ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE SISTEMAS

UNIDAD DE TITULACIÓN

DESARROLLO DE UN PROTOTIPO DE SISTEMA PARA EL ANÁLISIS DE OPINIONES BASADO EN TWEETS. CASO DE ESTUDIO: METRO DE QUITO

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

Jorge Luis Quilumba Toaquiza

jl_quilumba@hotmail.com

Kevin Joel Villacis Navarrete

k.joel2@hotmail.com

DIRECTORA: Hallo María PhD.

maria.hallo@epn.edu.ec

Quito, 2023

APROBACIÓN DEL DIRECTOR

Como director del presente trabajo DESARROLLO DE UN PROTOTIPO DE SISTEMA PARA EL ANÁLISIS DE OPINIONES BASADO EN TWEETS. CASO DE ESTUDIO: METRO DE QUITO que fue desarrollado por Quilumba Toaquiza Jorge Luis y Villacis Navarrete Kevin Joel, estudiantes de la carrera Ingeniería en Sistemas Informáticos y de Computación, habiendo supervisado la realización de este trabajo y realizado las correcciones correspondientes, doy por aprobada la redacción final del documento escrito para que prosiga con los trámites correspondientes a la sustentación de la defensa oral.

Hallo María PhD.

DIRECTORA

DECLARACIÓN DE AUDITORÍA

Nosotros, QUILUMBA TOAQUIZA JORGE LUIS y VILLACIS NAVARRETE KEVIN JOEL, declaramos bajo juramento que el trabajo aquí descrito es de nuestra 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.

JORGE LUIS QUILUMBA TOAQUIZA
Firmado digitalmente por JORGE LUIS QUILUMBA TOAQUIZA
Fecha: 2023.06.27 15:00:26 -05'00'

QUILUMBA TOAQUIZA JORGE LUIS

Firmado electrónicamente por:
KEVIN JOEL VILLACIS NAVARRETE
Razón:
Localización:
Fecha: 2023-06-28T09:01:21.077629-05:00

VILLACIS NAVARRETE KEVIN JOEL

DEDICATORIA

Este trabajo de grado se lo dedico a mis padres Luis Quilumba y Luz Toaquiza que con su amor y trabajo me educaron y apoyaron incondicionalmente en todo el proceso de formarme como profesional.

Jorge Luis Quilumba Toaquiza

Dedico este escalón académico a mi hermana Kary, confidente y amiga. Con tu amor y guía me has impulsado a creer que lo imposible se puede hacer realidad y que los sueños se van materializando con esfuerzo.

Joel Villacis Navarrete

AGRADECIMIENTO

A mi familia y amigos por estar a mi lado en cada paso de esta emocionante aventura. Hicieron que este camino se sienta más ligero, más llevadero. Madre, gracias por siempre creer en mí, por apoyarme en este camino. Padre, miraré al cielo y agradeceré por tu guía y enseñanzas de vida. Agradezco a mis hermanas y sobrinos por su infinito amor. Gracias Roxy por caminar de la mano y compartiendo un amor bonito. Agradezco a mis amigos, la familia que escogí por salvarme del caos interno y externo a Cesar, Vivito, Kevo, Violeta, Daniel, Pancho, Miguel, Andrés gracias por estar en esta etapa de mí vida. A mis hippiosas Nicole y Cami. Gracias Jorge por esta amistad por el compañerismo y profesionalismo en la elaboración de la tesis, por más proyectos juntos. Finalmente, a mi amigo de 4 patas Ghost, gracias por escogerme, por tu lealtad y compañía en las noches largas de estudio.

Joel Villacis Navarrete

A mis padres y hermanos Fausto, Fernanda y Mayra quienes me mostraron el camino hacia la superación, además de saber que mis logros también son los suyos. A mis queridos sobrinos Jordan, Aarón, Keyla, Jeremy y Daniela gracias por llenar mi vida de risas, abrazos y alegría. A mi primo Javi por su apoyo incondicional en todas las circunstancias. A Karen una persona muy especial en mi vida, gracias por tu apoyo incondicional en la carrera y en todas mis aspiraciones. A Joel, gracias por tu amistad, al igual que los consejos y los buenos momentos compartidos. A mi compañera de 4 patas Sheyka que siempre ha estado ahí para brindarme su cariño y su incondicionalidad. A mis compañeros por las maravillosas experiencias dentro y fuera de las aulas y a mis maestros por las enseñanzas impartidas de sus experiencias profesionales.

Jorge Luis Quilumba Toaquiza

ÍNDICE DE CONTENIDO

ÍNDICE DE FIGURAS viii
ÍNDICE DE TABLAS ix
RESUMEN xi
ABSTRACT xii
1 INTRODUCCIÓN 1
1.1. Descripción del problema 1
1.2. Propuesta 3
1.3. Objetivo general 4
1.4. Objetivos específicos 4
1.5. Trabajos relacionados 4
1.6. Marco teórico 6
1.6.1. Minería de datos 6
1.6.2. Procesamiento del Lenguaje Natural (NPL) 6
1.6.3. Análisis de sentimientos 6
1.6.4. Preprocesamiento de datos 7
1.6.5. Algoritmo de aprendizaje supervisado – clasificación 8
1.6.6. Extracción de tópicos 8
2 METODOLOGÍA 9
2.1. Comprensión del negocio 10
2.2. Comprensión de los datos 11
2.3. Preparación de los datos 11
2.3.1. Limpieza de datos 12
2.3.2. Estructuración de los datos 13
2.4. Modelamiento 15
2.4.1. Algoritmo de modelado de tópicos 15
2.4.2. Algoritmo de clasificación 16
2.5. Evaluación 17
2.5.1. Análisis de tópicos sobre el metro de Quito en la alcaldía de Augusto Barrera 17
2.5.2. Análisis de tópicos sobre el metro de Quito en la alcaldía de Mauricio Rodas 18
2.5.3. Análisis de tópicos sobre el metro de Quito en la alcaldía de Jorge Yunda 18
2.5.4. Análisis de tópicos sobre el metro de Quito en la alcaldía de Santiago Guarderas 19

2.6. Despliegue 20
3 DESARROLLO DEL PROTOTIPO 20
3.1. Análisis 20
3.1.1. Requerimientos 20
3.1.2. Otros requerimientos 24
3.2. Diseño 26
3.2.1. Diagrama de clases 26
3.2.2. Diagrama de casos de uso 27
3.2.3. Diagrama de secuencia 28
3.2.4. Estándares de diseño y desarrollo 29
3.3. Implementación 30
3.3.1. Pantalla inicio 30
3.3.2. Pantalla dashboard 30
3.4. Pruebas 34
3.4.1. Pruebas de rendimiento 34
3.4.2. Pruebas de usabilidad 37
3.5. Mantenimiento 38
4 RESULTADOS 39
4.1. Conclusiones 42
4.2. Recomendaciones 42
REFERENCIAS BIBLIOGRÁFICAS 44
ANEXOS 47

ÍNDICE DE FIGURAS

Figura 1. Encuesta realizada por Kdnuggets. 9
Figura 2. Niveles de CRISP-DM. 10
Figura 3. Ciclo de vida del proyecto. 10
Figura 4: Mockup de interfaz gráfica de la pantalla inicio. 25
Figura 5: Mockup de interfaz gráfica del dashboard. 25
Figura 6: Arquitectura del prototipo de sistema. 26
Figura 7: Diagrama de clases del dominio del problema. 27
Figura 8: Diagrama de casos de uso general. 27
Figura 9: Diagrama de casos de uso generación de resultados (dashboard). 28
Figura 10: Diagrama de secuencia del dominio del problema. 28
Figura 11: Diagrama de secuencia generación de resultados (dashboard). 29
Figura 12: Interfaz Web de línea de tiempo e indicadores. 30
Figura 13. Dashboard del análisis de opiniones sobre el metro de Quito en la alcaldía de Augusto Barrera. 31
Figura 14. Dashboard del análisis de opiniones sobre el metro de Quito en la alcaldía de Mauricio Rodas. 32
Figura 15. Dashboard del análisis de opiniones sobre el metro de Quito en la alcaldía de Jorge Yunda. 33
Figura 16. Dashboard del análisis de opiniones sobre el metro de Quito en la alcaldía de Santiago Guarderas. 34
Figura 17: Resultado de pruebas de rendimiento 35
Figura 18: Dimensión eficacia. 36
Figura 19: Tiempo de respuesta. 36
Figura 20: Uso de recursos. 37
Figura 21: Resultados de encuesta prueba de usabilidad. 38
Figura 22: Resultados encuesta de requerimiento de diseño 47
Figura 23: Reporte resumen. 48
Figura 24: Reporte medición eficacia. 49
Figura 25: Medición del tiempo de respuesta 50
Figura 26: Medición del tiempo de respuesta – gráfico de líneas. 51
Figura 27: Promedio de uso de memoria RAM. 52
Figura 28: Promedio de uso de CPU. 53
Figura 29: Promedio de uso de disco. 53

ÍNDICE DE TABLAS

Tabla 1. Cantidad de artículos obtenidos por sitios web. *4*
Tabla 2. Dataset inicial – datos crudos. *12*
Tabla 3. Dataset final. *13*
Tabla 4. Dataset entrenado. *14*
Tabla 5. Dataset con datos limpios. *14*
Tabla 6. Dataset de tweets del metro de Quito segmentado por alcaldías. *15*
Tabla 7. Comparación coherence score. *16*
Tabla 8. Score de algoritmos de clasificación. *17*
Tabla 9. Resultados del análisis de tópicos sobre el metro de Quito en la alcaldía de Augusto Barrera. *18*
Tabla 10. Resultados del análisis de tópicos sobre el metro de Quito en la alcaldía de Mauricio Rodas. *18*
Tabla 11. Resultados del análisis de tópicos sobre el metro de Quito en la alcaldía de Jorge Yunda. *19*
Tabla 12. Resultados del análisis de tópicos sobre el metro de Quito en la alcaldía de Santiago Guarderas. *19*
Tabla 13. Requerimiento visualización de línea de tiempo de las alcaldías. *21*
Tabla 14. Requerimiento visualización de imagen de los alcaldes. *21*
Tabla 15. Requerimiento visualización de nombres de los alcaldes. *21*
Tabla 16. Requerimiento visualización del periodo de la alcaldía. *21*
Tabla 17. Requerimiento visualización de tweets totales por alcaldías. *22*
Tabla 18. Requerimiento visualización de tweets positivos por alcaldía. *22*
Tabla 19. Requerimiento visualización de tweets negativos por alcaldía. *22*
Tabla 20. Requerimiento visualización de tópicos con su polaridad por alcaldía. *22*
Tabla 21. Requerimiento visualización de nube de palabras por alcaldía. *22*
Tabla 22. Requerimiento visualización de tendencias por alcaldía. *23*
Tabla 23. Requerimiento visualización de tweets según la fecha de publicación. *23*
Tabla 24. Requerimiento escalabilidad. *23*
Tabla 25. Requerimiento mantenibilidad. *24*
Tabla 26. Requerimiento confiabilidad. *24*
Tabla 27. Pruebas de rendimiento. *35*
Tabla 28. Dimensión eficacia. *35*
Tabla 29. Uso de recursos. *37*
Tabla 30. Resultado del análisis de opiniones sobre el metro de Quito por cada alcaldía. *39*
Tabla 31. Tendencias del análisis de opiniones sobre el metro de Quito en la alcaldía de Augusto Barrera. *41*
Tabla 32. Tendencias del análisis de opiniones sobre el metro de Quito en la alcaldía de Mauricio Rodas. *41*
Tabla 33. Tendencias del análisis de opiniones sobre el metro de Quito en la alcaldía de Jorge Yunda. *41*
Tabla 34. Tendencias del análisis de opiniones sobre el metro de Quito en la alcaldía de Santiago Guarderas. *41*
Tabla 35. Características del servidor *52*

GLOSARIO DE TÉRMINOS

LDA (Asignación de Dirichlet Latente). - Técnica para extraer y descubrir temas subyacentes en un conjunto de documentos asignando probabilidades a las palabras basadas en temas.

GSDMM (Mezcla de Muestreo Gibbs Dirichlet Multinomial). - Algoritmo eficiente para el clustering de documentos completos.

SVM (Máquinas de vectores de soporte). - Algoritmo de aprendizaje supervisado utilizado para la regresión y clasificación de texto, análisis de sentimiento, reconocimiento de imágenes, entre otros.

Storytelling. - Habilidad de transmitir información, emociones y valores a través de diversos ambientes como canvas, dashborad, entre otros.

Prototipo. - Es una versión preliminar del software creada con el fin de probar y mejorar el diseño antes de la implementación.

Lematización. - Proceso lingüístico para reducir una palabra a su forma base o raíz. Ejemplo: construyendo → construir.

Web scraping. - Técnica para extraer datos de páginas web a través de la estructura del HTML.

Part of speech. - Técnica de clasificación de palabras según su función y sus características gramaticales (verbos, adjetivos, preposiciones, sustantivos, pronombres, adverbios, conjuciones.) en un idioma especifico.

Dashboard. - Interfaz visual que presenta información resumida y accesible sobre datos clave o métricas.

KPI (indicador de clave de rendimiento). - Métricas que proporcionan información cuantificable y objetiva para el monitoreo de desempeño y toma de decisiones.

Algoritmo. – Es una secuencia de pasos puntuales para resolver un problema o realizar una tarea específica.

Front-end. – Es la parte visible y accesible de una aplicación con la que interactúan los usuarios.

Back-end. – Es la parte no visible de una aplicación encargada de la estructura y funcionamiento del sistema.

RESUMEN

Twitter es una red social cuyo enfoque es compartir información inmediata de distintos temas y posturas sobre asuntos políticos, sociales, problemáticas nacionales, entre otros. En ese sentido, el presente proyecto estuvo enfocado en realizar un análisis de opiniones sobre tweets relacionados al metro de Quito con base en la metodología CRISP-DM, que consistió en recopilar varios tweets y procesar los datos para garantizar la calidad de la información. Posteriormente, fue aplicado el modelo GSDMM para la generación de tópicos e identificación de los temas dominantes; esta fase implicó un previo análisis entre el GSDMM y el LDA, lo que permitió determinar que el primero genera mejores resultados en cuanto al manejo de texto no estructurado. Finalmente, se entrenó un modelo de aprendizaje supervisado, SVM, con datos en español, debido a que tuvo mejor score frente algoritmos como *decision tree, naive bajes* y *logistic regression*; de este modo, fue posible efectuar el análisis de sentimientos y determinar así la polaridad de los tweets.

Para presentar los resultados de manera visual, se elaboró un dashboard como prototipo de desarrollo web mediante un framework, como Flask y Dash, y con Python como lenguaje de programación. También se utilizó Pandas para estructurar los datos y Plotly para la creación de los gráficos a mostrarse en el dashboard. Así mismo, el análisis tiene fundamentos de *visual storytelling* que permite mostrar fácilmente los temas más relevantes sobre el avance de la obra, presupuesto, administración del metro, algunos inconvenientes en la construcción, entre otros, lo que da paso a identificar el punto de vista de los usuarios y la polaridad de sentimientos: positivo y negativo. Cabe indicar que el dashboard está dividido en función de los diferentes periodos de las alcaldías de Quito entre 2015 y 2023.

Palabras Clave: *Data Mining*, Preprocesamiento de Datos, Algoritmo GSDMM, Aprendizaje supervisado SVM, *Visual Storytelling*, *Flask*, *Plotly*, *Dash*, *Pandas*.

ABSTRACT

Twitter is a social media platform designed to share immediate information about various topics, including political, social, and national issues. This project aimed to analyze opinions on Quito's subway through the CRISP-DM methodology, which involved collecting tweets and processing the data to ensure information quality. The GSDMM model was then used for topic generation and identification of dominant topics, and a supervised learning model, SVM, was trained with Spanish data for sentiment analysis to determine tweet polarity.

A dashboard was developed as a prototype for web development using Flask, Dash, and Python as the programming language. Pandas was used to structure the data, and Plotly was used to create graphics to be displayed on the dashboard. The analysis uses visual storytelling to easily display the most relevant topics, such as progress, budget, subway administration, and construction issues, to identify user viewpoints and polarity of feelings (positive and negative). The dashboard is divided according to the different periods of Quito's mayoralties between 2015 and 2023.

Keywords: Data Mining, Data Preprocessing, GSDMM Algorithm, Supervised Learning SVM, Visual Storytelling, Flask, Plotly, Dash, Pandas.

1 INTRODUCCIÓN

1.1. Descripción del problema

La expansión territorial que se ha experimentado en Quito a lo largo de los años, especialmente en la última década, ha ocasionado que la movilidad de los quiteños se vea afectada tanto si eligen el transporte público o privado, lo que a su vez incide negativamente en la calidad de vida. La problemática que se observa en el día a día al momento de movilizarse es la presencia de embotellamiento en varios puntos en la ciudad, sobre todo en horas pico, tal como lo exponen los resultados de la Encuesta Domiciliaria de Movilidad del Distrito Metropolitano de Quito (EDM11) [1], en la que además se define que los horarios de mayor afluencia de tránsito son entre las 6h00 a 7h00, de 12h00 a 13h00 y de 18h00 a 19h00.

El panorama es bastante desalentador, las autoridades municipales han expuesto formalmente que el transporte público urbano está en emergencia y no existe al momento la capacidad para cubrir la demanda de pasajeros. Además, se suma que el estado mecánico de las unidades no es del todo óptimo y la frecuencia de las rutas tampoco es eficaz para abastecer a los usuarios. Las apreciaciones mencionadas anteriormente están sustentadas en el Plan Metropolitano de Desarrollo y Ordenamiento Territorial [2], documento que detalla el panorama actual y a largo plazo y en el que se afirma que, según la tendencia, la movilidad a futuro es un asunto insostenible.

Para tener una idea más clara de la problemática de movilidad, basta con recurrir a la información publicada por Lucero [3], con base en datos de la Secretaría de Movilidad, quien expone que en los últimos 10 años hubo un crecimiento promedio de vehículos de 7.5% anual, lo que corresponde a 35 000 carros que aumentan año tras año, hecho que ocasiona una saturación progresiva inminente. En este contexto, cabe enfatizar que el 35% de la red vial principal de la ciudad está saturada, y la tendencia continúa al alza pese a haber aplicado medidas de restricción como el Pico y Placa y Hoy no Circula [4].

Es importante mencionar que, durante la pandemia, los quiteños tuvieron que atenerse a determinados horarios de movilidad planteados en aras de evitar las aglomeraciones, y que de cierta forma aplacaron momentáneamente la saturación vehicular. No obstante, de nada sirvieron estas medidas a mediano plazo, dado que la dinámica post-COVID es caótica. Ante lo mencionado, se observa que la ciudadanía vive una tormentosa realidad vial al momento de movilizarse.

La alternativa de transporte que aportará a la movilidad de los quiteños a mediano y largo plazo es el metro de Quito, opción planteada en marzo de 2011 por la alcaldía de Augusto Barrera y cuya ejecución ha sido llevada a cabo por la empresa del Metro de Madrid, entidad con una larga trayectoria de experiencia en este ámbito [5]. En torno a este proyecto, el Municipio del Distrito Metropolitano de Quito (MDMQ), la Empresa Municipal de Movilidad y Obras Públicas (EPMMOP) y la Gerencia de Planificación de Movilidad estructuraron el denominado Plan Maestro de Movilidad Para el Distrito Metropolitano de Quito (PMM), con vigencia hasta 2025. Esta propuesta plantea una perspectiva integral para el manejo del transporte público y toma como eje articulador al metro, de tal manera que apunta a consolidar un sistema eficiente, sostenible e integrado, no solo a nivel físico sino en cuanto a las tarifas. De este modo, el enfoque es que la productividad y calidad de vida de los quiteños mejore sustancialmente [6].

El metro se estructura como parte de un plan sólido, el que constituye, una alternativa rápida, segura y sustentable, desde la que se despliega todo el sistema de transporte público de la ciudad [6]. La infraestructura del metro la conforman talleres, chocheras, el túnel de 22,6 km de extensión que conecta al terminal Quitumbe y el Labrador, el tramo en el que se distribuyen en total 15 estaciones, 5 de reserva, 13 pozos de ventilación, emergencia y bombeo, 11 subproyectos de sistemas de equipamiento e instalaciones y 18 trenes de 6 vagones cada uno. La ruta completa se puede recorrer en 34 minutos con una capacidad máxima por tren y por viaje de 1230 pasajeros [7].

Este proyecto ha generado una gran cantidad de opiniones encontradas entre sus detractores y partidarios, quienes critican o defienden su implementación. En este contexto, el presente estudio se enfoca en comprender las diversas posturas y opiniones sobre el metro de Quito, a través de un análisis de minería de datos en Twitter. Con el fin de abordar esta problemática, se plantea la siguiente interrogante: ¿cómo podemos analizar los tweets de los ciudadanos para detectar las diferentes opiniones acerca del metro de Quito?

1.2. Propuesta

La minería de datos, o *data mining*, consiste en un mecanismo de extracción de significativas cantidades de información sumamente útiles y entendibles [8]. Por consiguiente, este proceso responde totalmente a las nuevas dinámicas sociales en las que el Internet se abre campo como la fuente de conocimientos más grande del mundo. Dentro de esta coyuntura, María Isabel Magaña (docente de periodismo de datos de la universidad de La Sabana) afirma que los datos se consideran al día de hoy como el nuevo oro, en vista de que son el pilar para tomar decisiones trascendentales a distintos niveles, gracias a los tan variados recursos tecnológicos que están a disposición [9].

Es importante señalar que en la actualidad se comparte una infinidad de información cada segundo, a través de las distintas redes sociales como lo es Twitter. Esta red social es muy utilizada a nivel mundial para publicar contenido instantáneo de cualquier índole; además, en Ecuador, alrededor de 1.1 millones de personas son usuarios de esta red social [10].

El proyecto ha enfrentado diversos inconvenientes en todas sus etapas, desde la planificación hasta la inauguración. Entre estos problemas se incluyen retrasos en los plazos de entrega, la necesidad de rediseñar tres estaciones, lo que resultó en un aumento en el costo de la construcción. Además, la pandemia obligó a detener la obra por cerca de tres meses. Quizás uno de los inconvenientes más preocupantes es la afectación a varias viviendas en el sector de Solanda, ubicado en el sur de la ciudad, que presentan paredes cuarteadas, ventanas rotas y otros problemas que las convierten en espacios inhabitables.

Es importante aclarar que las publicaciones de Twitter son conocidas como tweets. Los tweets como fuente de información pueden utilizarse para analizar determinadas emociones con la implementación de diversos procesos computacionales. En otras palabras, es posible interpretar si una postura es positiva, negativa, objetiva o neutral [11] y así identificar la tendencia del sentimiento. El desarrollo del proyecto implica la extracción de tweets que mencionen el metro de Quito, en este proceso se busca recopilar una amplia variedad de tweets que reflejen diferentes perspectivas. El siguiente paso es agrupar los tweets que tienen relación entre sí con el fin de obtener los temas más relevantes y clasificarlos según su polaridad positiva o negativa. Para ello, es esencial implementar una estructura que permite identificar los distintos componentes de un tweet, como las menciones (@), los hashtags (#), enlaces (L) y las cadenas de texto. De esta manera, se puede obtener datos estructurados y con información relevante sobre las opiniones publicadas [12].

1.3. Objetivo general

Desarrollar un prototipo de sistema para visualización de datos de análisis de opiniones de los ciudadanos acerca del metro de Quito, usando la información extraída de la red social Twitter.

1.4. Objetivos específicos

Para el desarrollo del presente trabajo se consideran los siguientes apartados

- Investigar trabajos relacionados.
- Utilizar la metodología CRISP-DM para el desarrollo del proyecto de análisis de opiniones en Twitter sobre el metro de Quito.
- Desarrollar un prototipo de datos para la visualización de datos.
- Interpretar los resultados obtenidos en el análisis de opiniones.

1.5. Trabajos relacionados

Analizar los sentimientos es un camino trascendental que permite comprender a profundidad las opiniones publicadas entre una gran cantidad de información en plataformas digitales como: redes sociales, microbloggings, etc. Para Prabowo y Thelwall [13], este análisis es una tarea crucial dentro del procesamiento del lenguaje natural, dado que comprender la emocionalidad que está intrínseca en un escrito es sumamente útil para diversas aplicaciones. La importancia de comprender las emociones se hace evidente en el ámbito empresarial, ya que esto permite analizar el comportamiento del consumidor y su percepción de la marca.

La recopilación de estudios relacionados con este tema se llevó a cabo mediante la búsqueda en diversos sitios web, tales como *Google Scholar*, *Dialnet* y *World Wide Science*. Para ello, se utilizaron términos clave como "análisis de opiniones en tweets" y "minería de opiniones en redes sociales". Las búsquedas se realizaron en inglés y en español. La Tabla 1 detalla la cantidad de artículos obtenidos en los sitios web mencionados, utilizando un filtro de tiempo correspondiente a los últimos 5 años.

Tabla 1. Cantidad de artículos obtenidos por sitios web.

Sitio web	Total de artículos
Google Scholar	12.700
Dialnet	238
World Wide Science	2.359

La cantidad de artículos presentados en la tabla 1 fueron obtenidos en la primera búsqueda. Posteriormente, se filtraron los artículos con la opinión pública sobre el transporte urbano. A continuación, se detallan algunos de estos artículos.

El estudio de Hollander y Renski [14], afirma que el análisis de sentimientos es una herramienta útil para evidenciar las actitudes de las personas en entornos urbanos. En

su investigación, utilizaron el análisis de sentimientos de Twitter para estudiar la experiencia urbana y obtener conclusiones para la política pública. Después de analizar más de 300,000 publicaciones de Twitter de 50 ciudades de Estados Unidos, concluyeron que no hay una diferencia estadísticamente significativa en los sentimientos entre ellas. Esto sugiere que el análisis de sentimientos de Twitter puede ser una forma útil de comprender las actitudes urbanas.

El transporte es fundamental para la vida humana y resulta esencial comprender la opinión del público. Narayan C. et al. [15] presentaron en el 2019 un enfoque de minería de opiniones basado en tweets relacionados con el tráfico para determinar el sentimiento de los ciudadanos sobre los problemas de transporte urbano. Para ello, se realizó un proceso de extracción, procesamiento y clasificación de tweets basados en la ubicación de las cuatro ciudades indias como: Delhi, Bombay, Bangalore e Hyderabad. Finalmente emplearon un método basado en diccionarios para determinar el sentimiento y clasificar su polaridad, lo cual facilita la evaluación del nivel de satisfacción de los usuarios del transporte.

Igualmente, en el 2020, Qi B. et al. [16] presentaron un marco integral para extraer y analizar eficientemente las opiniones públicas sobre los servicios de transporte en Twitter. Para ello recopilaron y preprocesaron datos de Twitter de la zona de Miami-Dade durante el periodo del 2017 y 2018. A continuación, se filtraron las opiniones sobre el servicio de transporte utilizando modelos de clasificación de textos entrenados mediante un conjunto de datos etiquetados. El contenido semántico es extraído a partir de técnicas de modelado de temas (LDA) y tokenización. Finalmente, el método de clasificación de textos permitió reducir costes de tiempo y mejorar la precisión del nivel de satisfacción de los usuarios y la clasificación de lo términos pertenecientes a un tema.

En el año 2021, Moreno, A.; Inglesias, CA [17]. Realizaron un estudio sobre las opiniones de usuarios que utilizaron el medio de transporte Uber. Los datos fueron extraídos de comentarios en Twitter en el que se etiqueto la cuenta @Uber_Spport. En este estudio se utilizó el modelo probabilístico LDA para la extracción de tópicos y la utilización de un servicio gratuito para el análisis de sentimiento y emociones. Este tipo de estudio demuestra la importancia de las marcas en comprender las experiencias de sus clientes, de esta manera les permite mantener una ventaja competitiva en el sector.

Así mismo, en 2021 en *Business Information Systems Workshops* [18], una conferencia científica internacional, que se celebra anualmente, presentó el articulo “Desafíos de la extracción de datos de Twitter para analizar el rendimiento del servicio: un estudio de caso del servicio de transporte en Malasia”. Este estudio se centra en la minería de datos de Twitter para analizar el rendimiento del servicio de transporte en Malasia. Los autores discuten los desafíos que enfrentan al trabajar con grandes cantidades de datos de Twitter y describen los modelos utilizados para identificar temas y realizar el análisis de sentimiento. Los resultados muestran que el análisis de sentimiento es una herramienta útil para comprender las opiniones de los usuarios sobre el servicio de transporte y cómo mejorar su rendimiento. Además, se discuten las limitaciones de la minería de datos de Twitter y se sugieren futuras investigaciones para mejorar la precisión de los modelos utilizados.

Los trabajos expuestos anteriormente son la base para estructurar el presente proyecto que se centra en el entorno ecuatoriano específicamente en la ciudad de Quito, dado a que aportan con exactitud las técnicas y herramientas utilizadas. Además, profundizar los conceptos y metodologías que permitirán la extracción de tópicos y la clasificación de sentimientos positivos y negativos. Así mismo, pudo identificarse las limitaciones y desafíos presentes en esta área de estudio, de tal modo que fue posible aplicar un enfoque metodológico más preciso y efectivo.

1.6. Marco teórico

Con la finalidad de realizar un abordaje claro con respecto al enfoque del presente estudio, se lleva a cabo un análisis de diversos términos relacionados al tema, así como de algoritmos y técnicas aplicadas para su desarrollo.

1.6.1. Minería de datos

El *data mining* es un área enfocada en obtener información relevante de un gran conjunto de datos de cualquier índole: bancos, salud, marketing, comercio exterior, entre otros [19]. El propósito de este proceso es identificar determinados patrones y tendencias para transformarlos en conocimiento totalmente accesible, de gran utilidad y que aporta a contar con mayor sustento para tomar decisiones más acertadas. El camino a seguir consta de cuatro etapas: establecer el problema, preparar los datos, modelar los datos y analizar los resultados [20]; cada una de estas fases son descritas con mayor profundidad en el apartado de metodología.

1.6.2. Procesamiento del Lenguaje Natural (NPL)

El procesamiento del lenguaje natural (NPL) es un área interdisciplinaria relacionada a la informática y lingüística que está centrada en comprender la intercomunicación que se puede entablar entre las personas y máquinas. La premisa es que las computadoras sean capaces de entender y analizar el lenguaje humano oral y escrito y, por consiguiente, también implica tener una comprensión clara de los sentimientos que ello involucra [21].

1.6.3. Análisis de sentimientos

Como fue mencionado previamente, el análisis de sentimiento es un recurso de *data mining* para organizar las distintas opiniones escritas en categorías preestablecidas como por ejemplo positivas y negativas [22]. Este proceso es sumamente útil en marketing, finanzas, estudios sociales y cada vez más se lo aplica en otros ámbitos.

En la primera fase se recaban los datos a analizar como reviews de productos, comentarios publicados en plataformas digitales o, en este caso, los tweets. Posteriormente, esta información es procesada para descartar aquellos datos que no son necesarios y transformar el texto a un formato estructurado con el fin de tener una comprensión más clara. En tercer lugar, se determinan las características a tomar en

cuenta para el análisis de los sentimientos. Luego, para clasificar estos resultados de manera binaria (positivo o negativo), es aplicado un modelo de aprendizaje supervisado. Por último, es necesario analizar el modelo para corroborar su precisión y efectuar los ajustes correspondientes para potenciar su desempeño [23].

1.6.4. Preprocesamiento de datos

Una vez que los datos han sido recopilados, se lleva a cabo su procesamiento para transformarlos y depurarlos con el fin de contar con información de calidad y así aplicar modelos y algoritmos de análisis o entrenamiento. Por su puesto, es crucial emplear la técnica de preprocesamiento más conveniente para que el análisis sea más exacto y significativo [24].

Una de las técnicas más comunes para preprocesar datos textuales es la tokenización, que consiste en fraccionar todo el texto en palabras autónomas [25]; además existen varios factores a tomar en cuenta como eliminar aquellos datos incompletos o duplicados, normalizar los datos, entre otros. Otro método es la selección de características relevantes, que permite disminuir el volumen de un conjunto de datos y generar un subconjunto con información que comparte determinados rasgos y descartar los datos que realmente no brindan conocimiento útil y solo hacen ruido al análisis [26]; por consiguiente, mejora la eficiencia y precisión de los algoritmos al tener menor cantidad de datos.

1.6.4.1. Lematización

La lematización es una técnica de NPL que transforma las palabras a su estructura raíz, de tal modo que se busca disminuir la variabilidad morfológica y tener información más exacta. Al contrario de la derivación, que se enfoca en eliminar sufijos y prefijos, la lematización es más compleja en vista de que emplea un diccionario para corroborar la estructura canónica que tiene cada palabra del texto. Por eso, mediante esta técnica se puede realizar un tratamiento más acertado a verbos conjugados, sustantivos plurales y demás variaciones [27].

Diversas investigaciones han evidenciado el gran aporte que brinda este método para recuperar información, clasificar textos, para la minería de opiniones y la traducción automática. Por ejemplo, Liu et al. [28] estudiaron cómo impacta esta técnica en la precisión al momento de realizar la clasificación de textos y determinaron que hubo una mejora significativa en el desempeño del modelo en contraste con otros. Así mismo, Szymański et al. [29] aplicaron la lematización para contar con una mejor traducción automatizada de textos jurídicos, amenorar los errores y lograr mayor comprensión del contenido. Estas investigaciones son solo dos ejemplos de la efectividad de este modelo para el NPL y el potencial que tiene para alcanzar mayor precisión y calidad en los resultados.

1.6.4.2. Tokenización

La tokenización es un paso esencial en el NPL enfocado en segmentar un texto en unidades lingüísticas significativas, denominadas tokens, que pueden ser palabras, frases, símbolos u otros elementos utilizados para representar el texto en una estructura determinada a fin de que sea procesada por un algoritmo. Entonces, por nombrar un ejemplo práctico, en la frase *el perro está despierto*, los tokens son cada una de las palabras que conforman esta oración. Este paso es trascendental para preprocesar la información y tiene múltiples aplicaciones: recuperar datos, clasificar texto, extraer información o realizar traducciones automáticas [25].

En este proceso son aplicadas una serie de reglas para la segmentación del texto, entre ellas puede nombrarse la eliminación de puntuación o el separar palabras compuestas. También implica normalizar el texto, por lo que se suelen quitar caracteres no alfabéticos, convertir todas las palabras en minúsculas y eliminar aquellos términos irrelevantes para el análisis [30].

1.6.5. Algoritmo de aprendizaje supervisado – clasificación

Este es un modelo de aprendizaje automatizado que se entrena a raíz de una serie de datos que cuentan con pares de entrada-salida; el fin es clasificar nueva información gracias a datos entrenados [31]. En términos más simples, implica entablar una relación entre características de entrada y etiquetas de salida de tal modo que se puede clasificar la información adecuadamente. Cabe mencionar que hay diversas variaciones de este modelo, una de ellas y que tiene excelentes resultados es el *Support Vector Machine* (SVM).

1.6.5.1. Support Vector Machine (SVM)

El SVM es un algoritmo de aprendizaje utilizado para la clasificación de datos en varias categorías, hecho que es posible gracias a que se encuentra un hiperplano que separa los datos de distintas clases con el mayor espacio posible entre ellos. Cabe indicar que el hiperplano es una especie de línea que genera una división del espacio en dos segmentos; al utilizarse en un problema de clasificación binaria, no es más que una línea que divide los puntos de dos distintas clases [20].

Los conceptos analizados hasta el momento son los requeridos para llevar a cabo el análisis de sentimiento. En ese sentido, los dos subapartados siguientes realizan un breve repaso de los fundamentos para extraer los tópicos.

1.6.6. Extracción de tópicos

La extracción de tópicos consiste en identificar los temas claves (tópicos) que se encuentran en un conjunto de datos [32], de tal modo que se identifica y agrupa la información similar que forma los clústeres.

1.6.6.1. GSDMM

El *Gibbs Sampling Algorithm for the Dirichlet Multinomial Mixture Model* (GSDMM) es un modelo de clustering probabilístico aplicado para hallar tópicos en voluminosos conjuntos de datos no estructurados. Aquí se usa un algoritmo de muestreo de Gibbs que permite asignar documentos a tópicos a través de iteraciones. En cada iteración, todo documento es tratado de manera independiente y se calcula la probabilidad de que corresponda a un tópico específico [32], aunque también se toma en cuenta la posibilidad de que pertenezca a un tópico que no haya sido establecido. Además, el GSDMM emplea una matriz de similitud entre documentos con la finalidad de entablar un vínculo entre ellos y ajustar la distribución de tópicos en consecuencia [33].

En resumen, este algoritmo es sumamente efectivo para determinar tópicos a lo largo de un gran conjunto de datos, por ese motivo se lo utiliza en aplicaciones de minería de datos como por ejemplo para la extracción de tópicos [34].

2 METODOLOGÍA

Para el desarrollo de este proyecto se utilizó la metodología CRIPS-DM (*Cross Industry Standard Process for Data Mining*). Esta metodología es una de la más populares en la gestión de proyectos de *data mining*, actualmente es impulsada por IBM, según la encuesta realizada por *KDnuggets* desde 2002 hasta 2014, posiciona a CRISP-DM como una de las metodologías más usada seguida de SEMMA y KDD, sin embargo, en la encuesta se puede apreciar y destacar que en proyectos de minería y análisis de datos un gran porcentaje señala que utilizan una metodología propia, estos datos se pueden apreciar en la figura 1.

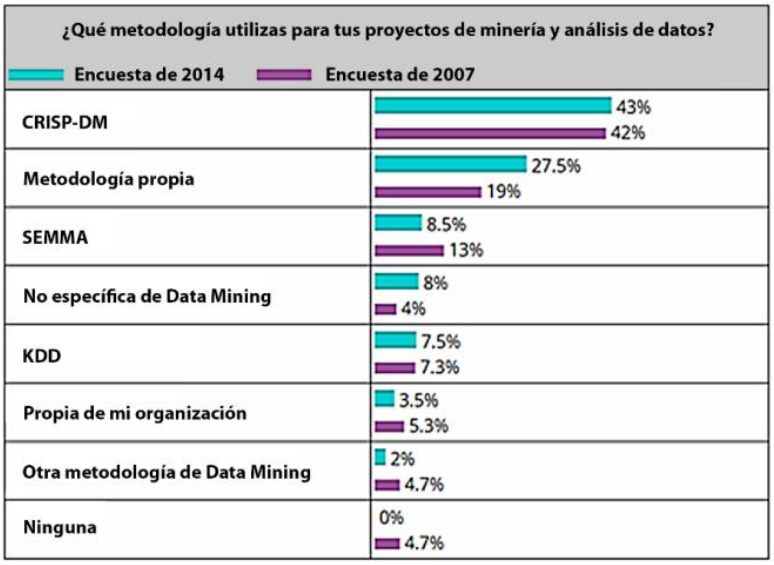

Figura 1. Encuesta realizada por Kdnuggets.

CRISP-DM consta de 4 niveles, como se puede visualizar en la figura 2. Estos niveles están organizados de forma jerárquica yendo de lo más general a lo más específico. Una de las características que más destaca en la metodología CRISP-DM es por ser

flexible con sus fases, ya que se puede modificar, repetir u omitir las fases adaptándose a casi cualquier proyecto [35].

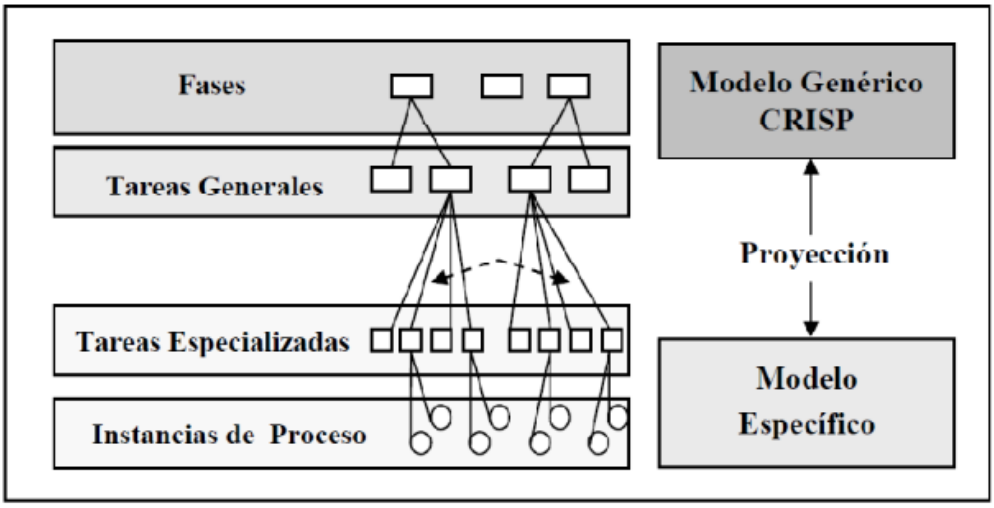

Figura 2. Niveles de CRISP-DM.

Teniendo en cuenta que las fases de la metodología CRISP-DM son flexibles, se propone como ciclo de vida del proyecto la figura 3. En este grafico se visualiza las fases generales de la metodología que se ha escogido los cuales contemplan la recolección de datos tanto para el análisis de tópicos como para el entrenamiento de los algoritmos. El preprocesamiento de datos es una fase crucial, donde se realiza la limpieza y estructuración de los datos, tanto para el análisis como para el entrenamiento. Tras el entrenamiento, se obtienen los tópicos y los clústeres correspondientes. Finalmente, se estructura el dashboard utilizando tecnologías como *Plotly*, *Dash* y HTML para presentar los resultados de manera clara y efectiva

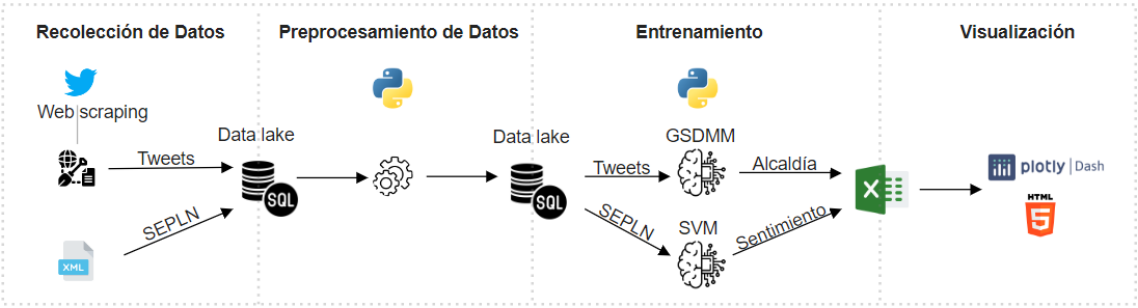

Figura 3. Ciclo de vida del proyecto.

2.1. Comprensión del negocio

Durante esta fase, se busca comprender a fondo los objetivos del problema y los requisitos del proyecto. En este sentido, se establece como punto de partida el objetivo general del proyecto.

> Desarrollar un prototipo de sistema para visualización de datos de análisis de opiniones de los ciudadanos acerca del metro de Quito, usando la información extraída de la red social Twitter.

Se plantea como objetivo del problema, realizar un análisis de opiniones de los usuarios que han publicado acerca del metro de Quito en la red social Twitter. Para realizar el análisis de opiniones, es necesario aplicar la técnica minería de opinión, está técnica permite detectar expresiones subjetivas en un grupo de texto, en este caso en la base de tweets que mencionen o tengan referencia al metro de Quito. Para el cumplimiento del objetivo, es necesario contar con una base de datos con tweets que tengan relación con el metro de Quito [36].

Se plantea los siguientes objetivos de *data mining.*

- Obtener tweets con relación al metro de Quito.
- Establecer un modelo de aprendizaje supervisado para el análisis de sentimientos.
- Establecer tópicos de los tweets recolectados utilizando algoritmos de clústeres.
- Analizar los tópicos y polaridad de sentimiento en cada periodo de alcalde de la ciudad de Quito.

2.2. Comprensión de los datos

En esta fase se accede a los datos con la finalidad de familiarizarse con ellos y realizar una exploración de estos. En este proceso se puede determinar la calidad de los datos útil para futuras fases. En este proceso se tiene un entendimiento de los datos y definir las primeras hipótesis.

Para poder acceder a los datos se utilizó la técnica *web scraping*, lo que permite extraer información de sitios web de manera automatizada [37]. Esta técnica permite absorber datos de la web y almacenarlos de manera organizada. Como primera opción se estableció utilizar la API de Twitter. Sin embargo, se identificó ciertas limitantes en la extracción de los datos. Ya que el método de búsqueda, por medio de una frase, “metro de Quito”, no recopilaba una gran cantidad de tweets. Alrededor de 18.000 tweets, además, otra limitante que se observo es que recuperaba tweets de máximo 7 días atrás.

Para solventar las limitantes presentadas con anterioridad. Se planteó usar la librería *snscrape* la cual nos permite realizar *web scraping*. Como resultado se obtuvo un total de 132.658 tweets, el rango de fechas que se tiene datos es desde 18/08/2009 hasta la 15/01/2023. Se considera que se tiene una cantidad considerable de tweets para el cumplimiento del objetivo planteado.

2.3. Preparación de los datos

En esta fase se tiene como objetivo estandarizar los datos aplicando criterio de calidad, estos criterios de calidad de los datos permiten un mayor entendimiento para la fase siguiente, la de modelado.

De los 132.658 tweets, el *dataset* inicial tiene los siguientes campos y tipos de datos. En la tabla 2 se expresa los siguientes datos, en la columna Twitter_Id es un identificador

por tweet, Date es la fecha de publicación del tweet. *User* es el nombre de usuario. Tweet es el texto publicado.

Tabla 2. Dataset inicial – datos crudos.

Columna	Tipo de dato
Twitter_Id	object
Date	datetime64[ns]
User	object
Tweet	object

2.3.1. Limpieza de datos

La limpieza de los datos se genera al seleccionar los datos que tendrán relevancia al momento de realizar el análisis, aplicar un algoritmo o graficarlos. En la limpieza se aplica varias técnicas con la finalidad de obtener datos de calidad y al mismo tiempo optimizar los modelos.

De la columna Tweet se procede a realizar la siguiente limpieza.
Estandarización de caracteres, convirtiendo el campo de tipo texto en minúscula.

- "El alcalde no toma en cuenta el tráfico en el sur" → "el alcalde no toma en cuenta el tráfico en el sur"

Eliminar los signos de puntuación.

- "tráfico" → "trafico"

Eliminar los hashtags

- "#metrodequito", "#yainaguren"

Eliminar las menciones a otras cuentas

- "@loromero", "@municipiodequito"

Eliminar los caracteres especiales

- @,.,/,*,$,&,(,),¿,¡,},_,:,;,°,|,<,>,\,^,%,!,",',)

Eliminar los enlaces o urls

- "https://www.quito.gob.ec", "http://turnosecuador.com/mdmq/".

Después de realizar la limpieza de los tweets y haberlo agregado al *dataset* original con el nombre de *Tweet_clean*, el siguiente paso es llevar a cabo la tokenización para dividir el texto en unidades semánticas significativas, como las palabras. Sin embargo, no todas las palabras en el texto son relevantes. Por lo tanto, es necesario eliminar las palabras comunes, también conocidas como *stopwords*, que no aportan información relevante y generan ruido. Una vez eliminadas las *stopwords* se realiza la lematización con el objetivo de reducir las palabras a su forma base y así simplificar el análisis de los datos.

A continuación, se procede a separar los sustantivos en una columna adicional sin perder el *dataset* original. Para ello, se utiliza la técnica de etiquetado de partes de la oración (POS) que asigna una etiqueta a cada palabra según su función gramatical en

la oración. De esta forma, es posible identificar los sustantivos en el texto y separarlos en una nueva columna, lo que facilita el análisis y la identificación de temas recurrentes en el conjunto de datos. Esta columna de sustantivos que se escribieron en el tweet se agrega al *dataset* original con el nombre de *Tweets_nouns*.

2.3.2. Estructuración de los datos

Del atributo Tweet presentada en la tabla 2, se extrae las cuentas mencionadas y los hashtags agregando estos datos en columnas nuevas, *Count_mentioned* y hashtags respectivamente. Esto permite comprender las relaciones que pueden existir entre los usuarios que se menciona y los temas de mayor tendencia que están hablando los usuarios.

Del atributo *Tweets_nouns* se realiza un filtro en donde se quedan en el *dataset* el conjunto de sustantivos mayor o igual 3. El *dataset* disminuye un total de registros de 87,035. Al realizar este filtro se mitigan aquellos tweets muy cortos que no aportan a los análisis a realizar en este proyecto.

Finalmente, se procede a establecer un formato adecuado para cada atributo, lo que facilita el manejo de los datos en los algoritmos a usar en este proyecto, se puede apreciar en la tabla 3.

Tabla 3. Dataset final.

Columna	Tipo de dato
Twitter_Id	object
Date	datetime64[ns]
User	object
Tweet	object
Tweet_clean	object
Tweets_noun	object
Coun_mentioned	object
Hashtags	object

Adicionalmente, para entrenar el algoritmo de aprendizaje supervisado, se necesita contar con un conjunto de datos previamente etiquetados con las categorías correspondientes. Este conjunto de datos se conoce como conjunto de entrenamiento, y es utilizado por el algoritmo para aprender a distinguir entre las diferentes categorías posteriormente, hacer predicciones precisas sobre nuevos textos que no han sido clasificados.

Se tiene un *dataset* entrenado extraído de la sociedad española del procesamiento de lenguaje natural, el cual cuenta con 37,000 registros, 25,000 registros con una clasificación de sentimiento positivo (P) y 12,000 registros negativos (N), el formato de la estructura de datos se puede evidenciar en la tabla 4.

Tabla 4. Dataset entrenado.

Columna	Tipo de dato
Id	str
Text	str
Sentiment	str

El dataset entrenado detallado en la tabla 4, contiene únicamente datos clasificados como positivos o negativos en la columna Sentiment. Esto se debe a que el análisis de sentimientos en español es un área de investigación en desarrollo y su complejidad aumenta al categorizar por más de dos categorías, lo que requiere más recursos [38]. Además, se ha establecido que una clasificación binaria es suficiente para cumplir con el objetivo de clasificación de sentimientos, dada la gran variedad y complejidad del idioma español.

Al *dataset* entrenado se aplica la limpieza realizada en el *dataset* inicial con la finalidad de mitigar el mismo ruido de los datos. Al finalizar esta limpieza se tiene una estructura de datos como se muestra en la tabla 5.

Tabla 5. Dataset con datos limpios.

Columna	Tipo de dato
Id	str
Text	str
Sentiment	str
Tweet_clean	str

Durante el proceso de entrenamiento de un modelo de aprendizaje, es fundamental garantizar un equilibrio en la cantidad de muestras entre las dos clasificaciones, positivo y negativo. Este proceso de equilibrar los datos es conocido como balanceo de datos. La importancia de balancear los datos radica en evitar que el modelo esté sesgado hacia la clasificación con mayor cantidad de datos y no sea capaz de aprender correctamente las características de la clasificación minoritaria [39]. Una vez que se han equilibrado los datos, la distribución de registros queda con 12,000 registros positivos y 12,000 registros negativos. En el entrenamiento del modelo, se utilizan 24,000 registros en total, donde el 20% se utiliza para realizar pruebas y el 80% restante se utiliza para entrenar el modelo. De esta forma, se garantiza que el modelo pueda aprender de manera adecuada las características de ambas clasificaciones y se pueda evaluar su rendimiento de manera efectiva.

Después de entrenar el conjunto de datos, es fundamental aplicar la técnica "*Bag of Words*" para transformar los datos de texto en datos numéricos. Esta técnica implica crear un vocabulario con todas las palabras únicas presentes en los datos de entrenamiento y representar cada documento como un vector de frecuencia de términos. Esta representación permite que los algoritmos de aprendizaje automático trabajen con los datos de texto como si fueran datos numéricos, lo que facilita el análisis y la modelización. La transformación de datos de texto en datos numéricos es un paso crucial en el análisis de sentimientos y otras tareas de procesamiento de lenguaje

natural. Como señala Zhang, "la representación numérica de los textos es un paso esencial para cualquier modelo de aprendizaje automático que se alimente de datos de texto" [40].

Finalmente, para la división de los datos por alcaldía lo que se hizo fue dividir los datos en los periodos de los diferentes alcaldes, como se muestra en la tabla 6. Presentando más datos en la alcaldía de Mauricio Rodas y menos en el periodo de Augusto Barrera.

Tabla 6. Dataset de tweets del metro de Quito segmentado por alcaldías.

Alcalde	Periodo mandato	Total tweets
Augusto Barrera	31/07/2009 - 14/05/2014	7,992
Mauricio Rodas	14/05/2014 - 14/05/2019	33,246
Jorge Yunda	14/05/2019 - 29/09/2021	15,476
Santiago Guarderas	30/09/2021 - 15/01/2023	15,476

2.4. Modelamiento

La cuarta fase tiene como objetivo seleccionar la técnica de modelado más apropiada para el cumplimiento de los objetivos planteados, por consiguiente, es necesario realizar dos tipos de algoritmos. El algoritmo de clasificación para realizar el análisis de sentimiento y el algoritmo de *clustering* que permita agrupar e identificar los tópicos.

2.4.1. Algoritmo de modelado de tópicos

Para la extracción de tópicos se establecen los modelos LDA y GSDMM siendo estos algoritmos de aprendizaje automático. Son usados comúnmente para modelamiento de tópicos para identificar los temas latentes en un conjunto de documentos [41].

LDA es un algoritmo de aprendizaje no supervisado que asume que cada documento en un conjunto de varias categorías conformando los tópicos, y cada tópico es una distribución de palabras [42]. LDA trabaja retrocediendo desde las palabras hasta los tópicos, por lo que primero asigna una distribución de tópicos a cada documento y, a continuación, asigna una distribución de palabras a cada tópico.

GSDMM también es un algoritmo de modelado de tópicos, pero a diferencia de LDA, es un algoritmo de aprendizaje supervisado. GSDMM se basa en el modelo de mezcla de Dirichlet [43], que es una forma de modelar la distribución de probabilidad de un conjunto de datos. Para seleccionar un modelo en la aplicación del proyecto se establecieron pruebas para ver cuál de los dos modelos es el adecuado.

Para seleccionar el mejor modelo entre GSDMM y LDA, se llevaron a cabo pruebas utilizando *unigram* y *bigram*. *Unigram* es una técnica que consiste en descomponer un texto en una secuencia de palabras individuales. Cada palabra se convierte en un *token* independiente que puede ser analizado y clasificado por el modelo. Por otro lado, *bigram* es una técnica que consiste en descomponer el texto en una secuencia de dos palabras consecutivas. En este caso, cada par de palabras se convierte en un token. En ambos

casos se creó un corpus de documentos a partir de los tweets recopilados y se estableció un diccionario que contiene todas las palabras que se encuentran en el corpus.
En el proceso de entrenamiento de ambos modelos se iteró 15 veces tanto para *unigram* como *bigram*, lo que significa que se realizaron 15 pasadas completas por todo el corpus de datos. Además, se especificó que se obtendrá 10 tópicos en cada modelo. Estos parámetros fueron elegidos en base a la literatura existente y se fue ajustando empíricamente para obtener los mejores resultados posibles en la clasificación de los datos de entrada.

Los resultados obtenidos al aplicar los modelos LDA y GSDMM se comparan en la tabla 7, en la que se evalúa la medida de coherencia de cada modelo. Se considera que un modelo de tópicos con una medida de coherencia más alta es mejor que uno con una medida de coherencia más baja, ya que indica que las palabras dentro de cada tópico están más relacionadas y tienen mayor sentido en conjunto. Específicamente para los modelos LDA y GSDMM, su comparación en la tabla 7 permitirá seleccionar el modelo que mejor se ajuste a los objetivos del estudio.

Tabla 7. Comparación coherence score.

Algoritmo	Coherence score
LDA - unigram	-7.571
LDA - bigram	-7.260
GSDMM - bigram	-3.685
GSDMM - unigram	-3.574

El algoritmo GSDMM - unigram obtuvo la mayor medida de coherencia con -3.57. En un estudio comparativo entre LDA y GSDMM titulado "Topic Modeling for Quality Prediction of Online Reviews: A Comparative Study of LDA and GSDMM" [44], se encontró que el modelo GSDMM con una medida de coherencia más alta superó significativamente al modelo LDA en términos de precisión y exhaustividad en la tarea de predicción de calidad de reseñas en línea. El modelo GSDMM se destacó por su capacidad de trabajar mejor en textos cortos y por tener una puntuación de coherencia más alta, lo que lo convierte en una mejor opción para la extracción de tópicos. Basándose en los resultados de este estudio y los resultados de la tabla 7, se decidió utilizar el modelo GSDMM - unigram para la extracción de tópicos en este proyecto.

2.4.2. Algoritmo de clasificación

En el proceso de clasificación, se evaluaron diferentes modelos de aprendizaje automático supervisado para determinar cuál era el mejor para predecir la polaridad entre positivo y negativo de los tweets. Se consideraron modelos como SVM, Árbol de decisión, *Naive Bayes* y *Logistic Regression*. Para cada modelo, se realizaron evaluaciones utilizando métricas de desempeño. En la tabla 8, se puede visibilizar el score de estos algoritmos. El score es una medida de evaluación para algoritmos de clasificación. Se utiliza para evaluar que tan bien el modelo puede predecir la categoría correcta para los datos de prueba [45].

Tabla 8. Score de algoritmos de clasificación.

Algoritmo	Score
SVM	0.888
Árbol de decisión	0.726
Naive Bayes	0.701
Logistic regression	0.878

Se puede atribuir el algoritmo con mejor rendimiento a SVM, este algoritmo mostro tener un mejor manejo de los datos y un buen desempeño con un gran conjunto de datos con alta dimensionalidad presente en los datos tipo texto de los tweets.

2.5. Evaluación

En esta fase se realiza una evaluación de los resultados obtenidos con la finalidad de establecer el cumplimiento de los objetivos planteados en la fase de comprensión del negocio.

Se recolectaron un total de 87,009 tweets relacionados con el metro de Quito después de llevar a cabo la limpieza de los datos. Esta cantidad de tweets resultó ser lo suficientemente considerable para realizar un análisis de data mining, por lo que se puede considerar que se cumplió el objetivo inicial de obtener tweets relacionados con el metro de Quito.

Para realizar el análisis de sentimientos se seleccionó el algoritmo de clasificación SVM, el cual comparado con otros tiene mayor score. El resultado de este proceso cumple el objetivo de realizar un modelo de aprendizaje supervisado para realizar el análisis de sentimiento.

Por otro lado, para generar los tópicos se utilizó el algoritmo GSDMM para establecer los tópicos presentes en los tweets, lo que permitió obtener información relevante acerca de las discusiones y opiniones en torno al metro de Quito. Para seleccionar el algoritmo GSDMM se realizaron pruebas en conjunto con el algoritmo LDA utilizando bigramas y unigramas.

Finalmente, para realizar un análisis por cada alcaldía se dividieron los datos, lo que permitió visualizar la evolución de la construcción del metro y la opinión de los ciudadanos en cada periodo. A continuación, se muestra para cada periodo el total de tweets, así como el porcentaje de sentimiento positivos y negativos respectivamente. También, en las tablas 9, 10, 11 y 12 se presentan las palabras más significativas, con las cuales se generaron los diferentes tópicos. De esta manera, se proporciona una visión sobre la opinión de los ciudadanos en las diferentes alcaldías.

2.5.1. Análisis de tópicos sobre el metro de Quito en la alcaldía de Augusto Barrera

Se recolecto 7,992 tweets relacionados, de los cuales 48.9% se clasificaron como positivos y el 51.1% como negativos. En la tabla 9 se evidencia los tópicos asignados a cada clúster.

Tabla 9. Resultados del análisis de tópicos sobre el metro de Quito en la alcaldía de Augusto Barrera.

Clúster	Palabras significativas	Tópico
0	metro', 'quito', 'concejal', 'costo', 'alcalde', 'proyecto', 'tema', 'pasaje', 'estudio', 'tarifa'	Desarrollo de un proyecto de metro con tecnología de tuneladora.
9	metro', 'geologo', 'obra', 'millón', 'quito', 'cuenta', 'codo', 'valor', 'accidente', 'muerte'	Estudio de impacto en el centro histórico y patrimonio de la ciudad.
6	metro', 'quito', 'alcalde', 'obra', 'gobierno', 'millón', 'proyecto', 'contrato', 'año', 'presidente'	Financiamiento del metro bajo la gestión del gerente general Jacome.
2	metro', 'quito', 'alcalde', 'ciudad', 'basura', 'obra', 'calle', 'año', 'señor', 'bogota'	Impacto en el desarrollo de la ciudad de Quito y solventar problemas de movilidad.
1	trabajo', 'metro', 'cierre', 'obra', 'calle', 'estacion', 'mes', 'tramo', 'mayo'	Inversión y financiamiento en proyecto de transporte.
8	metro', 'ciudad', 'obra', 'transporte', 'quito', 'movilidad', 'sistema', 'proyecto', 'tiempo', 'alcalde'	Proyecto de conexión entre el sur de Quito y el aeropuerto.
7	metro', 'quito', 'estacion', 'tunel', 'tren', 'obra', 'tuneladora', 'prueba', 'avance', 'ciudad'	Proyecto de construcción del metro de Quito.
4	metro', 'quito', 'logo', 'vagón', 'alcalde', 'acto', 'caso', 'empresa', 'gente', 'ciudad'	Proyecto de metro asesorado por empresas españolas.
5	metro', 'ciudad', 'alcalde', 'quito', 'gente', 'señor', 'daño', 'año', 'color'	Impacto en el desarrollo de la ciudad de Quito y solventar problemas de movilidad.
3	metro', 'quito', 'millón', 'consorcio', 'contrato', 'empresa', 'obra', 'fase', 'caso', 'fiscalia'	Inversión y financiamiento en proyecto de transporte.

2.5.2. Análisis de tópicos sobre el metro de Quito en la alcaldía de Mauricio Rodas

Se recolectó 33,246 tweets relacionados, de los cuales 37% se clasificaron como positivos y el 63% son negativos. En la tabla 10 se evidencia los tópicos asignados a cada clúster.

Tabla 10. Resultados del análisis de tópicos sobre el metro de Quito en la alcaldía de Mauricio Rodas.

Clúster	Palabras significativas	Tópico
0	'metro', 'quito', 'concejal', 'costo', 'alcalde', 'proyecto', 'tema', 'pasaje', 'estudio', 'tarifa'	El costo del pasaje será sin subsidio.
1	trabajo', 'metro', 'cierre', 'obra', 'calle', 'estacion', 'mes', 'tramo', 'mayo'	Obra del metro de Quito provoca cierres en las calles por la construcción de estaciones.
2	metro', 'quito', 'alcalde', 'ciudad', 'basura', 'obra', 'calle', 'año', 'señor', 'bogota'	Comparativa del metro de Quito con el problema de movilidad en Bogotá.
3	metro', 'quito', 'millón', 'consorcio', 'contrato', 'empresa', 'obra', 'fase', 'caso', 'fiscalia'	Especulación sobreprecio de la obra, ¿Interfiere la fiscalía?
4	metro', 'quito', 'logo', 'vagón', 'alcalde', 'acto', 'caso', 'empresa', 'gente', 'ciudad'	Se establece logo del metro de Quito.
5	'metro', 'ciudad', 'alcalde', 'quito', 'gente', 'bla', 'señor', 'daño', 'año', 'color'	Daño a la ciudad de Quito por construcción del metro.
6	'metro', 'quito', 'alcalde', 'obra', 'gobierno', 'millón', 'proyecto', 'contrato', 'año', 'presidente'	Proyecto de construcción del metro de Quito y su fuerte inversión.
7	'metro', 'quito', 'estacion', 'tunel', 'tren', 'obra', 'tuneladora', 'prueba', 'avance', 'ciudad'	Construcción en curso de los túneles y estaciones.
8	'metro', 'ciudad', 'obra', 'transporte', 'quito', 'movilidad', 'sistema', 'proyecto', 'tiempo', 'alcalde'	Desarrollo y mejora del sistema de transporte para la movilidad.
9	'metro', 'geologo', 'obra', 'millón', 'quito', 'cuenta', 'codo', 'valor', 'accidente', 'muerte'	Muerte de geólogo por accidente en las obras del metro.

2.5.3. Análisis de tópicos sobre el metro de Quito en la alcaldía de Jorge Yunda

Se recolectó 30,321 tweets relacionados, de los cuales 43% se clasificaron como positivos y el 57% son negativos. En la tabla 11 se evidencia los tópicos asignados a

cada clúster, se ha designado como * a los tópicos que no tienen relación con el metro de Quito o que no ha tenido sentido para establecer un tópico.

Tabla 11. Resultados del análisis de tópicos sobre el metro de Quito en la alcaldía de Jorge Yunda.

Clúster	Palabras significativas	Tópico
0	'metro', 'transporte', 'ciudad', 'sistema', 'quito', 'obra', 'movilidad', 'servicio', 'tren', 'estación'	Desarrollo y mejora del sistema de transporte para la movilidad.
1	'metro', 'transporte', 'quito', 'ciudad', 'año', 'servicio', 'modelo', 'empresa', 'tarifa', 'alcalde'	El modelo de gestión del servicio de metro enfrenta desafíos en la implementación de una tarifa.
2	'metro', 'ciudad', 'transporte', 'persona', 'estación', 'quito', 'espacio', 'sistema', 'gente', 'parte'	Sin espacios para el comercio en las estaciones del metro de Quito.
3	'metro', 'quito', 'trabajo', 'ciudad', 'calle', 'parque', 'espacio', 'zona', 'obra'	Metro de Quito influye en zonas aledañas a estaciones como parques, calles, espacios públicos.
4	'medida', 'quito', 'bioseguridad', 'metro', 'control', 'salud', 'uso', 'contagio', 'mascarilla', 'espacio'	Se paraliza la obra del metro de Quito a causa de la pandemia COVID19.
5	'metro', 'quito', 'alcalde', 'gerente', 'empresa', 'modelo', 'obra', 'caso', 'concejal', 'cargo'	Se designa nuevo gerente del metro de Quito, sin alcalde por caso de corrupción.
6	'metro', 'obra', 'año', 'quito', 'millón', 'mes', 'marzo', 'prueba', 'fecha', 'entrega'	Aumento de costo de la obra, se aplaza la fecha de entrega.
7	'metro', 'quito', 'alcalde', 'ciudad', 'obra', 'gente', 'año', 'señor', 'pueblo', 'ahi'	Se retoma obra del metro de Quito.
8	'kit', 'metro', 'quito', 'zona', 'apoyo', 'ciudad', 'persona', 'coordinacion', 'ayuda', 'entrega'	*
9	'vida', 'salud', 'quito', 'casa', 'tiempo', 'mano', 'persona', 'hogar', 'mujer', 'familia'	*

2.5.4. Análisis de tópicos sobre el metro de Quito en la alcaldía de Santiago Guarderas

Se recolectó 15,476 tweets relacionados, de los cuales 33.5% % se clasificaron como positivos y el 66.5% son negativos. En la tabla 12 se evidencia los tópicos asignados a cada clúster, se ha designado como * a los tópicos que no tienen relación con el metro de Quito o que no ha tenido sentido para establecer un tópico.

Tabla 12. Resultados del análisis de tópicos sobre el metro de Quito en la alcaldía de Santiago Guarderas.

Clúster	Palabras significativas	Tópico
0	'metro', 'medio', 'operación', 'gerente', 'viernes', 'agosto', 'jueves', 'quito', 'lunes'	*
1	'metro', 'ciudad', 'tren', 'quito', 'estacion', 'transporte', 'estación', 'tiempo', 'sistema', 'hora'	Tiempos de viaje entre estaciones del sistema de transporte metro de Quito.
2	'metro', 'quito', 'alcalde', 'ciudad', 'año', 'obra', 'gente', 'guardera', 'inauguracion', 'sinverguenza'	Alcalde Guarderas inaugura metro de Quito con controversias en la obra.
3	'metro', 'soborno', 'quito', 'contrato', 'millón', 'pago', 'empresa', 'obra', 'españa', 'fiscalia'	Fiscalía investiga soborno por parte de empresa española en la obra del metro de Quito.
4	'metro', 'empresa', 'proceso', 'quito', 'diciembre', 'gerente', 'operador', 'operacion', 'alcalde', 'año'	Gerente estructura el proceso de capacitación para el operador del metro.
5	'metro', 'estación', 'sistema', 'transporte', 'quito', 'diciembre', 'tren', 'movilidad', 'fase', 'seguridad'	Inicia fase de aprendizaje en algunas estaciones este 31 de diciembre.
6	'metro', 'quito', 'alcalde', 'ahi', 'persona', 'año', 'fin', 'guardera', 'tema'	*
7	'metro', 'quito', 'año', 'ciudad', 'obra', 'transporte', 'sistema', 'movilidad', 'proyecto', 'tiempo'	Tras años de construcción el sistema de movilidad metro de Quito presenta fallas.
8	'metro', 'millón', 'gobierno', 'quito', 'presidente', 'año', 'dinero', 'prioridad', 'obra', 'medicina'	Gobierno designa millones a la obra del metro, ¿Hay prioridad frente en la compra de medicinas?
9	'metro', 'trabajo', 'quito', 'zona', 'espacio', 'calle', 'movilidad', 'estacion', 'obra', 'ciudad'	Inicia trabajos en calles aledañas de distintas estaciones del metro.

Después de revisar los resultados obtenidos y compararlos con los objetivos planteados al inicio del proyecto, se determina el cumplimiento de todos los objetivos, logrando así los resultados esperados.

2.6. Despliegue

La fase final de la metodología CRISP-DM tiene como objetivo desarrollar un prototipo de sistema, que en este caso es un tablero de mando (dashboard), el cual muestra información de manera clara y amigable para el usuario. Para su desarrollo, se utilizó la metodología en cascada descrita en el siguiente capítulo. Además, se estableció que el Laboratorio de Analítica de Datos y Ciberseguridad - ADA LOVELACE absorberá y mantendrá el proyecto, así como su publicación en un servidor para que los datos puedan ser accesibles para su interpretación y análisis.

3 DESARROLLO DEL PROTOTIPO

El desarrollo del prototipo de sistema está basado en la metodología en cascada. La metodología en cascada se caracteriza por ser lineal y secuencial, lo que significa que cada fase se completa antes de avanzar a la siguiente [46].

3.1. Análisis

La primera fase tiene como objetivo detallar las necesidades y objetivos para posteriormente reunir los requisitos que se deben cumplir en el desarrollo del sistema. El objetivo es comprender los requisitos funcionales, los requisitos no funciones como la disponibilidad, escalabilidad, etc. y las interacciones del usuario en la pantalla principal y el dashboard.

3.1.1. Requerimientos

Los requerimientos se obtuvieron a través de una encuesta realizada por medio de formularios de Microsoft, la encuesta realizada se detalla en el ANEXO I. Los requerimientos de *look and feel* obtenidos por medio de la encuesta se aplican para una mejor experiencia de usuario y la visualización de información fundamental del análisis de opiniones sobre el metro de Quito en las diferentes alcaldías.

a) Funcionalidades del sistema

- **Pantalla inicio**

F1: El sistema mostrará una pantalla inicio con información general.
Esta pantalla consiste en una línea de tiempo con detalles generales de cada alcaldía y un hipervínculo hacia el dashboard.

- **Generación de resultados (dashboard)**

F2: El sistema generará un dashboard.
Este dashboard consiste en mostrar indicadores, graficas de barra, nube de palabras, etc., con información de la alcaldía.

b) Requerimientos funcionales

La encuesta detallada en el ANEXO I permitió obtener los requerimientos funcionales del prototipo de sistema para el análisis de opiniones sobre el metro de Quito, los cuales se encuentran detallados en las tablas 13 a la 23.

Tabla 13. Requerimiento visualización de línea de tiempo de las alcaldías.

ID:	RF-001	Relación:	
Descripción:	Visualización de línea de tiempo de las alcaldías.	Autor:	Jorge Quilumba Joel Villacis
• El sistema permitirá visualizar una línea de tiempo de las alcaldías relacionadas con el metro de quito.			

Tabla 14. Requerimiento visualización de imagen de los alcaldes.

ID:	RF-002	Relación:	
Descripción:	Visualización de imagen de los alcaldes.	Autor:	Jorge Quilumba Joel Villacis
• El sistema permitirá visualizar una imagen de cada alcalde.			

Tabla 15. Requerimiento visualización de nombres de los alcaldes.

ID:	RF-003	Relación:	
Descripción:	Visualización de nombres de los alcaldes.	Autor:	Jorge Quilumba Joel Villacis
• El sistema permitirá visualizar los nombres de cada alcalde.			

Tabla 16. Requerimiento visualización del periodo de la alcaldía.

ID:	RF-004	Relación:	
Descripción:	Visualización del periodo de cada alcaldía,	Autor:	Jorge Quilumba Joel Villacis
• El sistema permitirá visualizar el periodo de cada alcaldía.			

Tabla 17. Requerimiento visualización de tweets totales por alcaldías.

ID:	RF-005	Relación:	
Descripción:	Visualización de tweets totales por alcaldías.	Autor:	Jorge Quilumba Joel Villacis
• El sistema mostrará el número de tweets totales recolectados por cada alcaldía.			

Tabla 18. Requerimiento visualización de tweets positivos por alcaldía.

ID:	RF-006	Relación:	
Descripción:	Visualización de tweets positivos por alcaldía.	Autor:	Jorge Quilumba Joel Villacis
• El sistema mostrará el porcentaje de tweets positivos por cada alcaldía.			

Tabla 19. Requerimiento visualización de tweets negativos por alcaldía.

ID:	RF-007	Relación:	
Descripción:	Visualización de tweets negativos por alcaldía.	Autor:	Jorge Quilumba Joel Villacis
• El sistema mostrará el porcentaje de tweets negativos por cada alcaldía.			

Tabla 20. Requerimiento visualización de tópicos con su polaridad por alcaldía.

ID:	RF-008	Relación:	
Descripción:	Visualización de tópicos con su polaridad por alcaldía.	Autor:	Jorge Quilumba Joel Villacis
• El sistema permitirá visualizar un gráfico con una distribución de tópicos detallando la polaridad del tópico de cada alcaldía.			

Tabla 21. Requerimiento visualización de nube de palabras por alcaldía.

ID:	RF-009	Relación:	
Descripción:	Visualización de nube de palabras por alcaldía.	Autor:	Jorge Quilumba Joel Villacis
• El sistema permitirá visualizar una nube de palabras más frecuentes de cada alcaldía.			

Tabla 22. Requerimiento visualización de tendencias por alcaldía.

ID:	RF-010	Relación:	
Descripción:	Visualización de tendencias por alcaldía.	Autor:	Jorge Quilumba Joel Villacis
• El sistema permitirá visualizar un gráfico de las tendencias (hashtag) de cada alcaldía.			

Tabla 23. Requerimiento visualización de tweets según la fecha de publicación.

ID:	RF-011	Relación:	
Descripción:	Visualización de tweets según la fecha de publicación.	Autor:	Jorge Quilumba Joel Villacis
• El sistema mostrará el número de tweets agrupados por año y su polaridad en cada alcaldía.			

c) Requerimientos no funcionales

- **Restricciones de diseño e implementación**
  - Como *back-end* se utilizará el lenguaje de programación *Python* 3 y *Dash* con sus componentes en HTML.
  - Para el desarrollo del prototipo se utilizará el ambiente web de *Flask*.
  - Se utilizará archivos .xlsx como fuente de datos.
- **Requerimientos del producto**
  Los requerimientos no funcionales del prototipo de sistema para el análisis de opiniones sobre el metro de Quito se detallan en las tablas 24, 25, 26.

Tabla 24. Requerimiento escalabilidad.

ID:	RNF-001	Relación:	
Prioridad:		Autor:	Jorge Quilumba Joel Villacis
Descripción:	Escalabilidad.		
• El sistema será capaz de adaptar su rendimiento a medida que aumentan de manera significativa los datos.			

Tabla 25. Requerimiento mantenibilidad.

ID:	RNF-002	Relación:	
Prioridad:		Autor:	Jorge Quilumba Joel Villacis
Descripción:	Mantenibilidad.		
• El sistema estará estructurado con un código consistente, predecible y además comentado. Para que las funciones que son susceptibles a cambios en el tiempo sean de fácil implementación.			

Tabla 26. Requerimiento confiabilidad.

ID:	RNF-003	Relación:	
Prioridad:		Autor:	Jorge Quilumba Joel Villacis
Descripción:	Confiabilidad.		
• El sistema no presentará fallas durante su tiempo de operación.			

3.1.2. Otros requerimientos

a) Interfaces gráficas

Los bosquejos de cada interfaz del prototipo de sistema para la visualización de minería de opiniones basado en tweets caso de estudio metro de Quito, fueron usados de manera referencial por lo que no representan el boceto final de las interfaces ni la funcionalidad del prototipo de sistema para el análisis de opiniones sobre el metro de Quito.

La figura 4 muestra la interfaz principal que permite al usuario visualizar indicadores a nivel general de cada alcaldía y con un vínculo al dashboard que le permite obtener información más detallada.

Figura 4: Mockup de interfaz gráfica de la pantalla inicio.

La figura 5 muestra la interfaz del dashboard que permitirá al usuario visualizar la información de la alcaldía de manera más detallada por medio de gráficas.

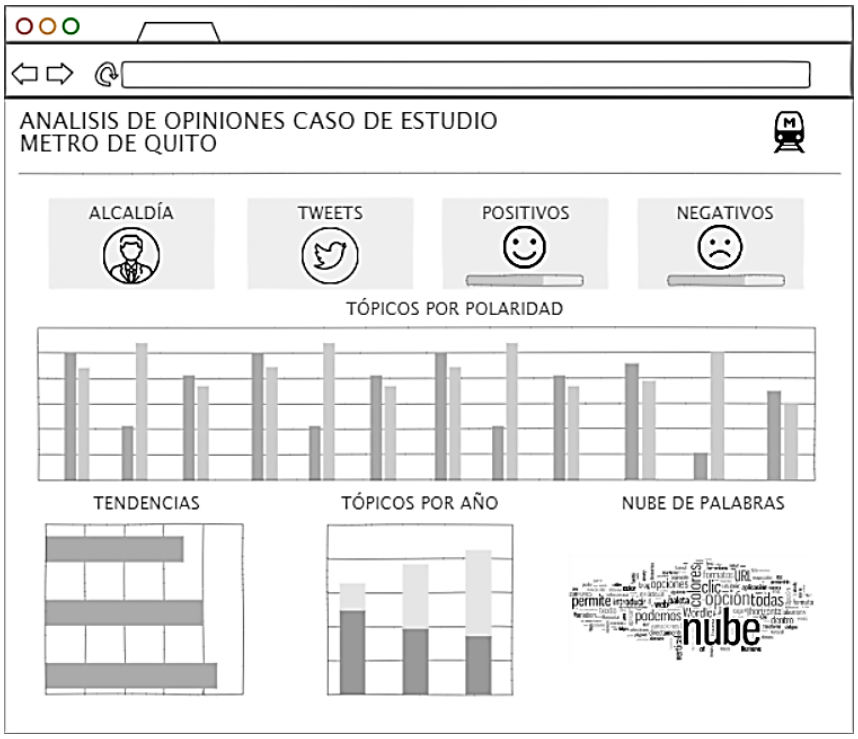

Figura 5: Mockup de interfaz gráfica del dashboard.

b) Requisitos de Hardware y Software

- **Arquitectura del prototipo**

En la figura 6 se detalla la arquitectura del prototipo de sistema.

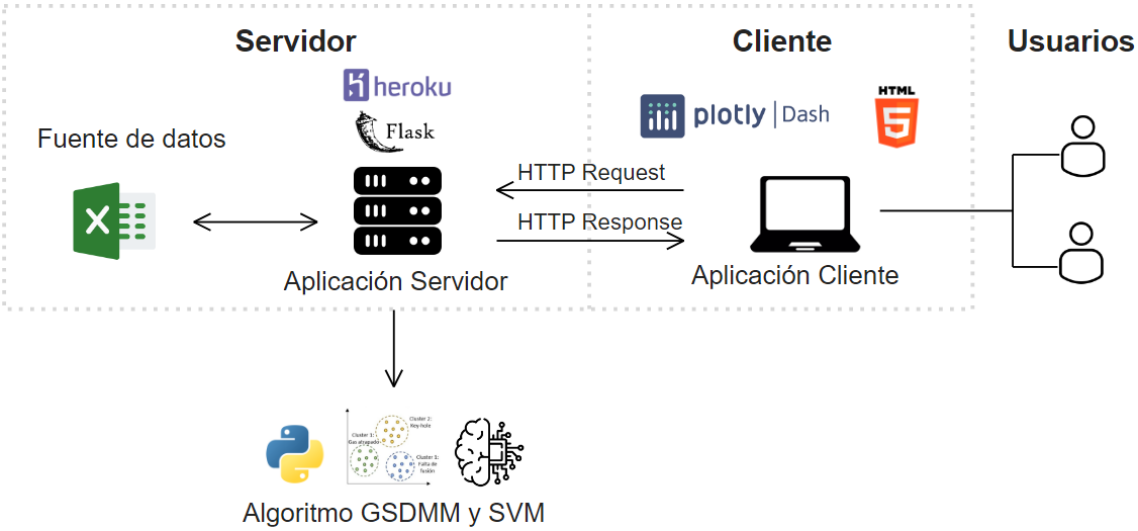

Figura 6: Arquitectura del prototipo de sistema.

3.2. Diseño

La segunda fase tiene como objetivo realizar el diseño de diagramas que permitirán tener un esquema más amplio de la funcionalidad del sistema y la interacción del usuario con cada elemento.

3.2.1. Diagrama de clases

La figura 7 muestra el diagrama de clases utilizado previo al desarrollo del prototipo de sistema para el análisis de opiniones sobre el metro de Quito.

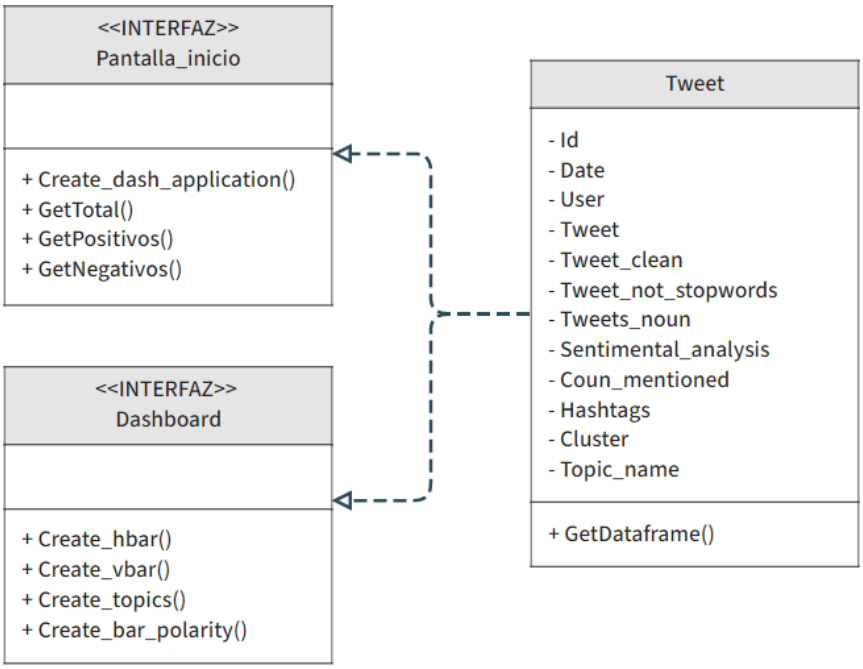

Figura 7: Diagrama de clases del dominio del problema.

3.2.2. Diagrama de casos de uso

La figura 8 muestra el caso de uso para el ingreso al sistema a través de la url que permite al usuario visualizar indicadores generales y el acceso al dashboard.

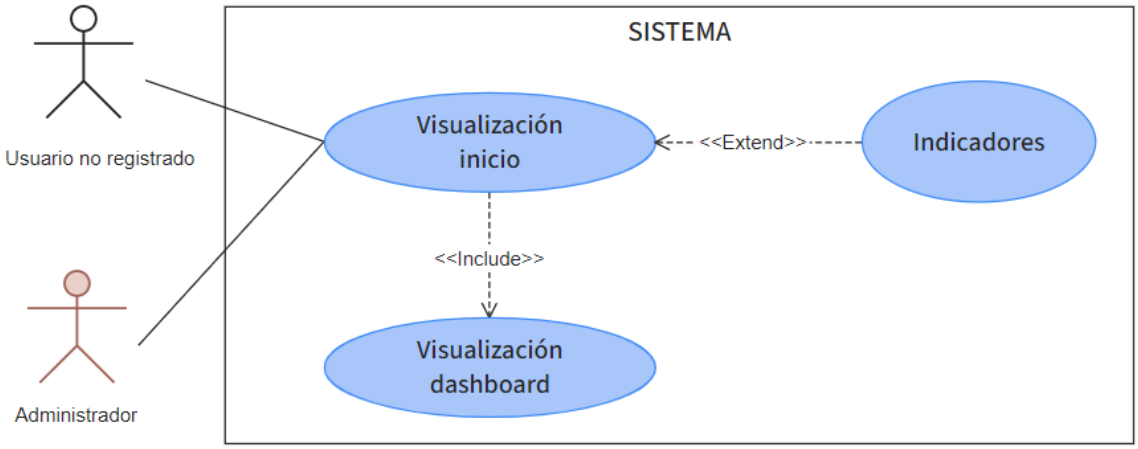

Figura 8: Diagrama de casos de uso general.

La figura 9 muestra el caso de uso para la presentación de gráficas de barras, indicadores y nube de palabras en un dashboard.

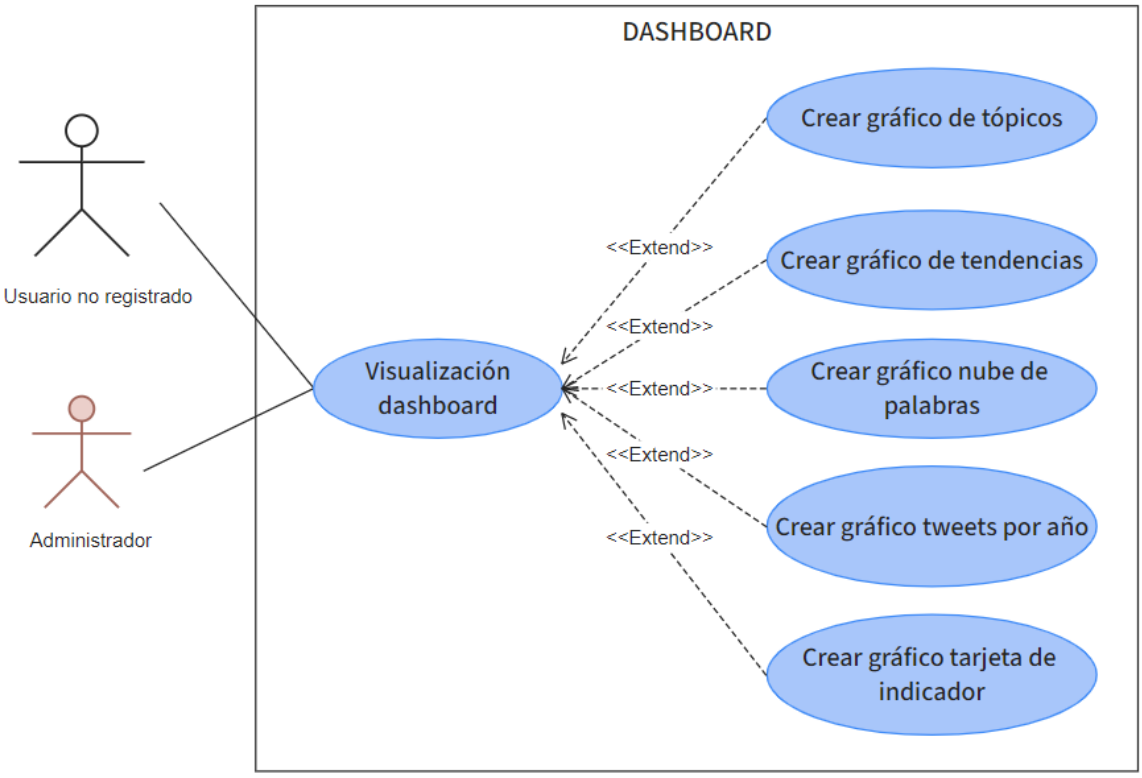

Figura 9: Diagrama de casos de uso generación de resultados (dashboard).

3.2.3. Diagrama de secuencia

F1: Pantalla inicio

La figura 10 muestra el diagrama de secuencia con el proceso que realiza el usuario para acceder a la pantalla de inicio del sistema.

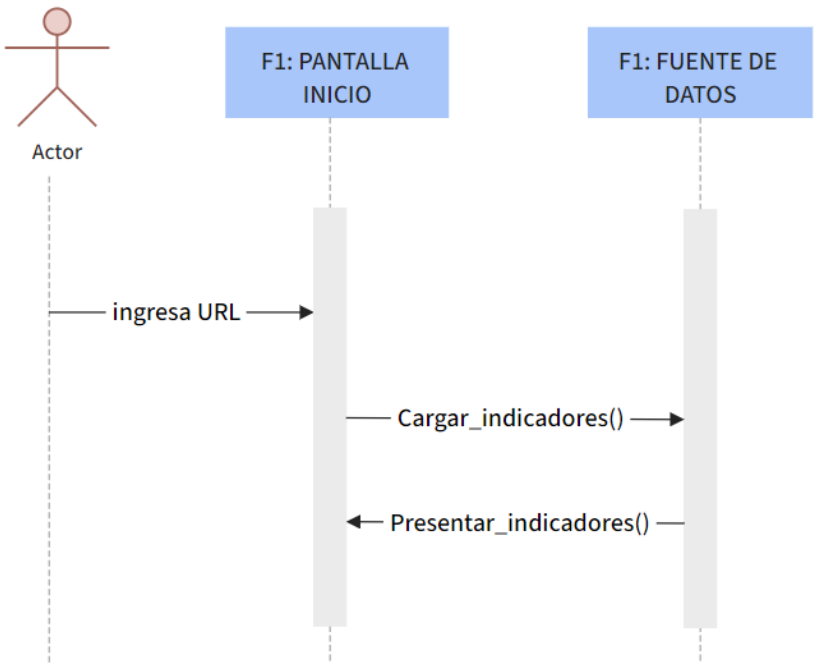

Figura 10: Diagrama de secuencia del dominio del problema.

F2: Generación de resultados (dashboard)

La figura 11 muestra del diagrama de secuencia para visualizar el dashboard con las gráficas e indicadores en el sistema.

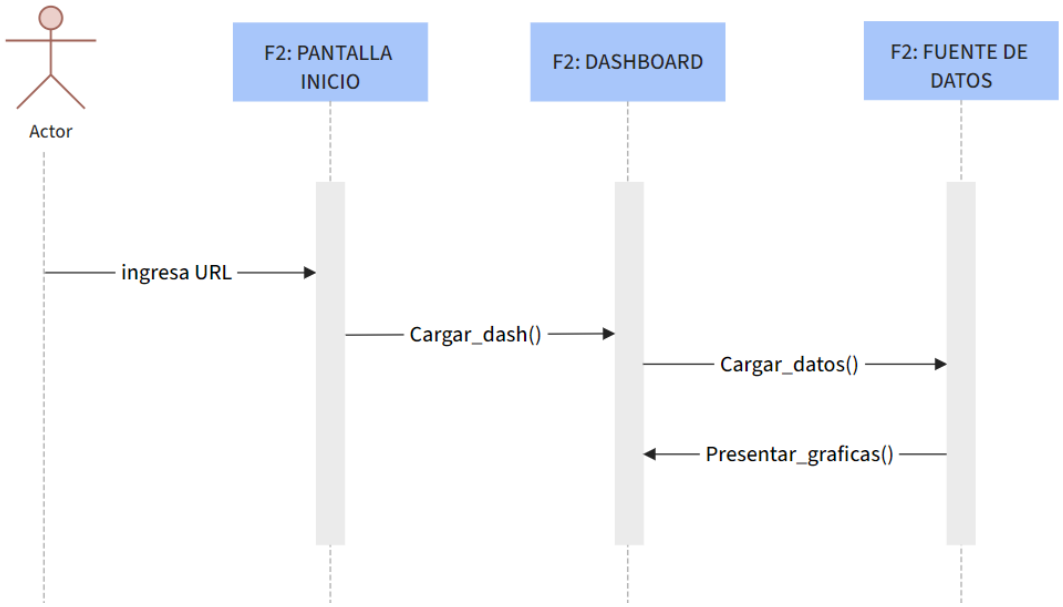

Figura 11: Diagrama de secuencia generación de resultados (dashboard).

3.2.4. Estándares de diseño y desarrollo

El estándar de desarrollo fue basado en la ISO/IEC 12207: Este estándar describe un ciclo de vida completo para el software y establece los procesos necesarios para su desarrollo, mantenimiento y retracción incluyendo procesos de diseño.

Estándares de codificación
Se utilizarán los estándares establecidos por *Flask* y las buenas prácticas de programación del lenguaje *Python*.

- Usar indentación en las funciones.
- Comentar la característica de las funciones y de los *templates*.
- Mantener menos de 80 caracteres de la codificación.
- Utilizar codificación utf-8.
- Utilizar cada elemento según la función de cada etiqueta HTML.

Estándares de interfaz gráfica
Se utilizarán las funcionalidades de *Flask* como render_template_html para la pantalla de inicio y los componentes de *Bootstrap* de la librería *Dash* para los dashboard. Esto permitirá el despliegue de resultados en las interfaces.

Mensajes de error
Se utilizarán cuadros de dialogo donde se mostrará los mensajes de error de los argumentos de código.

3.3. Implementación

La tercera fase tiene como objetivo escribir el algoritmo o código en lenguaje *Python* para el *back-end* y para el *front-end* se utilizó HTML con el framework Bootstrap y Dash. Esta fase permitirá llevar a cabo el proceso de convertir los requerimientos y diseño en un entorno de producción.

3.3.1. Pantalla inicio

La figura 12 muestra la página de inicio. Esta página contine una línea de tiempo de las alcaldías en las que se ejecutó el proyecto metro de Quito con detalles generales del alcalde, indicadores de polaridad y total de tweets, además de un hipervínculo en la imagen de cada alcalde y un botón de “más información” que tienen una redirección al dashboard.

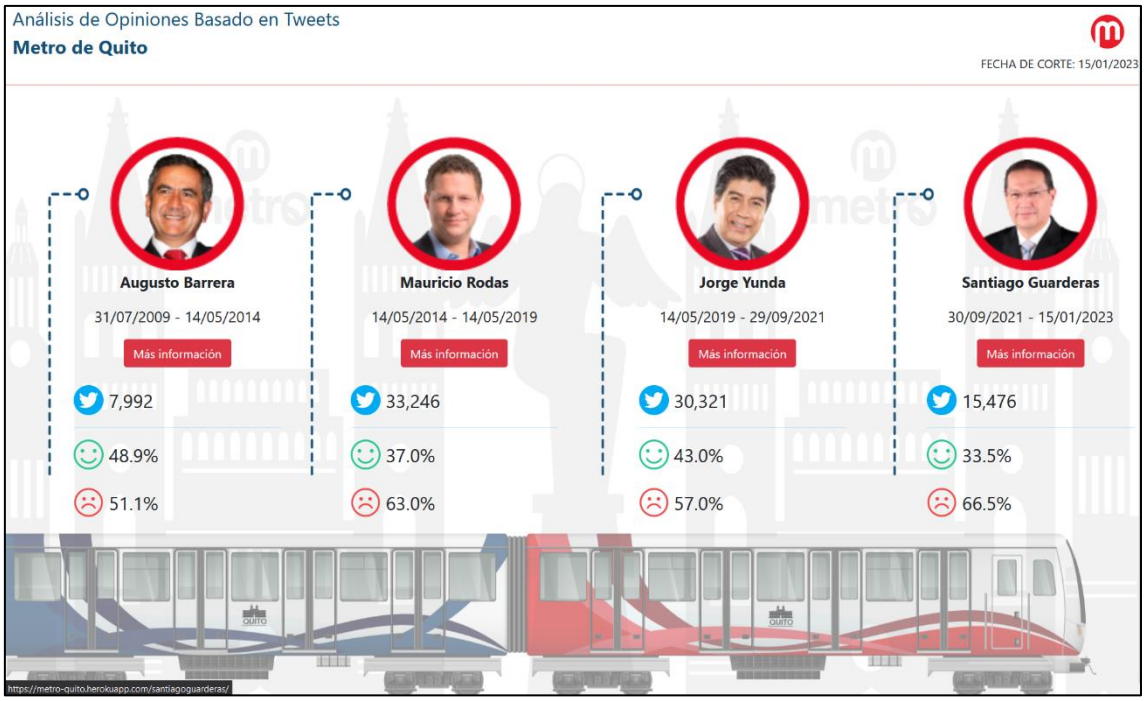

Figura 12: Interfaz Web de línea de tiempo e indicadores.

3.3.2. Pantalla dashboard

Las figuras 13,14,15 y 16 muestran las pantallas con detalle sobre las diferentes alcaldías estas pestañas contienen: información del periodo de la alcaldía, indicadores de polaridad (positivos, negativos) con una barra de progreso que indica la cantidad con su valor en porcentaje.

La gráfica de barras horizontal detalla los tópicos encontrados en la alcaldía los cuales fueron obtenidas en los puntos anteriores además de la polaridad de cada tópico (positivo, negativo) con valores en porcentaje.

La gráfica top de tendencias muestra los 5 hashtags más mencionados dentro del periodo de la alcaldía con la cantidad de veces que fueron mencionadas.
La gráfica de barras vertical muestra la cantidad de tópicos por año trascurrido en el periodo de la alcaldía y su polaridad estos agrupados a manera de pila.

La gráfica nube de palabras representa las palabras más frecuentes o las que más han sido escritos por los usuarios en los tweets, las palabras más frecuentes se pueden visualizar a una escala mayor, mientras que las menos frecuentes son de un tamaño menor.

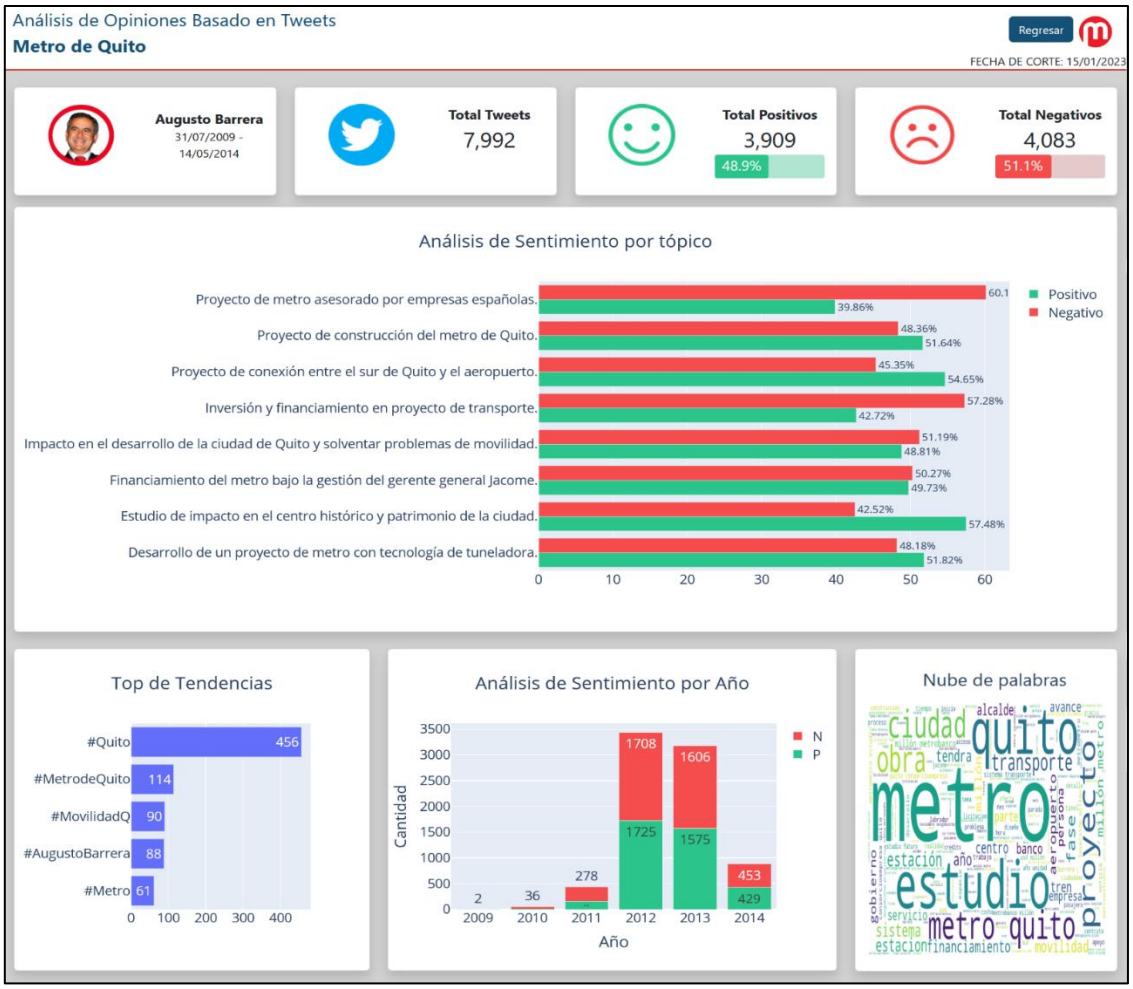

Figura 13. Dashboard del análisis de opiniones sobre el metro de Quito en la alcaldía de Augusto Barrera.

Figura 14. Dashboard del análisis de opiniones sobre el metro de Quito en la alcaldía de Mauricio Rodas.

Figura 15. Dashboard del análisis de opiniones sobre el metro de Quito en la alcaldía de Jorge Yunda.

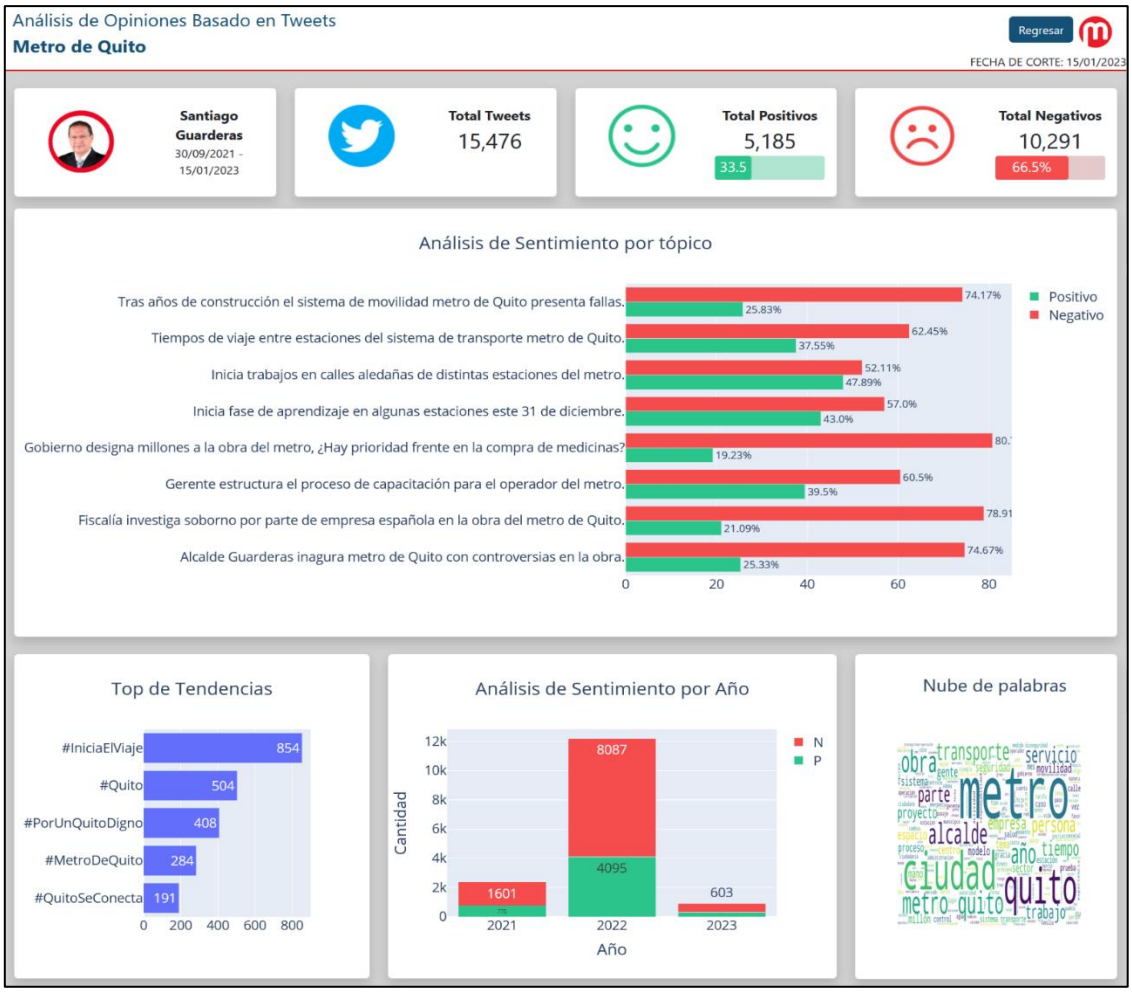

Figura 16. Dashboard del análisis de opiniones sobre el metro de Quito en la alcaldía de Santiago Guarderas.

Para visualizar de manera pública el dashboard del prototipo de sistema para el análisis de opiniones sobre el metro de Quito se encuentra alojado en Heroku, esta plataforma de servicios en la nube ofrece alojamiento gratuito, dominio y es de acceso libre lo cual permite que esté disponible en línea para cualquier persona que desee utilizarlo. Para visualizar se accede mediante el siguiente link. https://metro-quito.herokuapp.com/

3.4. Pruebas

Durante esta fase, se realizaron pruebas exhaustivas para identificar y corregir errores y defectos, con el fin de garantizar la calidad y la confiabilidad. Las pruebas permitieron verificar que el prototipo cumpla con los requisitos de usuario y que se ajuste a las especificaciones técnicas.

3.4.1. Pruebas de rendimiento

Las pruebas de rendimiento fueron realizadas para determinar la capacidad del software para manejar una carga determinada, capacidad de respuesta, medir su velocidad,

estabilidad y eficiencia. Las pruebas de rendimiento fueron basadas en el modelo de calidad de software FURPS y evaluadas con la herramienta Apache JMeter.

las pruebas fueron realizadas con un conjunto de 100 usuario los cuales forman un subgrupo de 10 usuarios y cada subgrupo ingresa cada segundo, estos usuarios están encargados de ingresar a la pantalla principal y al dashboard de manera simultánea.

Las pruebas que fueron ejecutadas se detallan en el ANEXO II. La tabla 27 y la figura 17 detallan de manera general los datos recopilados en esta prueba el cual demuestra que el 100% de casos fue exitoso.

Tabla 27. Pruebas de rendimiento.

Resultado pruebas de rendimiento		
Ejecutados	**Exitosos**	**Fallidos**
100	100	0

En la figura 17 se puede visualizar que las pruebas de rendimiento fueron exitosas en un 100%

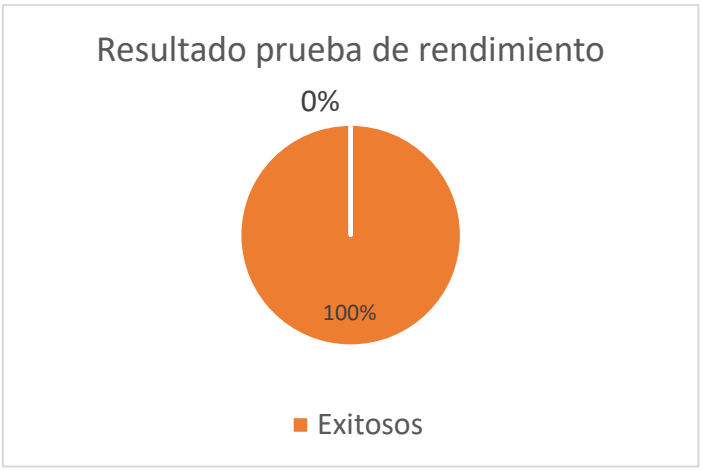

Figura 17: Resultado de pruebas de rendimiento

- **Análisis de los indicadores**

En la tabla 28 y la figura 18 se detalla la medición de eficacia realizada al prototipo de sistema de análisis de opiniones basado en tweets, alcanzando el 100% de las peticiones realizadas. Las peticiones realizadas se detallan en el ANEXO III.

Tabla 28. Dimensión eficacia.

Dimensión	Indicador	Peticiones
Eficacia	Número de peticiones realizada correctamente	200
	Promedio (%)	100%

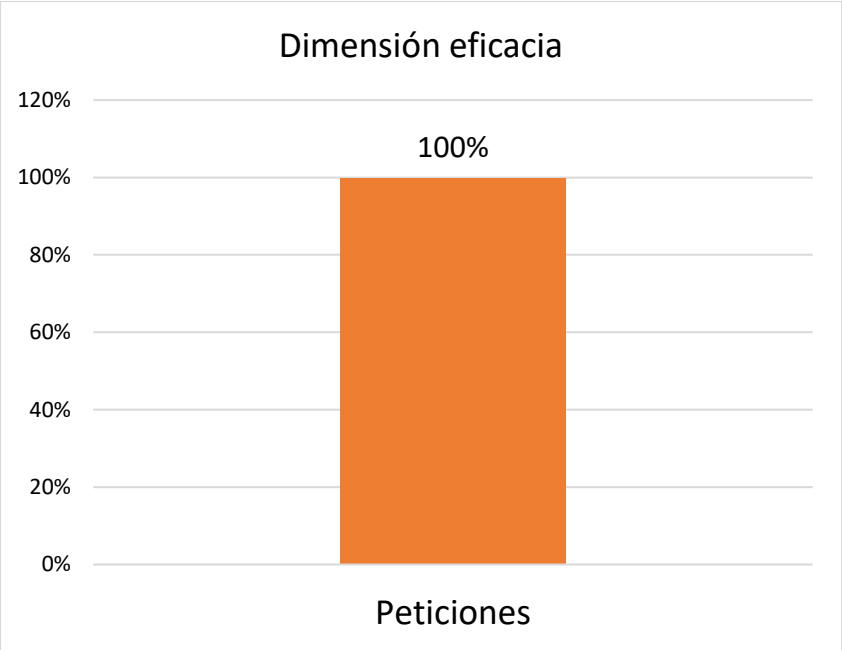

Figura 18: Dimensión eficacia.

- **Tiempo de respuesta**

La figura 19 detalla de manera general el promedio del tiempo de respuesta de las interacciones en milisegundos, con un numero de 100 usuarios conectados simultáneamente. Estas interacciones son detalladas en el ANEXO IV. Estos datos fueron calculados a partir de la siguiente fórmula.

$$n = \frac{1}{n}\sum_{i-1}^{n} a = \frac{a1 + a2 + a3 + a4 + a5 + \cdots an}{n}$$

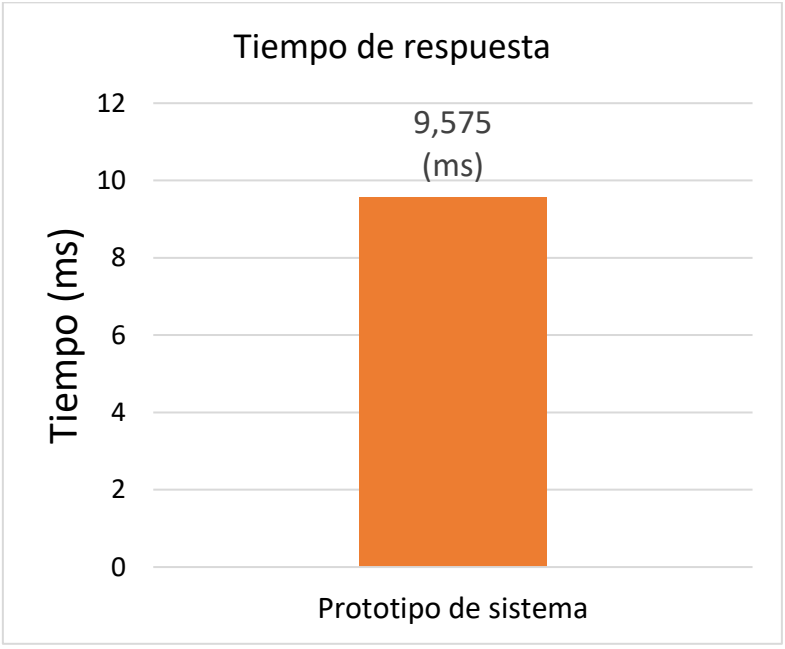

Figura 19: Tiempo de respuesta.

- **Utilización de recursos**

En la tabla 29 y la figura 20 podemos visualizar el consumo de recursos computacionales después de ejecutar las pruebas al prototipo de sistema de análisis de opiniones basado en tweets. El promedio de cada recurso usado de detalla en el ANEXO V.

Tabla 29. Uso de recursos.

Dimensión	Indicador	Prototipo de sistema de análisis de opiniones basado en tweets
Consumo de recursos	Promedio de uso de Memoria RAM %	44%
	Promedio de uso de CPU %	9%
	Promedio de uso de Disco %	2%

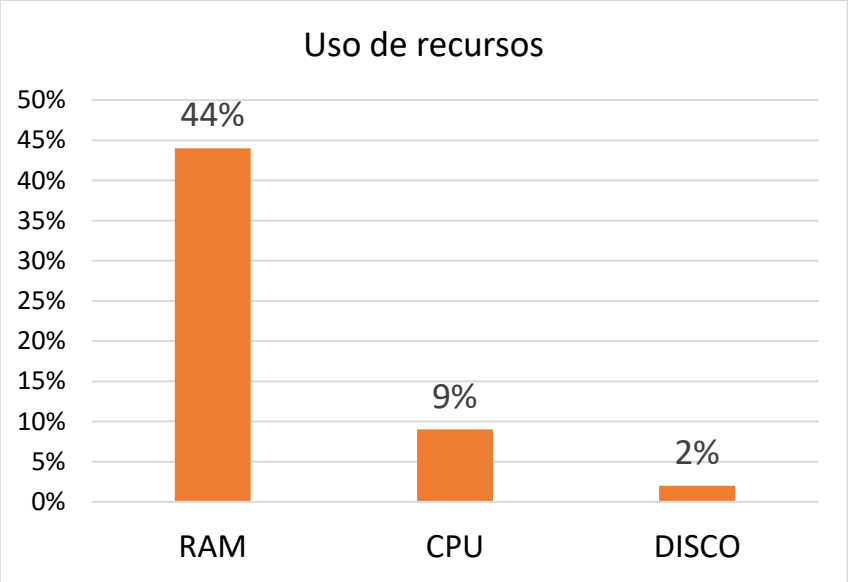

Figura 20: Uso de recursos.

3.4.2. Pruebas de usabilidad

El objetivo de las pruebas de usabilidad es comprender la experiencia del usuario al interactuar con el prototipo y asegurar que sea fácil de usar. Con estas pruebas se puede evaluar la facilidad de uso y eficacia del tablero en mostrar la información al usuario [47]. Las pruebas de usabilidad se realizaron mediante un formulario en la nube (Microsoft Forms) a 10 personas. Contine 6 preguntas las cuales fueron respondidas por 10 personas de la ciudad de Quito quienes navegaron y utilizaron el prototipo. Finalmente, los usuarios llenaron la encuesta. Las preguntas se enfocan en la usabilidad del prototipo y su forma de responder es en la escala de 1 a 5, en donde 1 totalmente en desacuerdo y 5 totalmente de acuerdo, las preguntas descritas a continuación:

- ¿Es fácil encontrar la información que se necesita en el prototipo?
- ¿Es fácil de usar el prototipo y las diferentes funcionalidades que ofrece?
- ¿La navegación del prototipo es intuitiva y coherente?
- ¿Las visualizaciones y gráficos del prototipo son claras y fáciles de interpretar?
- ¿El prototipo es atractivo y se ajusta a las expectativas de los usuarios?

- ¿El prototipo cumple con los objetivos y necesidades del usuario?

Los detalles de la encuesta se pueden visualizar en la figura 21. Los resultados de la encuesta indican que la mayoría de los usuarios encuentran el uso del prototipo sencillo y que la información es fácil de encontrar. Además, la navegación del prototipo se percibe como intuitiva y coherente por la mayoría de los encuestados. También se destaca que la mayoría de los usuarios consideran que el prototipo cumple con sus objetivos y necesidades. Estos resultados sugieren que el diseño y desarrollo del prototipo han sido efectivos y han cumplido con las expectativas de los usuarios. Los datos obtenidos a través de la encuesta serán útiles para mejorar el prototipo y asegurarse de que continúe cumpliendo con las necesidades y expectativas de los usuarios.

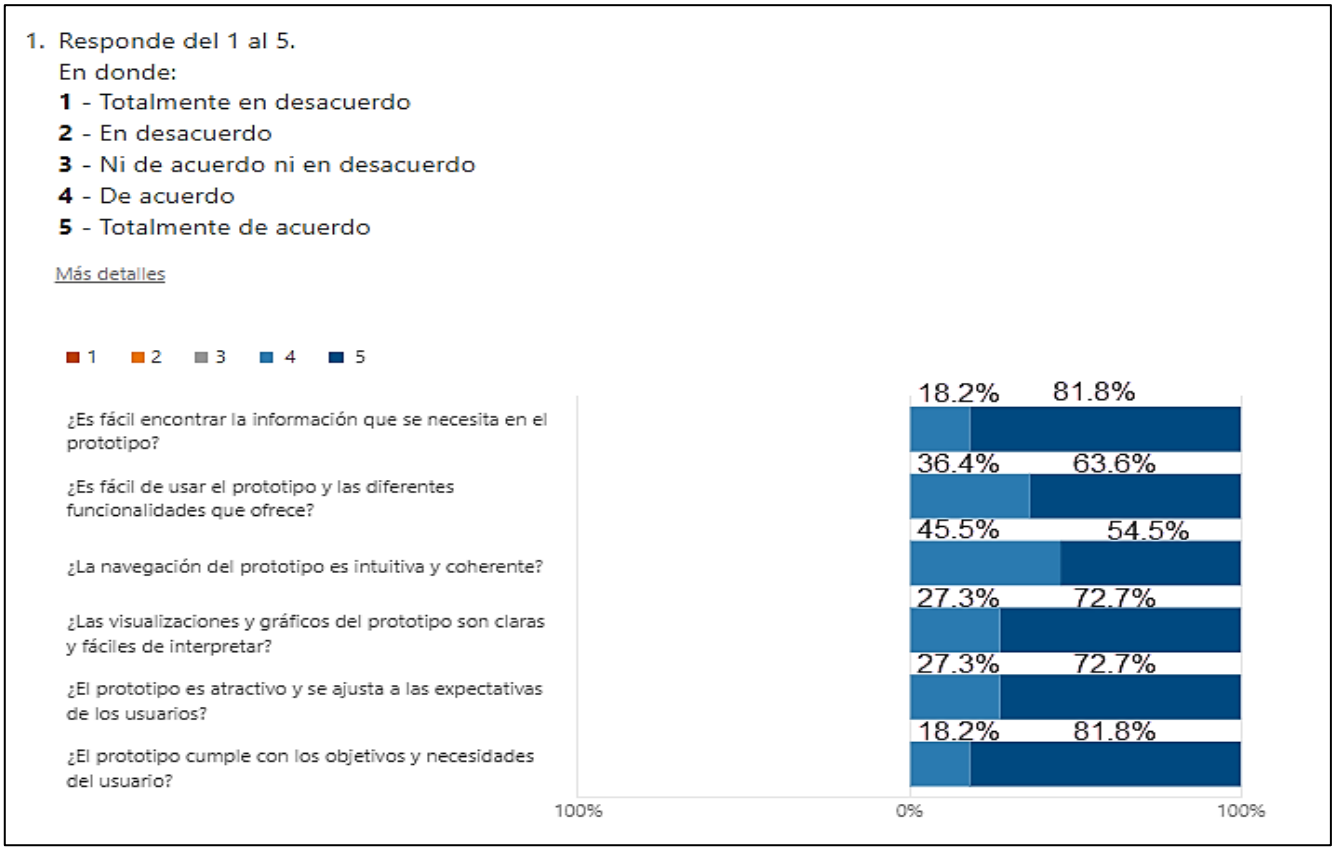

Figura 21: Resultados de encuesta prueba de usabilidad.

3.5. Mantenimiento

La fase final de la metodología en cascada tiene como objetivo garantizar que el prototipo de sistema de análisis de opiniones basado en tweets siga cumpliendo con los requisitos y expectativas del usuario después de su implementación. La fase se ejecutará cuando el sistema requiera un mantenimiento que implique: mejoras de rendimiento, solución de errores, la adaptación de un nuevo requerimiento por parte del usuario, mejorar la fiabilidad del sistema e incluso actualizaciones de las librerías utilizadas.

4 RESULTADOS

Luego de la recolección y limpieza de datos, se obtuvo un total de 87,035 tweets relacionados con el tema del metro de Quito en la red social Twitter. Al momento de capturar los datos se seleccionaron aquellos que únicamente menciona “metro de quito” o fue publicado desde la cuenta oficial @MetrodeQuito. Esto permitió tener una perspectiva de la población al momento de realizar el análisis de opiniones. Para tener un análisis comparativo, se decidió dividir el conjunto de datos por alcaldía como se muestra en la tabla 30.

Tabla 30. Resultado del análisis de opiniones sobre el metro de Quito por cada alcaldía.

Alcalde	Periodo mandato	Total tweets
Augusto Barrera	31/07/2009 - 14/05/2014	7,992
Mauricio Rodas	14/05/2014 - 14/05/2019	33,246
Jorge Yunda	14/05/2019 - 29/09/2021	15,476
Santiago Guarderas	30/09/2021 - 15/01/2023	15,476

En lo que respecta al periodo de alcaldía de Augusto Barrera, se observa que cuenta con el mayor porcentaje de tweets positivos. Además, se identificaron varios tópicos relevantes relacionados a los inicios de la construcción del metro, los cuales se indican a continuación.

- Estudio de impacto en el centro histórico y patrimonio de la ciudad.
- Financiamiento del metro bajo la gestión del gerente general Jácome.
- Impacto en el desarrollo de la ciudad de Quito y solventar problemas de movilidad.
- Inversión y financiamiento en proyecto de transporte.
- Proyecto de conexión entre el sur de Quito y el aeropuerto.
- Proyecto de construcción del metro de Quito.
- Proyecto de metro asesorado por empresas españolas.
- Impacto en el desarrollo de la ciudad de Quito y solventar problemas de movilidad.
- Inversión y financiamiento en proyecto de transporte.

En cuanto al periodo de Alcaldía de Mauricio Rodas, se puede apreciar tópicos relacionado con el avance de la obra del metro de Quito y algunos incidentes los cuales se indican a continuación:

- El costo del pasaje será sin subsidio.
- Obra del metro de Quito provoca cierres en las calles por la construcción de estaciones.
- Comparativa del metro de Quito con el problema de movilidad en Bogotá.
- Especulación sobreprecio de la obra, ¿Interfiere la fiscalía?
- Se establece logo del metro de Quito.
- Daño a la ciudad de Quito por construcción del metro.
- Proyecto de construcción del metro de Quito y su fuerte inversión.
- Construcción en curso de los túneles y estaciones.

- Desarrollo y mejora del sistema de transporte para la movilidad.
- Muerte de geólogo por accidente en las obras del metro.

Sobre el periodo de Alcaldía de Jorge Yunda, se puede destacar los desafíos que tuvo el proyecto a causa de la pandemia COVID-19. Al establecer estos tópicos se excluyó los que tenían relación directa con la pandemia, ya que no estaba en los objetivos del presente proyecto. Los tópicos encontrados fueron:

- Desarrollo y mejora del sistema de transporte para la movilidad.
- El modelo de gestión del servicio de metro enfrenta desafíos en la implementación de una tarifa.
- Sin espacios para el comercio en las estaciones del metro de Quito.
- metro de Quito influye en zonas aledañas a estaciones como parques, calles, espacios públicos.
- Se paraliza la obra del metro de Quito a causa de la pandemia COVID19.
- Se designa nuevo gerente del metro de Quito, sin alcalde por caso de corrupción.
- Aumento de costo de la obra, se aplaza la fecha de entrega.
- Se retoma obra del metro de Quito.

Tras la decisión del Consejo Metropolitano de Quito, Jorge Yunda es revocado de su puesto, posicionando como alcalde a Santiago Guarderas, en esta alcaldía los tópicos relevantes son los siguientes.

- Tiempos de viaje entre estaciones del sistema de transporte metro de Quito.
- Alcalde Guarderas inaugura metro de Quito con controversias en la obra.
- Fiscalía investiga soborno por parte de empresa española en la obra del metro de Quito.
- Gerente estructura el proceso de capacitación para el operador del metro.
- Inicia fase de aprendizaje en algunas estaciones este 31 de diciembre.
- Tras años de construcción del sistema de movilidad metro de Quito presenta fallas.
- Gobierno designa millones a la obra del metro, ¿Hay prioridad frente a la compra de medicinas?
- Inicia trabajos en calles aledañas de distintas estaciones del metro.

Otros resultados que se pudieron apreciar es los hashtags con mayor uso en las diferentes alcaldías. Con este análisis podemos establecer los hashtags con mayor número de mención.

Alcaldía Augusto Barrera, en su alcaldía también se visualiza que los años con más mención al metro de quito fue los años 2012 y 2013. En la tabla 31 se visualiza el top 5 de los hashtags con el mayor número de menciones.

Tabla 31. Tendencias del análisis de opiniones sobre el metro de Quito en la alcaldía de Augusto Barrera.

Hashtags	Número de mención
#Quito	456
#MetrodeQuito	114
#MovilidadQ	90
#AugustoBarrera	88
#Metro	61

Alcaldía Mauricio Rodas, en su alcaldía también se visualiza que los años con más mención al metro de quito fue los años 2017 y 2018. En la tabla 32 se visualiza el top 5 de los hashtags con el mayor número de menciones.

Tabla 32. Tendencias del análisis de opiniones sobre el metro de Quito en la alcaldía de Mauricio Rodas.

Hashtags	Número de mención
#ElMetroEstaLlegando	1331
#Quito	1211
#MetroDeQuito	1110
#Odebrecht	278
#Ecuador	227

Alcaldía Jorge Yunda, en su alcaldía también se visualiza que los años con más mención al metro de quito fue los años 2020. En la tabla 33 se visualiza el top 5 de los hashtags con el mayor número de menciones.

Tabla 33. Tendencias del análisis de opiniones sobre el metro de Quito en la alcaldía de Jorge Yunda.

Hashtags	Número de mención
#DisciplinaParaVolver	1299
#Quito	615
#MovilidadDeCalidad	553
#MetroCultura	335
#QuitoSolidario	329

Alcaldía Santiago Guarderas, en su alcaldía también se visualiza que los años con más mención al metro de quito fue los años 2022. En la tabla 34 se visualiza el top 5 de los hashtags con el mayor número de menciones.

Tabla 34. Tendencias del análisis de opiniones sobre el metro de Quito en la alcaldía de Santiago Guarderas.

Hashtags	Número de mención
#IniciaElViaje	1299
#Quito	615
#PorUnQuitoDigno	553
#MetroDeQuito	335
#QuitoSeConecta	329

4.1. Conclusiones

Se realizó una revisión bibliográfica de trabajos relacionados con la extracción de tópicos y análisis de sentimientos. En la literatura revisada se encontraron trabajos que abordaban cada proceso por separado. El presente trabajo tiene una perspectiva social y política al analizar los tweets relacionados con el metro de Quito, considerado eje articulador del sistema integrado de transporte público, lo que permitió obtener información valiosa sobre la percepción de la ciudadanía acerca de este importante proyecto de transporte y su impacto en la ciudad. La integración de la extracción de tópicos y análisis de sentimientos en este estudio permite un análisis más completo y enriquecedor de los datos recolectados convirtiéndose en una herramienta valiosa para comprender la percepción de los usuarios en las redes sociales.

La calidad y la integridad de los datos utilizados en el proceso son un factor importante para obtener resultados confiables en el análisis de datos. El proceso de preprocesamiento de datos ha demostrado la importancia de limpiar los datos antes de aplicar algoritmos, lo que permite una mejor comprensión, diagramación y toma de decisiones basadas en resultados. Por otro lado, es de gran importancia contar con datos de entrenamiento en español sin la necesidad de traducirlos al inglés para poder entrenar el modelo. De esta manera, se puede minimizar la pérdida de información y obtener análisis más relevantes y precisos. Sin embargo, es importante tener en cuenta que el idioma español presenta muchas variaciones que invitan a investigar más en el análisis de datos en español.

El proceso de creación de un dashboard efectivo no solo depende de la capacidad de manipular y visualizar datos, sino también es recomendable tener conocimiento sobre *storytelling*. La habilidad de contar una historia con los datos tiene relevancia para que el dashboard sea efectivo en el análisis de los resultados. También, es importante que al momento de diseñar el dashboard la información presentada sea comprensible y pueda transmitir un mensaje útil al usuario final.

4.2. Recomendaciones

Dado que la obtención de datos adicionales en los tweets fue limitada, se sugiere considerar la posibilidad de obtener información adicional, como la ubicación geográfica de los usuarios, para enriquecer el análisis. Esto permitiría la inclusión de nuevos gráficos en el dashboard y una narrativa más completa y precisa en el análisis de los datos.

Al seleccionar los temas o tópicos para los clústers, es recomendable tener en cuenta la relevancia de cada tema para el objetivo del análisis. Es decir, es importante elegir aquellos temas que sean relevantes para el problema que se está tratando de resolver. Además, es recomendable tener en cuenta la distribución de los datos en cada clúster, ya que algunos temas pueden tener más datos que otros. Por lo tanto, se recomienda realizar un análisis previo de los datos y de los temas posibles para determinar cuáles son los más relevantes y coherentes con la información que se desea extraer. Por último, es importante evaluar la calidad y precisión de los resultados obtenidos al

seleccionar los temas y clústers, ya que esto puede influir en la eficacia de las decisiones que se tomen a partir de los análisis realizados. Por lo tanto, se recomienda realizar una validación cruzada y comparar los resultados obtenidos con los datos originales para garantizar la calidad de la información obtenida.

Al aplicar storytelling en la creación de un dashboard, es fundamental definir un mensaje principal claro y conciso que se pueda transmitir a través de los datos. De esta manera, se evita perder el enfoque en la información más importante y se logra contar una historia coherente y efectiva. Además, es esencial utilizar visualizaciones efectivas, como gráficos y diagramas claros y fáciles de entender, que presenten la información de manera clara y coherente con el mensaje principal. Esto permitirá que el usuario final comprenda rápidamente la información presentada y se logren mejores resultados en el análisis de los datos.

REFERENCIAS BIBLIOGRÁFICAS

[1] Empresa Pública Metropolitana Metro de Quito, “Encuesta domiciliaria de movilidad (edm11) del distrito metropolitano de quito 13,” pp. 1–56, 2012.

[2] Municipio del Distrito Metropolitano de Quito, “Plan Metropolitano de Desarrollo y Ordenamiento Territorial 2015-2025: Componente Estratégico,” pp. 234–321, 2015, [Online]. Available: https://www.quito.gob.ec/documents/PMDOT.pdf

[3] K. Lucero, “Mientras el transporte público sea deficiente, el parque automotor seguirá engordando,” *revistagestion.ec*. p. 2, 2020.

[4] Periodista digital, “Quito: ¿Cuál es la causa de los embotellamientos?,” *ecuavisa*, Nov. 17, 2022.

[5] Carolina Ponce, “¿Cuándo y quién inició el proyecto del Metro de Quito?,” *GK*, 2019. https://gk.city/2019/09/05/quien-empezo-construccion-metro-quito/

[6] Empresa Pública Metropolitana Metro de Quito, “Proyecto Primera Línea Del Metro De Quito Marco De Politicas De Reasentamiento,” p. 4, 2013, [Online]. Available: https://www.metrodequito.gob.ec/wp-content/uploads/2018/01/Marco_de_Politicas_de_Reasentamiento.pdf

[7] metro de Quito, “Metrocultura,” Quito, 2020. Accessed: Apr. 17, 2023. [Online]. Available: https://metrodequito.gob.ec/wp-content/uploads/2021/01/Guia-de-Uso-Metro.pdf

[8] J. Mi. Moine, A. Haedo, and S. Gordillo, “Estudio comparativo de metodologías para minería de datos,” *XIII Workshop de Investigadores en Ciencias de la Computación*, pp. 278–281, 2011.

[9] U. de La Sabana, “El manejo de Datos por Internet,” *U. Sabana*. https://www.unisabana.edu.co/portaldenoticias/al-dia/el-manejo-de-los-datos-por-internet/

[10] N. Dávalos, “13 millones de personas tienen redes sociales en el Ecuador,” *PRIMICIAS*, 2020. https://www.primicias.ec/noticias/tecnologia/13-millones-personas-redes-sociales-ecuador/

[11] M. Jasso-Hernández, D. Pinto, D. Vilariño, and C. Lucero, “Análisis de sentimientos en Twitter: impacto de las características morfológicas,” *Research in Computing Science*, vol. 72, no. 1, pp. 37–45, 2014, doi: 10.13053/rcs-72-1-3.

[12] F. Gerrero-Solé and L. Mas-Manchón, “Estructura de los tweets políticos durante las campañas electorales de 2015 y 2016 en España,” *El Profesional de la Información*, vol. 26, no. 5, p. 805, 2017, doi: 10.3145/epi.2017.sep.03.

[13] M. T. R. Prabowo, “Sentiment analysis: A combined approach.,” *J Informetr*, pp. 123–157, 2009.

[14] J. B. Hollander and H. Renski, “Measuring Urban Attitudes Using Twitter,” 2015.

[15] N. Chaturvedi, D. Toshniwal, and M. Parida C, “Twitter to Transport: Geo-Spatial Sentiment Analysis of Traffic Tweets to Discover People’s Feelings for Urban Transportation Issues,” 2019.

[16] B. Qi, A. Costin, and M. Jia, “A framework with efficient extraction and analysis of Twitter data for evaluating public opinions on transportation services,” *Travel Behav Soc*, vol. 21, pp. 10–23, Oct. 2020, doi: 10.1016/j.tbs.2020.05.005.

[17] A. Moreno and C. A. Iglesias, “Understanding customers’ transport services with topic clustering and sentiment analysis,” *Applied Sciences (Switzerland)*, vol. 11, no. 21, Nov. 2021, doi: 10.3390/app112110169.

[18] H. N. Chua, A. W. Q. Liao, Y. C. Low, A. S. H. Lee, and M. A. Ismail, “Challenges of Mining Twitter Data for Analyzing Service Performance: A Case Study of Transportation Service in Malaysia,” 2022, pp. 227–239. doi: 10.1007/978-3-031-04216-4_21.

[19] k. Selcuk Candan, leman Akoglu, Xin Luna DOng, and Huan Liu, “WSDM ’22: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining,” New York, NY, USA: Association for Computing Machinery, 2022.

[20] M. Adedoyin-Olowe, M. M. Gaber, and F. Stahl, “A Survey of Data Mining Techniques for Social Network Analysis.” [Online]. Available: http://mashable.com/2012/06/22/data-created-every-minute/

[21] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed. Upper Saddle River: Prentice Hall, 2018.

[22] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, Dec. 2014, doi: 10.1016/j.asej.2014.04.011.

[23] C. C. Aggarwal, *Data Mining: The Texbook*. Cham: Springer International Publishing, 2015. doi: 10.1007/978-3-319-14142-8.

[24] W. Zhang, *Data pre-processing*. Data Mining and Knowledge Discovery for Geoscientists, 2014.

[25] C. Manning, P. Raghavan, and H. Schütze, “Introduction to Information Retrieval,” 2008.

[26] I. H. Witten, E. Frank, and M. A. Hall, “Data Mining: Practical Machine Learning Tools and Techniques,” 2016.

[27] M. Zubair Asghar, A. Khan, S. Ahmad, and F. Masud Kundi, “A Review of Feature Extraction in Sentiment Analysis Mitigation of DDOS attacks in DTNs View project personality recognition from textual content View project A Review of Feature Extraction in Sentiment Analysis,” *Article in Journal of Basic and Applied Research International*, vol. 4, no. 3, pp. 181–186, 2014, [Online]. Available: www.textroad.com

[28] J. Liu, Y. Wang, and Z. Han, “Effect of Text Preprocessing on Sentiment Classification for Chinese Micro-blogs,” *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1117–1126, 2018.

[29] P. Szymański, P. Rychlikowski, and P. Gawrysiak, “Lemmatization in Translation of Legal Texts – Translation Quality vs. Efficiency,” *Proceedings of the 17th International Conference on Artificial Intelligence and Law*, pp. 249–253, 2019.

[30] Srinivas Chakravarthy, “Tokenization for Natural Language Processing,” Jun. 2020, Accessed: Nov. 04, 2022. [Online]. Available: https://towardsdatascience.com/tokenization-for-natural-language-processing-a179a891bad4

[31] T. Inoue and S. Abe, “Fuzzy support vector machines for pattern classification,” in *IJCNN’01. International Joint Conference on Neural Networks. Proceedings (Cat. No.01CH37222)*, IEEE, pp. 1449–1454. doi: 10.1109/IJCNN.2001.939575.

[32] M. Ji and L. Carin, “Gibbs sampling for a Dirichlet process mixture model,” *Journal of Machine Learning Research*, 2007.

[33] J. Yin, Z. Wang, Q. Li, S. Wang, and S. Wu, “A new generative model for topic modeling based on a mixture of Dirichlet trees.,” *Inf Sci (N Y)*, vol. 502, pp. 96–109, 2019.
[34] K. Shu, S. Wang, H. Chen, and X. Li, “Toward a general-purpose model for clustering-based topic modeling,” *IEEE Trans Knowl Data Eng*, 2018.
[35] J. A. Gallardo, “CRISP-DM Metodologia para el Desarrollo de Proyectos de Mineria de Datos,” *EPB 603 Sistena del conocimiento*, 2006.
[36] IBM, *spss modeler*. 2022.
[37] Marq Martí, “Qué es el Web scraping? Introducción y herramientas,” Apr. 29, 2016.
[38] R. P.-R. and A. J. R.-R. J. M. Royo-Letelier, “Evaluating Sentiment Analysis Techniques in Spanish: A Case Study,” *the 2nd Workshop on Evaluation of Human Language Technologies for Iberian Languages*, pp. 52–58, 2017.
[39] H. , & G. E. A. He, “Learning from imbalanced data,” *IEEE Trans Knowl Data Eng*, pp. 1263–1284, 2009.
[40] Y. Zhang and B. Wallace, “A Sensitivity Analysis of (and Practitioners’ Guide to) Convolutional Neural Networks for Sentence Classification,” Oct. 2015.
[41] J. M. L.-R. A. M.-G. and R. V.-G. María del Pilar Salas-Zárate, “Topic modeling for text mining: a review,” *Expert Syst Appl*, 2019.
[42] D. M. Blei, A. Y. Ng, and J. B. Edu, “Latent Dirichlet Allocation Michael I. Jordan,” 2003.
[43] Micheal Heck, Sakriani Sakti, and Satoshi Nakamura, “A Dirichlet Mixture Model for Improved Unsupervised Document Clustering,” *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 155–164, 2014.
[44] Y. , & Z. Referencia: Zhang, “Topic modeling for quality prediction of online reviews: A comparative study of LDA and GSDMM,” *Inf Process Manag*, pp. 1256–1271, 2018.
[45] N. , & S. Japkowicz, “The class imbalance problem: A systematic study,” *Intelligent Data Analysis*, pp. 429–449, 2002.
[46] A. I. Khan, “Comparative study on so ware development methodologies,” *Database Systems*, vol. 5, pp. 37–64, 2014.
[47] M. , & N. J. Hertzum, “The Encyclopedia of Human-Computer Interaction,” *The Interaction Design Foundation*, 2019.

ANEXOS

ANEXO I: Encuesta requerimiento de diseño

Los requisitos de interfaz recopilados mediante la encuesta se aplican para asegurar que el prototipo de sistema de análisis de opiniones basado en tweets satisfaga las necesidades y expectativas de los usuarios.

La encuesta fue realizada por 22 personas de la ciudad de Quito. Consta de 7 preguntas y se puede responder en la escala del 1 a 5, en donde 1 es no me interesa y 5 muy interesado. Los resultados de la encuesta revelan el interés de los participantes en visualizar el análisis obtenido de manera organizada, que esta información sea pública. Así mismo, mayor interés en visualizar las temáticas más relevantes y el análisis de sentimiento (positivo/negativo). El resultado de esta encuesta se visualiza en la figura 22.

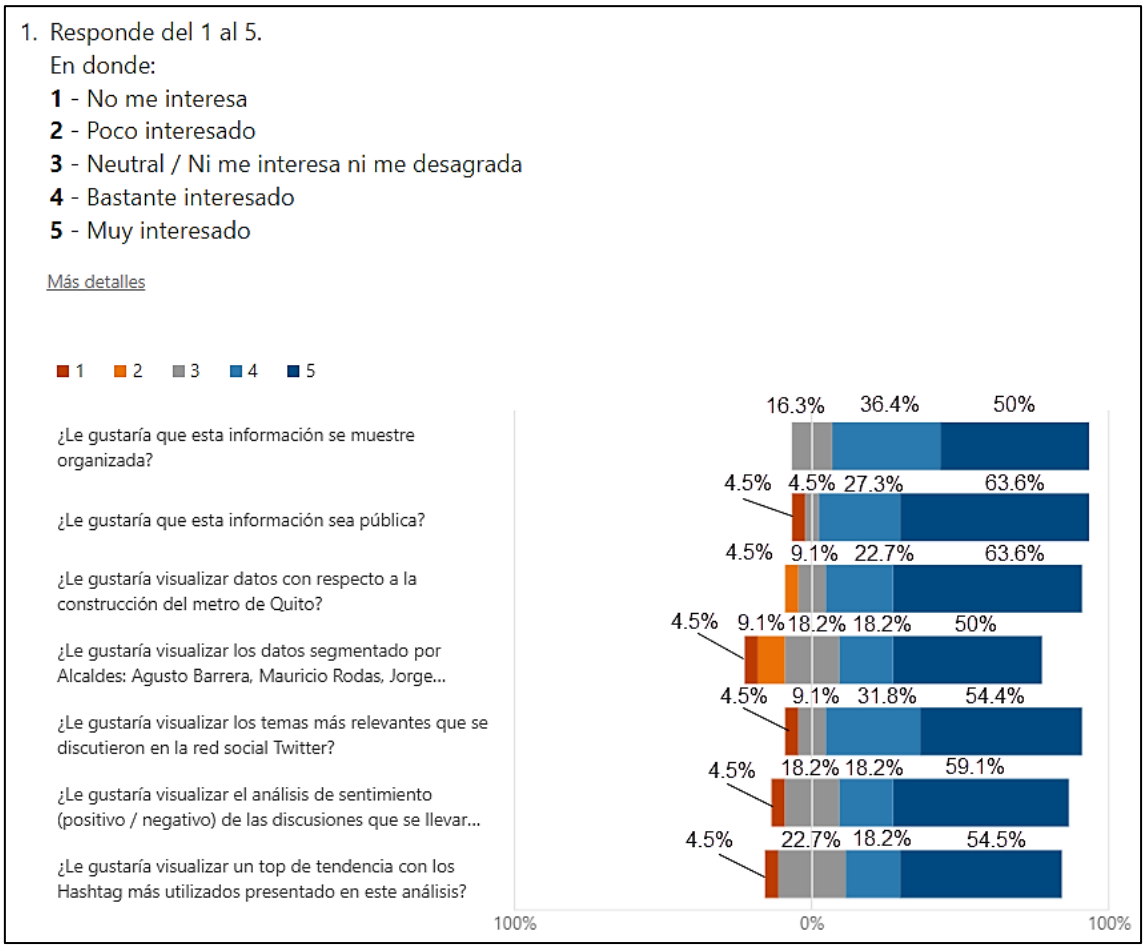

Figura 22: Resultados encuesta de requerimiento de diseño

ANEXO II: Resultado de pruebas de rendimiento del prototipo de sistema de análisis de opiniones basado en tweets

Las pruebas de rendimiento realizadas permitieron determinar el tiempo de respuesta del sistema y la capacidad máxima de rendimiento para minimizar riesgos de fallos y problemas futuros.

La figura 23 muestra el resultado de realizar 100 peticiones a cada pantalla del prototipo de sistema de análisis de opiniones basado en tweets. Las peticiones son realizadas de manera interactiva entre la pantalla principal y la pantalla dashboard, se puede observar que cada etiqueta (*label)* tiene 100 muestras (*samples)* el cual da un total de 200 muestras (*samples*), A*verage* nos muestra el tiempo promedio transcurrido de las 100 muestras (*samples*) por cada pantalla, obteniendo 14 milisegundos para la pantalla inicio y 11 milisegundos para la pantalla dashboard dando un promedio de 12 milisegundos, Min nos indica que el tiempo más bajo trascurrido para las 100 muestras (*samples*) se presentó en la pantalla dashboard con 7 milisegundos y 10 milisegundos para la pantalla inicio dando un total de 7 milisegundos, Max nos indica que el tiempo más largo trascurrido para las 100 muestras (*samples*) se presentó en la pantalla dashboard con 57 milisegundos y 34 milisegundos en la pantalla inicio dando un total de 57 milisegundos, la desviación estándar (Std. Dev.) nos indica que el tiempo transcurrido de la muestra (*sample*) de la pantalla inicio es 4.78 milisegundos y la pantalla dashboard 6.48 milisegundos dando un total de 5.88 milisegundos, finalmente el porcentaje de error nos muestra que 0.00 % de las etiquetas (*labels*) fueron fallidas es decir que todas se ejecutaron de manera exitosa.

Label ↓	# Samples	Average	Min	Max	Std. Dev.	Error %
Pantalla Inicio	100	14	10	34	4.78	0.00%
Pantalla Dashboard	100	11	7	57	6.48	0.00%
TOTAL	200	12	7	57	5.88	0.00%

Figura 23: Reporte resumen.

ANEXO III: Resultado de medición de eficacia del prototipo de sistema de análisis de opiniones basado en tweets.

La medición de la eficacia es fundamental para evaluar el desempeño de las 100 muestras y establecer las áreas de mejora y oportunidades de optimización, permitiendo que el prototipo de sistema de análisis de opiniones basado en tweets sea más efectivo y eficiente.

La figura 24 detalla que para cada pantalla se usaron 100 muestras (*samples*) los cuales realizaban peticiones de manera simultánea dando un total de 200 muestras, el porcentaje de error es de 0.00% lo cual indica que el prototipo de sistema de análisis de opiniones basado en tweets realiza de manera exitosa las peticiones a cada usuario.

Label	# Samples	Error %
Pantalla Inicio	100	0.00%
Pantalla Dashboard	100	0.00%
TOTAL	200	0.00%

Figura 24: Reporte medición eficacia.

ANEXO IV: Resultado de medición del tiempo de respuesta del prototipo de sistema de análisis de opiniones basado en tweets.

El resultado de medición del tiempo de respuesta mantiene un estatus de éxito, esto indica que el tiempo que tarda el sistema en responder a una solicitud del usuario es satisfactorio evitando así cuellos de botella y problemas de rendimiento.

La figura 25 especifica el detalle de cada muestra (*sample*) para la pantalla inicio y la pantalla dashboard dando un total de 200 muestras. *Star time* es el tiempo de inicio de cada muestra (*sample*), *thread name* indica que el grupo de subprocesos es 1 y que este grupo tiene un número de subproceso, ejemplo 1-2, en este caso 1 es el grupo de subproceso y 2 es el número de proceso, *Sample time* es el tiempo de la muestra (*sample*) medido en milisegundos, estatus nos indica que el estado fue exitoso para cada una de las muestras (*sample*).

Sample # ↑	Start Time	Thread Name	Label	Sample Time(ms)	Status
1	15:19:12.998	Pantalla inicio 1-1	Pantalla Inicio	13	
2	15:19:13.011	Pantalla inicio 1-1	Pantalla Dashboard	8	
3	15:19:13.104	Pantalla inicio 1-2	Pantalla Inicio	11	
4	15:19:13.117	Pantalla inicio 1-2	Pantalla Dashboard	8	
5	15:19:13.198	Pantalla inicio 1-3	Pantalla Inicio	10	
6	15:19:13.208	Pantalla inicio 1-3	Pantalla Dashboard	9	
7	15:19:13.296	Pantalla inicio 1-4	Pantalla Inicio	10	
8	15:19:13.307	Pantalla inicio 1-4	Pantalla Dashboard	9	
9	15:19:13.398	Pantalla inicio 1-5	Pantalla Inicio	11	
10	15:19:13.409	Pantalla inicio 1-5	Pantalla Dashboard	8	
11	15:19:13.497	Pantalla inicio 1-6	Pantalla Inicio	10	
12	15:19:13.507	Pantalla inicio 1-6	Pantalla Dashboard	8	
13	15:19:13.598	Pantalla inicio 1-7	Pantalla Inicio	10	
14	15:19:13.608	Pantalla inicio 1-7	Pantalla Dashboard	8	
15	15:19:13.698	Pantalla inicio 1-8	Pantalla Inicio	11	
16	15:19:13.709	Pantalla inicio 1-8	Pantalla Dashboard	8	
17	15:19:13.798	Pantalla inicio 1-9	Pantalla Inicio	10	
18	15:19:13.809	Pantalla inicio 1-9	Pantalla Dashboard	9	
19	15:19:13.897	Pantalla inicio 1-10	Pantalla Inicio	12	
20	15:19:13.909	Pantalla inicio 1-10	Pantalla Dashboard	8	
21	15:19:13.997	Pantalla inicio 1-11	Pantalla Inicio	11	
22	15:19:14.008	Pantalla inicio 1-11	Pantalla Dashboard	9	
23	15:19:14.098	Pantalla inicio 1-12	Pantalla Inicio	10	
24	15:19:14.109	Pantalla inicio 1-12	Pantalla Dashboard	8	
25	15:19:14.197	Pantalla inicio 1-13	Pantalla Inicio	11	
26	15:19:14.208	Pantalla inicio 1-13	Pantalla Dashboard	7	
27	15:19:14.298	Pantalla inicio 1-14	Pantalla Inicio	11	
28	15:19:14.309	Pantalla inicio 1-14	Pantalla Dashboard	8	
29	15:19:14.398	Pantalla inicio 1-15	Pantalla Inicio	11	
30	15:19:14.409	Pantalla inicio 1-15	Pantalla Dashboard	9	
31	15:19:14.498	Pantalla inicio 1-16	Pantalla Inicio	11	
32	15:19:14.509	Pantalla inicio 1-16	Pantalla Dashboard	9	
33	15:19:14.597	Pantalla inicio 1-17	Pantalla Inicio	11	
34	15:19:14.611	Pantalla inicio 1-17	Pantalla Dashboard	9	

Figura 25: Medición del tiempo de respuesta

La gráfica de líneas permite visualizar las solicitudes de cada muestra (*sample*) a lo largo del tiempo, nos permite comparar la velocidad de respuesta que tiene el sistema ante la solicitud del usuario para interactuar con la pantalla inicio y la pantalla dashboard.

La figura 26 muestra en el eje X el tiempo transcurrido para la ejecución de todas las muestras o procesos y en el eje Y los picos de respuesta del prototipo de sistema de

análisis de opiniones basado en tweets medido en milisegundos. Podemos observar que le toma más tiempo a la pantalla inicio responder a las peticiones mientras que en la pantalla dashboard el tiempo es más corto.

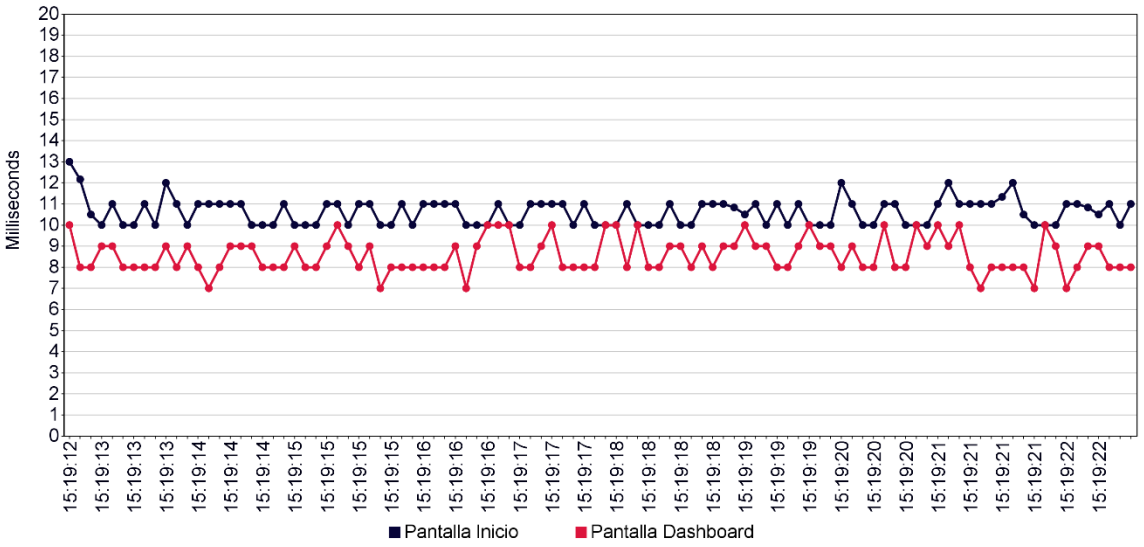

Figura 26: Medición del tiempo de respuesta – gráfico de líneas.

ANEXO V: Resultado de medición de uso de recursos computacionales.

La medición de los recursos computacionales, como la memoria, el procesador, el ancho de banda y el almacenamiento, permiten optimizar la capacidad del prototipo de sistema de análisis de opiniones basado en tweets e identificar los recursos que pueden estar siendo subutilizados o sobrecargados.

Los recursos computacionales del servidor utilizado se detallan en la tabla 35.

Tabla 35. Características del servidor

Características del servidor
Procesador Core i7 6ta generación
Memoria RAM 8GB
Disco SSD 500GB

El prototipo de sistema de análisis de opiniones basado en tweets usó el 44% de memoria RAM como de detalla en la figura 27.

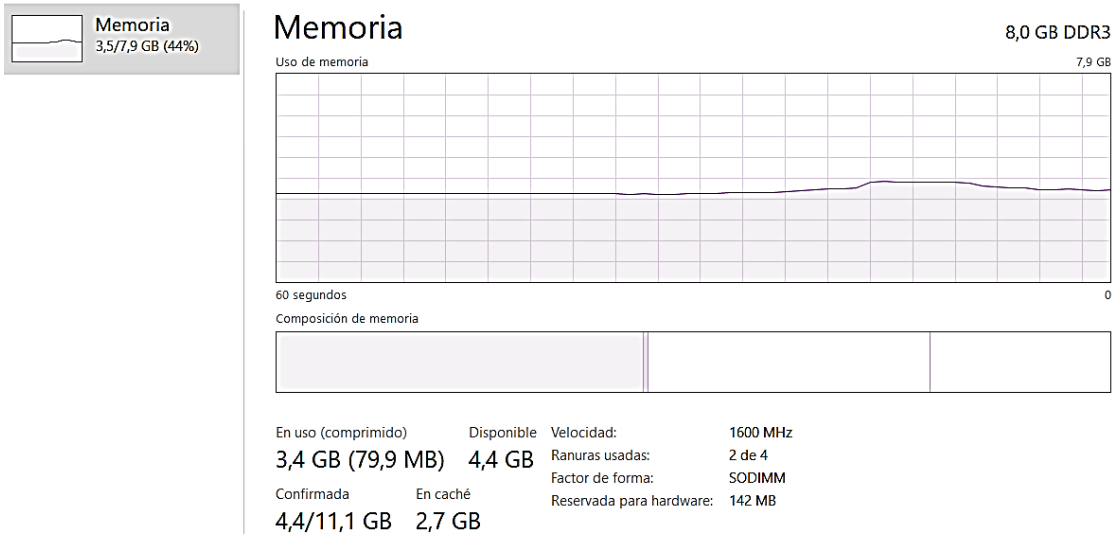

Figura 27: Promedio de uso de memoria RAM.

El prototipo de sistema de análisis de opiniones basado en tweets usó el 9% del procesador (CPU) como de detalla en la figura 28.

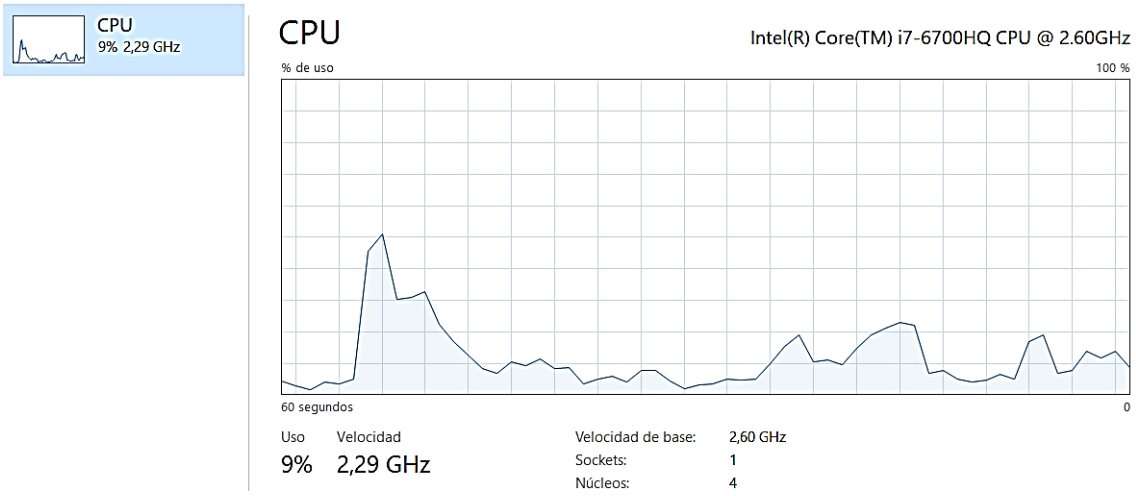

Figura 28: Promedio de uso de CPU.

El prototipo de sistema de análisis de opiniones basado en tweets usó el 2% del disco SSD como se detalla en la figura 29.

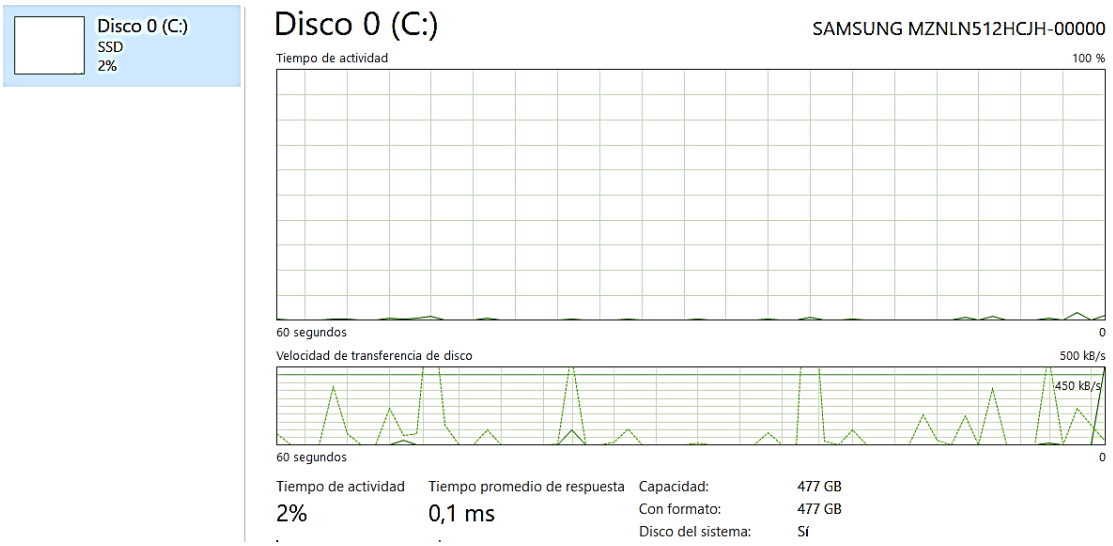

Figura 29: Promedio de uso de disco.

ANEXO VI: Repositorio Web.

Enlace al repositorio web: https://github.com/jorgequilumba/metrodequito

ANEXO VII: Servidor de Prueba del Sistema.

Enlace al servidor web: https://metro-quito.herokuapp.com/

ANEXO VIII: Manual de Usuario.

Enlace del manual de usuario: https://github.com/jorgequilumba/metrodequito/blob/main/Manual%20de%20usuario.pdf

ANEXO IX: Manual de Instalación.

Enlace del manual de instalación:
https://github.com/jorgequilumba/metrodequito/blob/main/Manual%20de%20instalaci%C3%B3n.pdf