, 9(4), 5756-5761. https://doi.org/10.30534/ijatcse/2020/231942020

Moine, J., Haedo, A., & Gordillo, S. (2011). Estudio comparativo de metodologías para minería de datos. *XIII Workshop de Investigadores en Ciencias de la Computación, 278-281*, 11. http://sedici.unlp.edu.ar/bitstream/handle/10915/20034/Documento_completo.pdf?sequence=1&isAllowed=y

Murthy, D., Gross, A., & Pensavalle, A. (2016). Urban Social Media Demographics: An Exploration of Twitter Use in Major American Cities. *Journal of Computer-Mediated Communication*, 21(1), 33-49. https://doi.org/10.1111/jcc4.12144

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5), 554-551. https://doi.org/10.1136/amiajnl-2011-000464

Patil, P., & Yalagi, P. (2016). Sentiment Analysis Levels and Techniques: A survey. *International Journal of Innovations in Engineering and Technology (IJIET)*, 6(4), 523-528. http://ijiet.com/wp-content/uploads/2016/05/72.pdf

Pedrycz, W., & Chen, S.-M. (2016). *Sentiment Analysis and Ontology Engineering.* Cham Springer. https://link.springer.com/book/10.1007/978-3-319-30319-2

Prabowo, R., & Thelwall, M. (2009). Sentiment Analysis: A Combined Approach. *Journal of Infometrics, 3(2),* 143-157. https://doi.org/10.1016/j.joi.2009.01.003

Rosner, F., Hinneburg, A., Roder, M., Nettling, M., & Both, A. (2014). Evaluating topic coherence measures. arXiv preprint. http://arxiv.org/abs/1403.639

Saunders, A. (2017). *La era de la Perplejidad: Repensar el mundo que conocíamos*. Penguin Random House Grupo Editorial. https://www.bbvaopenmind.com/wp-content/uploads/2018/01/BBVA-OpenMind-La-era-de-la-perplejidad-repensar-el-mundo-que-conociamos.pdf

Schwaber, K. & Sutherland, J. (2020). The Scrum Guide. https://scrumguides.org/docs/scrumguide/v2020/2020-Scrum-Guide-US.pdf#zoom=100

SEPLN. (2017). *TASS: Workshop on Semantic Analysis at SEPLN*. http://tass.sepln.org/tass_data/download.php?auth=Ft0aSs2sV4peVv2eor9

Srivastava, D. K., & Bhambhu, L. (2009). Data Classification Using Support Vector Machine. *Journal of Theoretical and Applied Information Technology*. http://www.jatit.org/volumes/research-papers/Vol12No1/1Vol12No1.pdf

Subasi, A. (2020). *Practical Machine Learning for Data Analysis Using Python*. Academic Press. https://doi.org/10.1016/C2019-0-03019-1

Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making, 19(1), 281*. https://doi.org/10.1186/s12911-019-1004-8

Umachandran, K., Jurčić, I., Corte, V., & Ferdinand-James, D. (2019). Industry 4.0.: The new industrial revolution. *Big Data Analytics for Smart and Connected Cities, 138-156.* https://doi.org/10.4018/978-1-5225-6207-8.ch006

Wang, Z., Bai, G., Chowdhury, S., Xu, Q., & Seow, Z. L. (2017). TwiInsight: Discovering Topics and Sentiments from Social Media Datasets. arXiv preprint. http://arxiv.org/abs/1705.08094

Yin, J., & Wang, J. (2014, August). A Dirichlet Multinomial Mixture Model-based Approach for Short Text Clustering. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* (pp. 233-242). https://doi.org/10.1145/2623330.2623715

Zavattaro, S. M., French, P. E., & Mohanty, S. D. (2015). A sentiment analysis of U.S. local government tweets: The connection between tone and citizen involvement. *Government Information Quarterly*, 32, 333-341. https://doi.org/10.1016/j.giq.2015.03.003

7. ANEXOS

Anexo I. Se presentan las cuentas de Twitter relacionadas con el Municipio de Quito de las cuales se extrajeron los tweets recolectados.

INSTITUCIÓN	CUENTA DE TWITTER
Municipio de Quito	@MunicipioQuito
Empresa de Logística para la Seguridad de Quito	@epemseguridad
Zonal Manuela Sáenz	@ZonalmSaenz
Zonal Calderón	@Zonalcalderon
Hábitat y Vivienda	@HabitatUio
Urbanimal Quito	@urbanimalquito
Fundación Museos	@Museos_Quito
Secretaría de Desarrollo Productivo	@DesarrolloQuito
Museo Carmen Alto	@MuseoCarmenAlto
Empresa de Gestión de Residuos Sólidos Quito	@EMGIRSEP
Agentes de Control Quito	@agentesdequito
OMSC Quito	@OMSCQuito
Metro de Quito	@MetrodeQuito
CONQUITO	@conquitouio
Museo Alberto Mena Caamaño	@museodeceraAMC
Casa de las Artes La Ronda	@casalaronda
Centro Cultural Benjamín Carrión	@casacarrion
Relaciones Internacionales Quito	@QuitoGlobal
Cultural Itchimbia	@itchimbia_uio
Centro Cultural Cumandá	@quitocumanda
Teatro Capitol Quito	@capitolquito

Consejo Metropolitano de Responsabilidad Social Quito	@RSQuito
Centro de Arte Contemporáneo	@CentroArteQ
YAKU	@yakuquito
Centro Cultural Metropolitano	@CentroCulturalQ
Museo Interactivo de Ciencia	@MICmuseo
Teatro Sucre	@TeatroSucreQ
Quito Honesto	@quitohonesto
Museo de la Ciudad	@MuseoCiudadUIO
Zonal Tumbaco	@zonaltumbaco
Secretaría de Territorio, Hábitat y Vivienda Quito	@territorio_uio
Concejo de Quito	@ConcejoQuito
Secretaría de Coordinación Territorial	@zonalesquito
Comercio Quito	@QuitoComercio
aQUITOdos	@aQUITOdosEC
AMC Quito	@amcquito
COE Quito	@coequito
Quito Informa	@quitoinforma
Secretaría de Movilidad del Municipio de Quito	@movilidadquito_
Pacha FM 102.9	@pachafmquito
Empresa de Pasajeros Quito	@TransporteQuito
AMT Quito	@AMTQuito
Empresa Aseo Quito	@EmAseoQuito
Ambiente Quito	@ambientequito
Secretaría de Seguridad y Gobernabilidad Quito	@SeguridadeQuito
Bomberos Quito	@BomberosQuito
Epmaps – Agua de Quito	@aguadequito
Jorge Yunda Machado	@LoroHomero
Obras Quito	@ObrasQuito
Secretaría de Cultura Quito	@culturaquito
Zonal La Delicia	@zona_ladelicia
Zonal Quitumbe	@zonalquitumbe
Registro de la Propiedad	@RegistroQuito
Zonal Los Chillos	@zona_loschillos
Zonal La Mariscal	@zonalamariscal
Zonal Eugenio Espejo	@ZonalEspejo

Zonal Eloy Alfaro	@zonaeloyalfaro
Secretaría de Educación, Recreación y Deporte	@EducacionQuito
Secretaría de Salud Quito	@saludquito
Quito Turismo	@EPMQuitoTurismo
IMP Patrimonio Quito	@PatrimonioQuito
Empresa de Servicios Aeroportuarios Quito	@ServiAeroQuito
Visit Quito	@VisitQuito
Secretaría de Inclusión de Quito	@InclusionQuito
Patronato San José	@PatronatoSJ

Anexo II. Se presenta la lista de barrios de la ciudad de Quito.

5 esquinas	Alangasí	Atucucho	Bellavista	Carcelén
Caupichu	Centro Histórico	Chilibulo	Chillogallo	Chimbacalle
Ciudadela del Ejército	Ciudadela Ibarra	Comité del Pueblo	Conocoto	Cornejo
Cotocollao	Cumbayá	El Batán	El Beaterio	El Calzado
El Camal	El Condado	El Dorado	El Ejido	El Inca
El Panecillo	El Pintado	El Tejar	El Troje	Guajalo
Guamaní	Guápulo	Iñaquito	Kennedy	La Argelia
La Bota	La Ecuatoriana	La Ferroviaria	La Floresta	La Florida
La Forestal	La González Suárez	La Guaragua	La Libertad	La Loma Grande
La Magdalena	La Marín	La Mariscal	La Mena	La Ronda
La Tola	La Vicentina	La Victoria	Las Casas	Lucha de los Pobres
Luluncoto	Manuelita Saenz	Mena de Hierro	Miraflores	Monjas
Nueva Aurora	Oriente Quiteño	Pifo	Ponceano	Puembo
Puengasí	Quito Norte	Quito Sur	Quito Tennis	Quitumbe
Reino de Quito	Rumiñahui	San Roque	San Bartolo	San Carlos
San Diego	San Juan	San Marcos	San Martin	San Rafael
Santa Rita	Solanda	Tababela	Toctiuco	Tumbaco
Turubamba	Villaflora			