Análisis de sentimientos y percepción de la seguridad en la ciudad de barranquilla

Cesar Caro
Departamento de Ingeniería de Sistemas
Universidad del Norte
Barrranquilla - Colombia
eccaro@uninorte.edu.co

Camilo Salgado Rada
Departamento de ingenieria de sistemas
Universidad del Norte
Barranquilla-Colombia
camilosalgado@uninorte.edu.co

Juan Pablo Uribe Pulgarin
Departamento de ingenieria de sistemas
Universidad del norte
Barranquilla - Colombia
jpuribe@uninorte.edu.co

***Abstract*—Este proyecto tuvo como objetivo el hacer un análisis de la percepción ciudadana expresada en las redes sociales. Dicha investigación fue realizada mediante distintos algoritmos de machine learning para analizar cual se comporta de la mejor manera para este tipo de análisis. Se realizó con el uso de una base de datos extraída con la ayuda de X (previamente Twitter) Developer API. Los resultados encontraron que hay una gran cantidad de personas que se sienten inconformes con la seguridad en la ciudad de Barranquilla, tal como lo expresan en sus publicaciones, y que el algoritmo más efectivo para este caso es el de regresión logística. Se concluyó que el uso de este tipo de herramientas permite conocer eficazmente el panorama actual de la sensación de seguridad en la ciudad de Barranquilla.**

Palabras clave—Análisis de sentimientos, machine learning, seguridad, aprendizaje supervisado

I. INTRODUCCIÓN

La percepción de la seguridad ciudadana es un factor crucial para la calidad de vida y el desarrollo sostenible de cualquier urbe. En el contexto de la ciudad de Barranquilla, saber y comprender cómo los ciudadanos evalúan su seguridad resulta fundamental para diseñar e implementar políticas públicas efectivas. Sin embargo, los métodos tradicionales de medición, como las encuestas, pueden presentar limitaciones en términos de alcance y vigencia (Suhaimin et al., 2023).

El análisis de sentimientos en redes sociales ha emergido como una herramienta poderosa para entender la opinión pública sobre muchos temas, donde la seguridad ciudadana no es la excepción (Yue et al., 2022). X (antes Twitter), se ha convertido en una fuente invaluable de datos para este tipo de análisis, por su facilidad de acceso a los datos y la gran cantidad de información que genera en tiempo real (Muhammed et al., 2022). Diversos estudios han demostrado la utilidad del análisis de sentimientos utilizando los datos de Twitter para monitorear la percepción de seguridad en diferentes contextos geográficos y temporales (Yang et al., 2022). Según J.F. et al. (2023), "el análisis de sentimientos ha demostrado su utilidad en la predicción de eventos delictivos. Asimismo, esta técnica ha sido empleada para evaluar el impacto de políticas públicas en la percepción de seguridad ciudadana" (Lidia & Sabar, 2021).

La mayoría de las investigaciones en este campo utilizan técnicas de aprendizaje automático y Deep learning para clasificar la polaridad de los tweets y extraer información relevante (T. L. Bene, et al, 2023). Este proyecto aborda la necesidad de un análisis dinámico y en tiempo real de la percepción de seguridad en barranquilla. A través del análisis de sentimientos en Twitter, se busca capturar lo que sienten los ciudadanos y conocer las perspectivas sobre sus preocupaciones y experiencias en materia de seguridad. (Plata D., et al, 2023)

Finalmente, analizaremos los datos obtenidos de todos los métodos para finalmente escoger el mejor modelo planteado, y ver finalmente la percepción que tiene la gente sobre la seguridad en Barranquilla en las redes sociales junto con las palabras asociadas a estos sentimientos

II. PROBLEMA

En la ciudad de Barranquilla, la percepción de seguridad de los ciudadanos ha sido un tema de gran relevancia, influenciando no solo la calidad de vida de sus habitantes, sino también el desarrollo económico y social de la región. A pesar de los esfuerzos por medir y mejorar la seguridad mediante métodos tradicionales, como encuestas y estadísticas oficiales, estos métodos presentan limitaciones significativas en cuanto a alcance y capacidad para reflejar la realidad dinámica y cambiante de la percepción ciudadana.

El análisis de sentimientos en redes sociales, particularmente en plataformas como X (anteriormente conocida como Twitter), ha emergido como una herramienta poderosa para capturar de manera más precisa y en tiempo real las preocupaciones, miedos y expectativas de los ciudadanos respecto a la seguridad en su entorno. Estudios recientes han demostrado que el análisis de grandes volúmenes de datos no estructurados, como los comentarios en redes sociales, puede ofrecer una visión más matizada y actualizada de la percepción pública, superando así las limitaciones de las encuestas tradicionales (Wang et al., 2022; Suhaimin et al., 2023).

Sin embargo, aunque la aplicación del análisis de sentimientos ha sido explorada en diversos contextos, como la evaluación de la percepción pública sobre políticas de

confinamiento (Yue et al., 2022), la implementación de ciudades inteligentes (Ahmed, s.f.) y el impacto del entorno urbano en el bienestar de los ciudadanos (Yang, Duarte, & Ciriquián, 2022), sigue existiendo una brecha en la aplicación específica de estas técnicas para monitorear y entender la percepción de seguridad en ciudades como Barranquilla. Este proyecto busca llenar este vacío, proponiendo un modelo de análisis de sentimientos que permita identificar patrones y tendencias en la percepción de la seguridad.

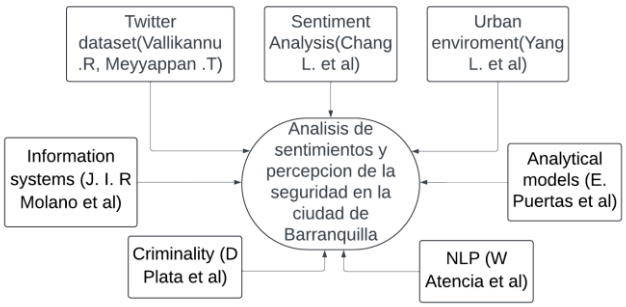

Figura 1. Árbol descriptivo del problema.

III. OBJETIVOS

1. Objetivo general: Desarrollar un modelo de análisis sentimental basado en machine learning que permita evaluar la percepción de los habitantes de Barranquilla sobre la seguridad en la ciudad, utilizando datos obtenidos a través de APIs de redes sociales para identificar patrones y tendencias que apoyen la toma de decisiones en políticas públicas de seguridad.
2. Objetivos específicos:

- Elaborar la revisión sistemática de la literatura relacionada con el análisis de sentimientos y el uso de machine learning en el contexto de la percepción de seguridad urbana.
- Desarrollar el modelo y el diseño de la solución para el análisis sentimental.
- Desarrollar el prototipo de la solución de análisis sentimental.
- Validar el prototipo de la solución de análisis de sentimientos, mediante el uso de las métricas para problemas de clasificación.

IV. METODOLOGIA

Debido a que el problema propuesto para este proyecto cae en el área de análisis de datos, se ha propuesto el uso de una metodología CRISP-DM,

a) Entendimiento del problema:

Antes de iniciar el desarrollo se debe dedicar tiempo a explorar las expectativas del proyecto con respecto a la minería de datos. Consultas sobre investigación de análisis de sentimientos aplicado a la seguridad en ciudades y, por otro lado, propuestas de soluciones familiares a la propuesta.

b) Entendimiento de los datos:

La fase de entendimiento es esencial para evitar inconvenientes inesperados durante la preparación de los datos, esto implica un análisis detallado de los datos que están disponibles para minería.

c) Preparación de los datos:

La preparación de los datos es la fase que generalmente toma más tiempo de la minería de datos, esta fase puede verse acelerada y disminuir el número de problemas con un buen desarrollo de las dos fases anteriores. En este caso, es importante realizar la limpieza de los datos de texto recolectados.

d) Modelado:

A partir de este punto, todo lo hecho anteriormente comienza a tener forma, en esta sección se crearán varios modelos para ser entrenados bajo ciertos parámetros definidos por nosotros, para esto se usará la librería de python scikit-learn.

e) Evaluación:

La evaluación de lo obtenido de las fases anteriores es un paso crucial, aquí es donde se define si lo desarrollado es verdaderamente efectivo contra la problemática que fue planteada inicialmente, si lo extraído de los datos representa verdaderamente el sentimiento de la comunidad expresado a través de redes sociales, en este caso, X (anteriormente Twitter). Esta fase demuestra si se deben realizar ajustes al modelo planteado y cambiarlo en ese caso, para luego ser evaluado nuevamente.

f) Despliegue:

Una vez se tengan los datos evaluados, sin presencia de algún sesgo u otro problema de depuración, se procede con el despliegue. En este caso el despliegue consiste en mostrar mediante el uso de gráficos y métricas el desempeño del modelo a los usuarios.

V. JUSTIFICACIÓN

Entender cómo perciben los ciudadanos la seguridad en su entorno urbano es de gran importancia, centrándonos en la ciudad de Barranquilla, la percepción de seguridad no solo influye en la calidad de vida de los habitantes, sino que también impacta en el desarrollo económico y social de la ciudad. Chang L. et al (2021) nos afirman que los métodos tradicionales de medición, como las encuestas, a menudo son limitados en alcance y no siempre reflejan el sentir inmediato de la población debido a su carácter estático y ocasional, además que, por otro lado, el análisis de sentimientos en redes sociales, particularmente en Twitter, ofrece una alternativa dinámica y en tiempo real para captar la percepción ciudadana.

Este enfoque permite monitorear de manera continua y más precisa las preocupaciones y sentimientos de los ciudadanos sobre la seguridad, Abdou & Benelallam (2023) mencionan que cuando las personas se sienten afectadas tienden a expresarse a través de medios de comunicación, particularmente en redes sociales, proporcionando datos valiosos que pueden ser utilizados para diseñar políticas públicas más efectivas y ajustadas a las necesidades reales de la población.

Finalmente, podremos ver qué palabras se asocian a cada sentimiento, comprendiendo el qué actividades o cosas son aquellas que presentan más problemas alrededor de la inseguridad en la ciudad. Este contraste es crucial para entender cómo se interpreta la seguridad a nivel ciudadano ya que, según Muhammed S. E. et al (2022), examinar y analizar data sets de redes sociales a gran escala es crucial para tanto las empresas como para los gobiernos, asegurando que las políticas públicas y privadas no solo sean efectivas en términos técnicos, sino también en la percepción y aceptación social.

VI. Marco Teórico

La percepción de la seguridad es un factor crucial que influye en la calidad de vida de los ciudadanos, la inversión, el turismo y el desarrollo de una ciudad. (J. I. R. Molano et al., 2023) Este marco teórico explora los fundamentos conceptuales y metodológicos del análisis de sentimientos, su aplicación específica en la medición de la percepción de la seguridad en Barranquilla a través de las redes sociales, y las consideraciones éticas y metodológicas relevantes.

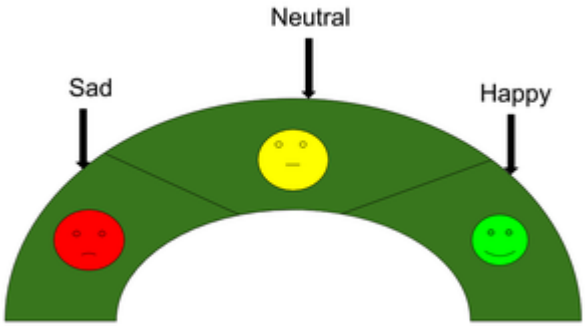

Figura 2. Sentimiento

La percepción de la seguridad es crucial para la confianza ciudadana, la inversión, el turismo y el desarrollo económico en Barranquilla. Las redes sociales se han convertido en una plataforma importante donde los ciudadanos comparten sus opiniones y experiencias sobre la seguridad. (W Atencia et al., 2020) El análisis de sentimientos en estas plataformas permite monitorear y medir la percepción de la seguridad en tiempo real. Las métricas clave para evaluar esta percepción incluyen el volumen de menciones relacionadas con la seguridad, el sentimiento general expresado, la identificación de áreas o situaciones percibidas como inseguras, y el impacto de usuarios influyentes y líderes de opinión en la percepción general. (Said-Hung et al., 2021)

El análisis de sentimientos enfrenta varios desafíos, como la detección del sarcasmo, la ironía y el lenguaje coloquial, así como las variaciones culturales. Estos desafíos requieren abordar los sesgos algorítmicos y asegurar la transparencia y equidad en el proceso de análisis. Además, es esencial respetar la privacidad de los usuarios de redes sociales y cumplir con las regulaciones de protección de datos, obteniendo el consentimiento informado y utilizando la información de manera ética. El contexto local de Barranquilla también debe ser considerado al interpretar los resultados, ya que factores culturales, sociales y económicos pueden influir en cómo se expresan las opiniones sobre la seguridad. (D Plata et al., 2023)

El análisis de sentimientos ofrece una herramienta valiosa para comprender la percepción de la seguridad en Barranquilla a través de las redes sociales. Este marco teórico establece una base sólida para futuras investigaciones y aplicaciones prácticas en el campo del análisis de sentimientos, proporcionando puntos de vista que pueden contribuir a la toma de decisiones informadas para mejorar la seguridad y la calidad de vida en la ciudad. El uso adecuado de estas técnicas permitirá una mejor comprensión de las preocupaciones ciudadanas y facilitará el desarrollo de estrategias efectivas para abordar los problemas de seguridad. (W Atencia et al., 2020)

La arquitectura extremo a extremo para el análisis de sentimientos sobre la percepción de seguridad en Barranquilla comienza con la recopilación de datos de redes sociales mediante APIs, seguido del preprocesamiento, que incluye limpieza y normalización del texto. Luego, se aplica el análisis de sentimientos utilizando métodos basados en léxico o aprendizaje automático, para clasificar el sentimiento en positivo, negativo o neutral. Los resultados son visualizados mediante herramientas de análisis de datos para identificar patrones clave y retroalimentar el proceso con ajustes según las interpretaciones obtenidas. (E. Puertas et al 2022)

El foco de tu investigación es medir y analizar la percepción de la seguridad en Barranquilla utilizando técnicas de análisis de sentimientos en redes sociales. El objetivo principal es identificar cómo los ciudadanos expresan sus opiniones sobre la seguridad, qué factores son más discutidos (delincuencia, presencia policial, etc.), y cómo varía la percepción según diferentes eventos o contextos (económicos, sociales). Esta investigación permitirá generar recomendaciones para mejorar la seguridad y la confianza ciudadana. (Suhaimin, M. S. M et al, 2023)

El análisis de sentimientos ha sido ampliamente utilizado para estudiar la percepción en temas de política, mercadeo y seguridad. A nivel internacional, proyectos similares han sido implementados en lugares como Estados Unidos, África, China, España, Oman, Portugal, etc.; donde las redes sociales son utilizadas para medir la percepción de los ciudadanos sobre la seguridad y el crimen. (A. M. Naira et al, 2023, Yue, A et al, 2022, Yang, L et al, 2022, V. Ramanathan et al, 2019, J. I. R. Molano et al, 2023)

En Colombia, estudios recientes han abordado la relación entre seguridad y redes sociales en ciudades como Bogotá y Medellín, donde se han identificado patrones de miedo y

zonas de riesgo basados en el análisis de sentimientos. Sin embargo, hay poca literatura enfocada específicamente en Barranquilla, lo que resalta la novedad y relevancia de tu estudio. (D Plata et al, 2023)

Para la realización del prototipo y lograr los objetivos de este proyecto, se puede llevar a cabo a través de diferentes metodologías y tecnologías:

- **Métodos basados en léxico**: Usan diccionarios de palabras clasificadas por sentimiento (positivo, negativo, neutral). Ejemplo: SentiWordNet. (T. L. Ben et al, 2023)
- **Métodos de aprendizaje automático**: Entrenan modelos predictivos para clasificar el sentimiento de nuevos textos. Algoritmos comunes incluyen Naive Bayes, SVM, y Redes Neuronales. (J. F. Sossa Rojo et al, 2023)
- **Técnicas híbridas**: Combinan métodos basados en léxico y aprendizaje automático, mejorando la precisión del análisis. (GeeksforGeeks, 2024)
- **Análisis de aspecto**: No solo analiza el sentimiento general, sino también aspectos específicos relacionados con la seguridad, como delitos, violencia, o presencia policial. (D Plata, 2023)

La metodología, tecnologías e implementación usados para este proyecto, se encuentran en el apartado de prototipo.

VII. Marco Conceptual

1. Análisis de sentimientos

La clasificación de textos por medio del proceso de análisis de sentimiento busca el análisis de opiniones de las personas, de manera que se concluya si es de característica positiva, negativa o neutra. El análisis de sentimiento no está centrado únicamente en la polaridad, sino también en las emociones (triste, feliz, enfadado) a la cual se llega por medio de algoritmos de procesamiento de lenguaje natural. Los resultados de este proceso pueden ser usados por empresas y gobiernos en la toma de decisiones que involucren al ciudadano.

2. Preprocesamiento de Datos

Si necesitamos procesar datos para extraer alguna información de ellos el preprocesamiento de estos es de suma importancia. Tal como no comerías sin antes lavar la comida, cortar y cocinar, este proceso es esencial para la ciencia de datos. Básicamente implica limpiar, transformar y organizar los datos antes de ser analizados, asegurándonos que estén libres de errores e inconsistencias.

Algunas de las técnicas que encontramos comúnmente en trabajos relacionados con el análisis de sentimientos son los siguientes:

- **Tokenización:** Dividir el texto en unidades significativas como palabras o frases.
- **Eliminar información irrelevante**, como emojis, etiquetas HTML, caracteres especiales y entre otros.
- **Lematización y Stemming:** Reducción de las palabras a su forma raíz para simplificar el análisis.
- **Eliminación de stop words:** Filtrar palabras comunes que no aportan valor semántico significativo.

3. Machine learning

El machine learning es el campo de estudio que le da a los computadores la capacidad de aprender cosas que no les han sido programadas específicamente, es una de las tecnologías más importantes en los últimos tiempos. Así como sugiere su nombre, les da a los computadores el poder de aprender, una habilidad que distingue los humanos especialmente. Este tipo de tecnologías se aplican actualmente en numerosos ámbitos, incluso en algunos en los que no se esperaría.

4. Modeling

La fase de modelado para el modelo CRISP-DM es cuando el proceso de machine learning ocurre. Para construir un rango de predicción de modelos se utilizan distintos algoritmos de machine learning, de los cuales se escoge el mejor modelo para desplegar.

5. Evaluation

Es necesario que los modelos estén completamente evaluados y se demuestre que cumplen acordemente para resolver el problema antes de ser desplegados para su uso dentro de una organización. Durante esta fase de CRISP-DM se cubre toda la evaluación requerida para mostrar que el modelo de predicción será capaz de hacer predicciones precisas luego de ser desplegado.

6. Deployment

La última fase de CRISP-DM cubre todo el trabajo que se necesita para integrar acordemente procesos de machine learning en una organización, dado que los modelos están hechos para servir un propósito específico dentro de esta.

VIII. Arquitecturas

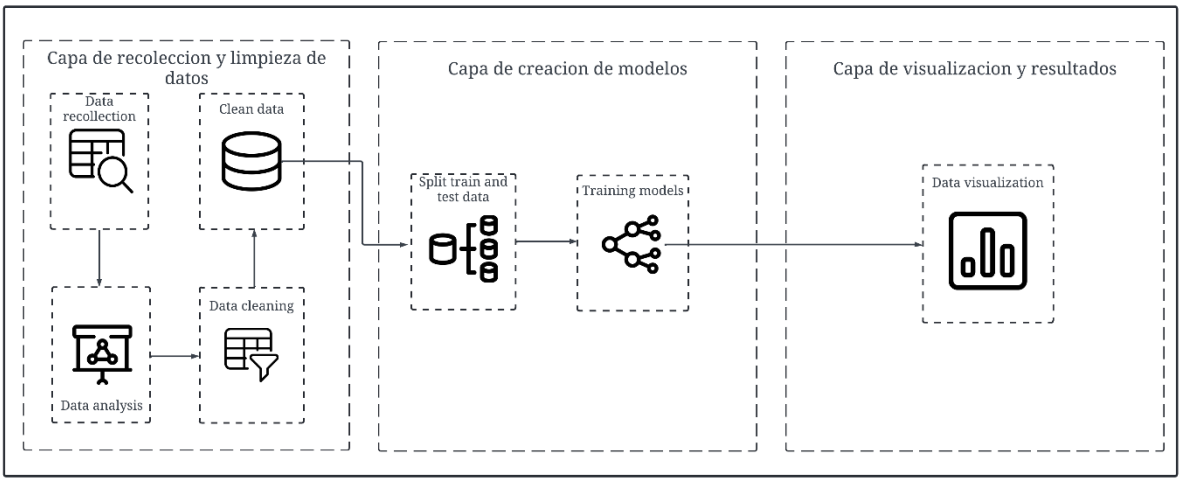

Figura 3. Diagrama de la arquitectura lógica de la solución

La figura 3, muestra la arquitectura lógica planteada para la solución del problema, dividiendo por capas los procesos que se llevaran a cabo, Recolección/limpieza de datos, creación de modelos y visualización de resultados. Es una solución optimizada que busca, en primer lugar, obtener los datos de trabajo que serán utilizados para crear y entrenar los modelos, modelos que utilizaremos para realizar el proceso de análisis de sentimientos y obtener los resultados deseados.

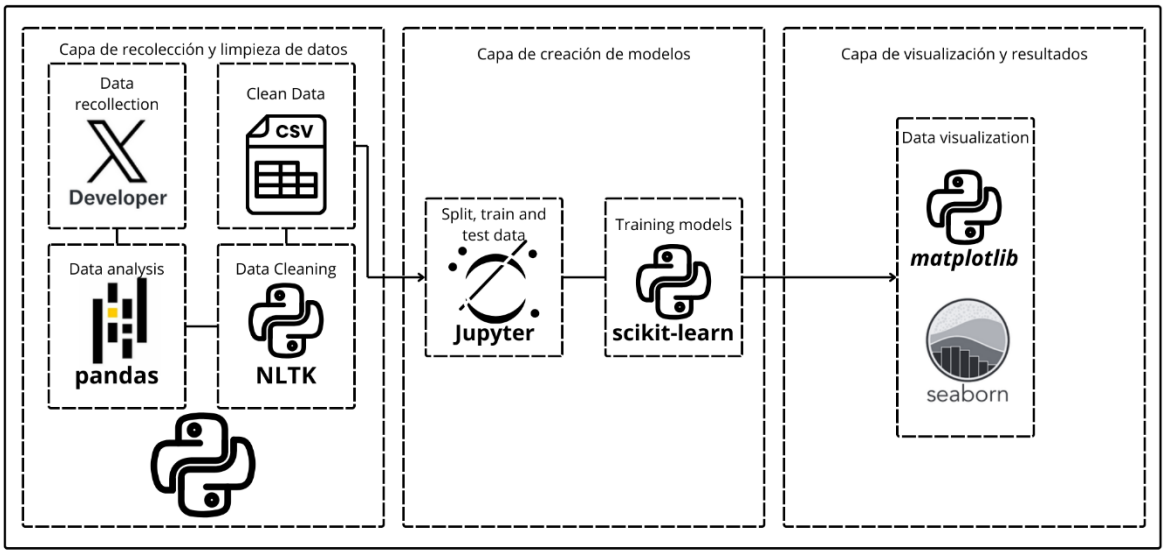

Figura 4. Diagrama de la arquitectura física de la solución

La figura 4, muestra la arquitectura física de la solución donde, además de la estructura general del proyecto, se incluye las tecnologías usadas en el para llevarlo a cabo. En primer lugar, la recolección de datos se realizará mediante el uso de X Developer API, que permite la recolección de tweets dados ciertos parámetros, así podemos recolectar información para la base de datos que sea acorde al proyecto para almacenarla en un CSV. Esta información será limpiada por medio de varios recursos de la librería NLTK, tales como, SnowballStemmer, para aplicar Stemming; stopwords, para importar lista de estas y word_tokenize, para la tokenización de las palabras. Esto se realiza para evitar redundancias y errores que tiñan la imparcialidad del análisis. En segundo lugar, con los datos procesados, se procede al entrenamiento de los modelos con el uso de las librerías de scikit-learn. Durante el entrenamiento de modelos mediremos las métricas con estas mismas librerías, las cuales nos darán como resultado qué modelo es más apto para nuestro proyecto. Por último, utilizamos seaborn y matplotlib para tener evidencia gráfica de cómo se comporta cada modelo a comparación de los demás, lo que nos facilitará el análisis y resultados del prototipo.

IX. PROTOTIPO

El prototipado consiste en un código funcional que tanto analice el sentimiento como mostrar gráficamente los resultados de estos análisis a nivel porcentual entre los sentimientos, de un dataset provisto y extraído previamente de X (previamente Twitter).

Inicialmente importamos las librerías que utilizaremos tanto para la realización del análisis como para la impresión de resultados a manera de gráficos que faciliten su comprensión. Gráficos de tipo circular, tablas de confusión, wordclouds entre otros.

Pandas se implementa para la manipulación de los datos que vamos a utilizar. Para entrenar el modelo necesitamos de usar un porcentaje de la base de datos como prueba, para esto utilizamos `train_test_split` de `sklearn.model_selection`. Por otra parte, necesitamos convertir el texto a parámetros numéricos para que puedan ser analizados por el modelo, para eso utilizamos `CountVectorizer` y `TfidfTransformer`.

A. Explicación de los datos:

Los datos que vamos a usar son textos extraídos de X (previamente Twitter), los cuales son strings que contienen una publicación realizada por un usuario de Twitter. Al Twitter ser una red social las publicaciones generalmente expresan una opinión sobre un tema en específico, este tema podemos elegirlo al momento de extraer los datos, por lo cual podemos elegir el tema que unirá la base de datos, en este caso la seguridad en Barranquilla.

Los datos se almacenarán en un CSV que inicialmente contendrá una columna, la que contiene el tweet a analizar, posteriormente a este CSV se le añadirá una nueva columna que contendrá el sentimiento asociado a dicho tweet.

B. Toma de datos:

La toma de datos se realizó mediante la API de X (previamente Twitter), la cual tiene un límite de unos 100 mensajes por cuenta, por lo que el resultado de la toma de datos es una base de datos un poco corta de contenido. Sin embargo, para motivos de prueba del prototipo también usamos una base de datos conseguida en la web.

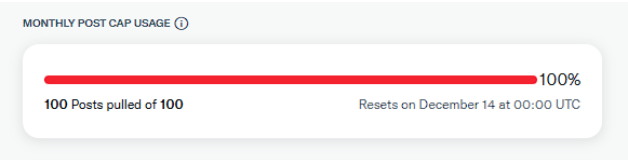

Figura 5. Límite de tweets extraíbles por cuenta

Para la extracción mediante el uso de la API debemos de utilizar la librería requests, que básicamente nos permite acceder al endpoint de la API para hacer las solicitudes. Una vez inicializado debemos de cargar las variables de entorno, principalmente el token que debemos suministrar de nuestra cuenta de X Developer, este Bearer Token es el que verifica nuestra identidad y finalmente nos permite hacer las solicitudes.

```
load_dotenv()
BEARER_TOKEN =
os.getenv("TWITTER_BEARER_TOKEN")
def create_headers(bearer_token):
    headers = {"Authorization": f"Bearer
{bearer_token}"}
    return headers
```

Figura 6. Código para inicializar el token a partir de un archivo, y así no subir las credenciales en GitHub

Luego básicamente agregamos creamos una función para obtener los tweets, donde accedemos al endpoint y dado el query, solicita los datos a la API para luego guardarlos en un CSV que será nuestra base de datos.

```
def get_tweets(query, bearer_token,
max_results):
    url =
"https://api.twitter.com/2/tweets/search/recen
t"
    headers = create_headers(bearer_token)
    params = {
        'query': query,
        'max_results': max_results,
        'tweet.fields': 'id,text,created_at'
    }
    response = requests.get(url,
headers=headers, params=params)
    if response.status_code ≠ 200:
        raise Exception(f"Error al obtener
tweets: {response.status_code},
{response.text}")
    return response.json().get('data', [])
def save_tweets_to_csv(tweets,
filename="tweetsCamilo.csv"):
    df = pd.DataFrame(tweets)
    df.to_csv(filename, index=False)
    print(f"Tweets guardados en {filename}")
```

Figura 7. Código para obtener los datos del endpoint de la API y su respectivo almacenamiento

Finalmente, hacemos el main que va a llamar a la función que buscará los tweets, este main es donde se ingresa el query que da los parámetros de la solicitud, específicamente que los tweets estén asociados a Barranquilla e inseguridad.

```python
if _name_ == "_main_":
    query = "(inseguridad OR robo OR atracos
OR violencia OR delincuencia) Barranquilla -
is:retweet"
    try:
        tweets = get_tweets(query,
BEARER_TOKEN, max_results=70)
        if tweets:
            save_tweets_to_csv(tweets)
        else:
            print("No se encontraron tweets
para la consulta especificada.")
    except Exception as e:
        print(f"Error: {e}")
```

Figura 8. Código con el query

Esta base de datos de la web mencionada tiene un inconveniente, los datos no tienen un hilo conductor, por lo tanto, los resultados del análisis de sentimientos en esta resultan bastante más generales.

Por otra parte, la base de datos extraída mediante la API de Twitter, a pesar de ser limitada por la cantidad de datos, se pueden ver claras tendencias que explicaremos más adelante.

	text
0	@henriqu28521949 @AtlanticoEmi @expresbrasilia...
1	@JhonjairoRend13 @AlejandroChar Ninguno \nVivo...
2	#Barranquilla #Violencia #Criminalidad \n\nLas...
3	#Atlántico #Violencia #Criminalidad\n\nUn jove...
4	#Cumpleaños #Tragedia 🚨 Tragedia en #Barranqui...

Figura 9. Primeros 5 datos extraídos mediante X Developer API

C. Preparación de datos:

1) Asignación de sentimientos:

Con los datos obtenidos se realiza la asignación del sentimiento de estos. Mediante el uso de la biblioteca Transformer se le asigna un puntaje a cada dato, este puntaje lego definirá qué sentimiento tiene la oración.

El puntaje va de 1 a 5 estrellas, donde 1 o 2 estrellas implica un sentimiento negativo, 3 estrellas significa que el texto tiene un sentimiento neutro y de 4 o 5 estrellas es asociado a un sentimiento positivo.

Este sentimiento conseguido al analizar los datos se le atribuye al CSV en forma de una nueva columna, la cual contendrá ‘neg’, ‘neu’ y ‘pos’ que identifican negativo, neutro y positivo respectivamente.

2) Limpieza de datos:

Una vez con los datos obtenidos y su sentimiento asignado, se procederá a hacer una preparación de estos. La preparación incluye la eliminación de formatos no necesarios para el análisis del sentimiento.

Principalmente el proceso de eliminación incluye:

1) Etiquetas HTML y URLs.
2) Menciones de usuarios (subcadenas que comiencen con @).
3) Emojis y caracteres especiales.
4) Puntuación en las palabras.
5) Mayúsculas, se normaliza el texto a minúsculas.

```python
def preprocess_text(text):
    text = re.sub(r'<.*?>|http\S+', '', text)
    text = re.sub(r'@\w+', '', text)
    text = re.sub(r'[^\w\s,]', '', text, flags=re.UNICODE)
    text = text.translate(str.maketrans('', '', string.punctuation))
    text = text.lower()
    tokens = word_tokenize(text)
    stop_words = set(stopwords.words('spanish'))
    tokens = [word for word in tokens if word not in stop_words]
    stemmer = SnowballStemmer('spanish')
    tokens = [stemmer.stem(word) for word in tokens]
    cleaned_text = ' '.join(tokens)
    return cleaned_text
```

Figura 10. Código para procesar y limpiar los datos

Luego de limpiar las cadenas te texto se procede a tokenizar el texto y posteriormente eliminar las stopwords (palabras que no añaden ningún tipo de sentimiento a la oración), para reducirlo a las palabras que deben ser analizadas.

Una vez identificadas las palabras clave, se aplica un proceso de Stemming. Este procedimiento reduce las palabras a su raíz, eliminando variaciones derivadas de conjugaciones o formas gramaticales. De esta manera, todas las formas de una palabra se consideran equivalentes, facilitando su conteo y análisis.

El último paso del proceso de preparación de datos es recomponer el texto tras realizar los procedimientos anteriores y guardarlo en un nuevo CSV que contenga tanto los textos limpios como su respectivo sentimiento.

D. Proporción de datos:

La proporción de datos asignada en el prototipo varía en base a qué base de datos estemos usando, para la base de

datos extraída por medio de la API de Twitter utilizamos una proporción de 70% para entrenamiento 30% para testeo, para la base de datos de la web utilizamos un 80% para entrenamiento 20% para testeo.

En el caso específico de la base de datos proveniente de la API de Twitter, se opta por una división más equilibrada entre entrenamiento y testeo, debido a la limitación en la cantidad de datos disponibles. Esto nos permite garantizar un conjunto de testeo más representativo y robusto.

```
X_train, X_test, y_train, y_test =
    train_test_split(
    tfidf_vectors,
    df['sentiment'],
    test_size = 0.3,
    random_state = 42)
```

Figura 11. Código de la proporción de datos para la base de datos sacada de X API

Por otra parte, decidimos que, al usar la base de datos proveniente de la web y esta tener un número alto de datos la proporción de 80% entrenamiento y 20% testeo.

```
X_train, X_test, y_train, y_test =
    train_test_split(
    tfidf_vectors,
    df['sentiment'],
    test_size = 0.2,
    random_state = 42)
```

Figura 12. Código de la proporción de datos para la base de datos sacada de la web

En ambos casos, el test_size indica la proporción de uso de la base de datos para la prueba, podemos ver que no se indica la cantidad que se va a usar para entrenamiento, sin embargo, esta es inferida por la librería en base a la de test_size.

E. Preparación de modelos:

Los modelos que vamos a usar van a ser del tipo SML (Supervised Machine Learning), siendo que le proveemos tanto los datos como su sentimiento.

Para iniciar el entrenamiento de los modelos que utilizaremos mediante la librería Scikit-learn haremos que se ejecuten cada uno mediante un for que incluya los modelos:

```
for model in models:
    model.fit(X_train, y_train)
    y_pred_woGsCv = model.predict(X_test)
    acc = accuracy_score(y_test,y_pred_woGsCv)
    accuracy_scores.append(acc)
    prec = precision_score(y_test, y_pred_woGsCv, average='weighted')
    rec = recall_score(y_test, y_pred_woGsCv, average='weighted')
    f1 = f1_score(y_test, y_pred_woGsCv, average='weighted')
```

Figura 13. Código de la preparación de los modelos

Donde básicamente ajustamos los datos y realizamos el entrenamiento de los modelos.

F. Evaluación y resultados de modelos:

Para medir la efectividad de estos modelos vamos a calcular varias métricas:

a. Accuracy: Representa el porcentaje total de valores correctamente clasificados, en este caso, si los sentimientos fueron correctamente clasificados como “pos”, “neu” y “neg”.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

b. Precission: La métrica de precisión es utilizada para poder saber qué porcentaje de valores que se han clasificado como positivos son realmente positivos.

$$precision = \frac{TP}{(TP + FP)}$$

c. Recall: Utilizada para saber cuántos valores positivos son correctamente clasificados.

$$Recall = \frac{TP}{(TP + FN)}$$

d. F1-score: Esta métrica combina el precission y el recall, para obtener un valor mucho más objetivo

$$F1 = 2 \cdot \frac{(recall \cdot precision)}{(recall + precision)}$$

TP indica los verdaderos positivos; TN, verdaderos negativos; FP, falsos positivos y FN, falsos negativos.

Una vez se entrenen los modelos y se hagan as pruebas, se miden las métricas acordes de cada modelo y finalmente se imprimen los resultados.

```
LogisticRegression()
Accuracy Score: 0.7631578947368421
Precision Score: 0.7292738082211766
Recall Score: 0.7631578947368421
F1 Score: 0.743498452012384

SVC()
Accuracy Score: 0.6710526315789473
Precision Score: 0.7066769865841073
Recall Score: 0.6710526315789473
F1 Score: 0.6481600342319213

RandomForestClassifier()
Accuracy Score: 0.7105263157894737
Precision Score: 0.6906236178681999
Recall Score: 0.7105263157894737
F1 Score: 0.6901514059120404
```

Figura 14. Resultados de las métricas de las pruebas

Para poder evidenciar mejor estos resultados, hacemos unas matrices de confusión que nos permitan evidenciar, dado el número de datos utilizados para las pruebas, cuántos fueron en realidad aciertos.

Para Random Forest:

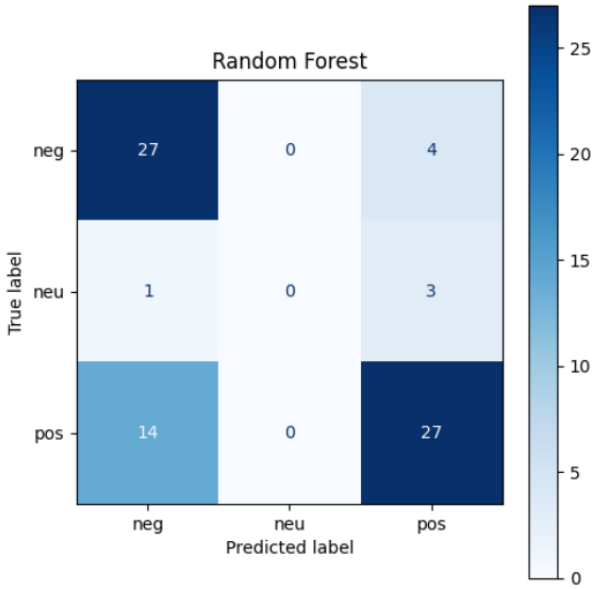

Figura 15. Matriz de confusión para Random Forest

Para SVC

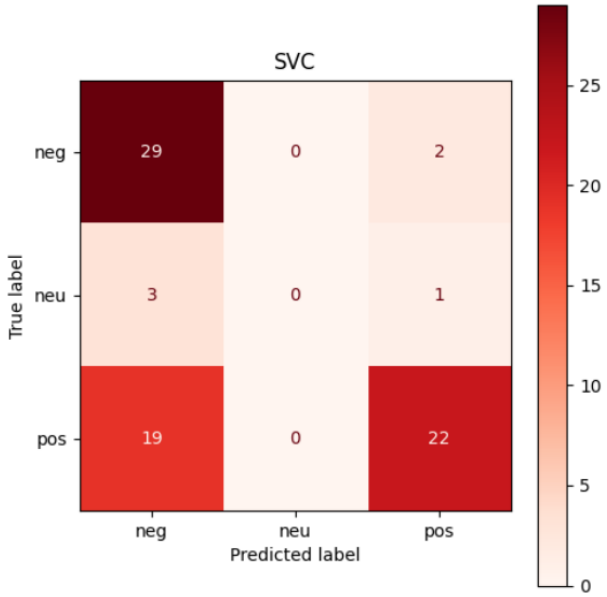

Figura 16. Matriz de confusión para SVC

Para regresión logística:

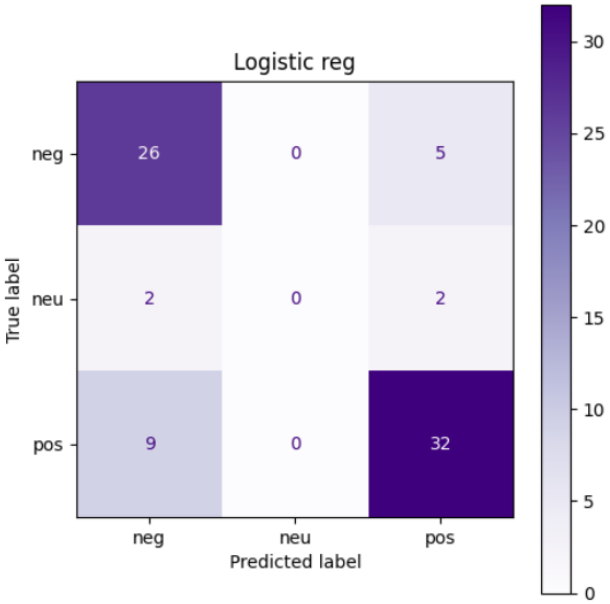

Figura 17. Matriz de confusión para Regresión logística

Por tanto, se puede evidenciar tanto en las métricas calculadas como en las matrices de confusión la superioridad de la regresión logística en el caso de predicción de sentimiento de una oración tras el entrenamiento.

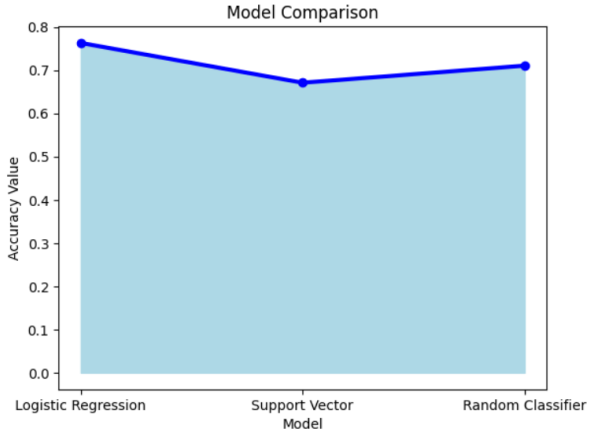

Figura 18. Tabla de comparación para el valor de Accuracy de modelos

G. **Grid search:**

Una vez con los resultados obtenidos de los tres modelos analizados, decidimos realizar los modelos nuevamente, pero con el uso de grid search, donde básicamente lo que se hace es optimizar los parámetros usados en los modelos y maximizar el rendimiento de cada uno. Esta forma de realizarlo es bastante pesada en términos de consumo, pero puede cambiar los resultados de los modelos.

```
pipeline = Pipeline([
    ('classifier', SVC())  # Placeholder; será reemplazado en el GridSearchCV
])

# Define el grid de parámetros para cada modelo
param_grid = [
    # Hiperparámetros para SVC
    {
        'classifier': [SVC()],
        'classifier__C': [10, 100],
        'classifier__gamma': [1, 0.1, 0.01, 0.001],
        'classifier__kernel': ['rbf', 'linear']
    },
    # Hiperparámetros para RandomForestClassifier
    {
        'classifier': [RandomForestClassifier()],
        'classifier__n_estimators': [50, 100, 200],
        'classifier__max_depth': [None, 10, 20],
        'classifier__min_samples_split': [2, 5, 10]
    },
    # Hiperparámetros para LogisticRegression
    {
        'classifier': [LogisticRegression(max_iter=1000)],
        'classifier__C': [0.1, 1, 10, 100],
        'classifier__penalty': ['l2'],
        'classifier__solver': ['lbfgs', 'liblinear']
    }
]
```

Figura 19. Código de la definición de los modelos y sus parámetros

Asignamos los hiperparámetros para cada modelo que vamos a probar. Una vez asignados, se realiza el grid search para con cada uno.

```
grid_search = GridSearchCV(pipeline, param_grid,
cv=5, scoring='f1_weighted', verbose=2)
```

Con el grid search realizado se hacen los cálculos de las métricas, así evaluamos cuál resultó con el mayor valor de precisión y poder tomar una decisión.

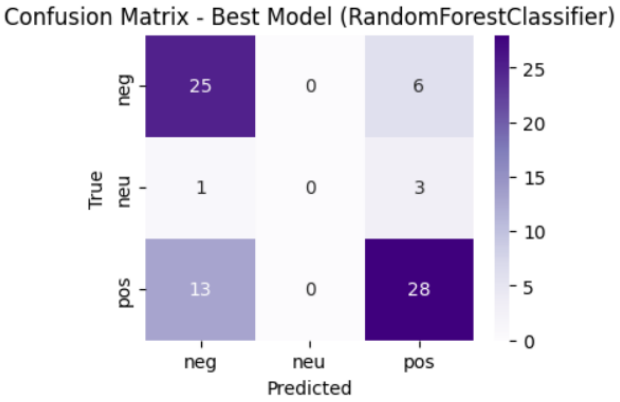

Figura 20. Matriz de confusión del mejor modelo con grid search: Random Forest

accuracy			0.70	76
macro avg	0.47	0.50	0.48	76
weighted avg	0.67	0.70	0.68	76

Figura 21. Métricas del mejor modelo con grid search

Como podemos observar, los resultados de las métricas resultan con valores más bajos que regresión logística sin el uso de grid search, el cual obtuvo los mejores resultados en las pruebas anteriores. ¿Por qué sucede esto? básicamente es porque al tener una base de datos tan limitada, esta limita el entrenamiento de grid search, el cual cuanto mayor sea el tamaño de la base de datos, mejores serán los resultados. Por lo tanto, el uso de esta herramienta podría ser importante al conseguir una database más robusta.

H. Validación del prototipo:

Caracteristica	Definición o descirpción	1	2	3	4	5
Understandability	¿Facil de comprender?					5
Documentation	¿Documentación de usuario completa, apropiada y bien estructurada?				4	
Buildability	¿Fácil de construir en un sistema compatible? (Close-Open)					5
Installability	¿Fácil de instalar en un sistema compatible?					
Learnability	¿Fácil de aprender a usar sus funciones?				4	
Identity	¿La identidad del proyecto / software es clara y única?				4	
Copyright	¿Es fácil ver quién posee el proyecto / software?					
Licencing	Adopción de la licencia apropiada?					
Governance	¿Fácil de entender cómo se ejecuta el proyecto y cómo se gestiona el desarrollo del software?					
Community	¿Evidencia de comunidad actual / futura?			3		
Accessibility	¿Evidencia de capacidad de descarga actual / futura?					
Testability	¿Fácil de probar la corrección del funciones caja negra?					
Portability	¿Utilizable en múltiples plataformas?					
Supportability	¿Evidencia de soporte para desarrolladores actuales / futuros?			3		
Analysability	¿Fácil de entender a nivel fuente?				4	
Changeability	¿Fácil de modificar y aportar cambios a los desarrolladores?					5
Evolvability	¿Evidencia de desarrollo actual / futuro?				4	
Interoperability	¿Interoperable con otro software requerido / relaciona					
modelo de evaluación basado en el estandar ISO 9126 ISO 15504 + ESCALA DE LIKERT						

Figura 22. Tabla de validación del prototipo

X. CONCLUSIONES

Finalizando el proyecto, el cual aborda el uso de herramientas para la extracción de publicaciones de redes sociales, otras especializadas para el tratamiento y preparación de esta información, para luego utilizar esta base de datos y aplicarle modelos de machine learning de tipo supervisado es importante resaltar varios puntos sobre los resultados obtenidos.

El prototipo desarrollado de un programa que tiene el propósito de analizar los sentimientos encontrados dada una base de datos mediante el uso de distintas técnicas de aprendizaje supervisado, dado que la base de datos incluye los datos tanto de entrenamiento como de prueba, y finalmente decidir en el análisis mediante métricas qué técnica funciona mejor acorde a lo establecido previamente. Esto nos ayudó a evidenciar que estas técnicas de aprendizaje permiten un análisis certero de los sentimientos encontrados en las publicaciones realizadas por los habitantes de la ciudad de Barranquilla mediante X (previamente Twitter). Estas técnicas, con sus limitaciones, son una gran herramienta que puede ser usada tanto por entidades gubernamentales como privadas para impulsar el desarrollo de la región. Durante el proyecto pudimos implementar distintas soluciones, unas más eficaces que otras, siendo la de Regresión Logística aquella que mostró mejores puntajes a nivel de métricas. Sin embargo, aquellas como SVC y Random Forest no son del todo fallidas. Por otra parte, la inclusión de Grid Search a pesar de que no fue del todo un éxito puede usarse a futuro si se cumplen las condiciones requeridas.

a. **Modelos implementados:** Los modelos utilizados mencionados con anterioridad son modelos que demuestran ser de gran utilidad al realizar este tipo de análisis de sentimientos. Principalmente, Regresión Logística cuyo accuracy superó aquellos de los demás modelos implementados con un valor del 0.7631 demostró que aún con una base de datos ajustada puede predecir eficazmente el sentimiento implícito de una oración dada. Los otros modelos, aunque no con valores tan altos como los de la Regresión Logística, también son una buena alternativa para la implementación de este tipo de herramientas. El Grid Search obtuvo unos resultados menores a los esperados, resultados que pueden cambiar dadas mejores condiciones al algoritmo.

b. **Limitantes y desafíos:** A pesar de que pudimos encontrar buenos resultados con los modelos planteados, notamos que la base de datos con la que trabajamos principalmente resultó demasiado pequeña. Esto se dio por la dificultar para acceder a la API específica de X (previamente Twitter) la cual requiere una mensualidad para poder extraer más de 100 tweets por cuenta. A pesar de que intentamos utilizar distintas plataformas de redes sociales, finalmente la única que cumplía con los requerimientos para saciar los datos necesarios, dado que X por su naturaleza es una red que busca que las personas se expresen libremente, no pudimos encontrar otra fuente de información. Con esto, ya que tenemos una cantidad muy limitada de datos para trabajar, los modelos no funcionan a su máximo potencial, ya que la baja cantidad afecta tanto su entrenamiento como sus pruebas. Todo esto resulta en un rendimiento menor a lo esperado en las métricas generales de todos los modelos.

c. **Futuras mejoras:** Finalizando, es importante mencionar que este prototipo realizado tiene espacio para mejoras. Contemplando que los datos no fueron suficientemente grandes podemos implementar el uso de redes neuronales para realizar un análisis más profundo, el cuál analice no únicamente el texto plano si no aquellos matices que caracterizan al ser humano al comunicarse, tales como el sarcasmo e ironía, así como este tipo de características del lenguaje natural. Por otra parte, se podría realizar el desarrollo con una mejor cantidad de datos, utilizando el método pago de X Developer API, el cual aumenta hasta 15.000 la cantidad de posts que se pueden extraer mensualmente. Con esta mayor cantidad de datos se pueden hacer análisis más profundos y con mejores resultados que resulten en conclusiones más exactas. También podríamos explorar el uso de otras redes sociales para complementar la principal fuente de información.

XI. BIBLIOGRAFÍA

[1] Ahmed, K. B. (s. f.). Sentiment Analysis for Smart Cities: State of the Art and Opportunities - ProQuest. https://www.proquest.com/openview/8897dc1c4858b3271a2e009a280a6288/1?pq-origsite=gscholar&cbl=1976348

[2] Evaluating ESG Impacts in African Cities through Topic-Level Sentiment Analysis. (2023, 26 octubre). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/10322894

[3] López-Chau, A., Valle-Cruz, D., & Sandoval-Almazán, R. (2020). Sentiment Analysis of Twitter Data Through Machine Learning Techniques. En Computer communications and networks (pp. 185-209). https://doi.org/10.1007/978-3-030-33624-0_8

[4] Suhaimin, M. S. M., Hijazi, M. H. A., Moung, E. G., Nohuddin, P. N. E., Chua, S., & Coenen, F. (2023). Social media sentiment analysis and opinion mining in public security: Taxonomy, trend analysis, issues and future directions. Journal Of King Saud University - Computer And Information Sciences, 35(9), 101776.https://doi.org/10.1016/j.jksuci.2023.101776

[5] Vohra, A., & Garg, R. (2022). Deep learning based sentiment analysis of public perception of working from home through tweets. Journal Of Intelligent Information Systems, 60(1), 255-274. https://doi.org/10.1007/s10844-022-00736-2

[6] Wang, Y., Guo, J., Yuan, C., & Li, B. (2022). Sentiment Analysis of Twitter Data. Applied Sciences, 12(22), 11775. https://doi.org/10.3390/app122211775

[7] Yang, L., Duarte, C. M., & Ciriquián, P. M. (2022). Quantifying the relationship between public sentiment and urban environment in Barcelona. Cities, 130, 103977. https://doi.org/10.1016/j.cities.2022.103977

[8] Yue, A., Mao, C., Chen, L., Liu, Z., Zhang, C., & Li, Z. (2022). Detecting Changes in Perceptions towards Smart City on Chinese Social Media: A Text Mining and Sentiment Analysis. Buildings, 12(8), 1182. https://doi.org/10.3390/buildings12081182

[9] Zengin, M. S., Arslan, R., & Akgün, M. B. (2022). **Distributed sentiment analysis for geo-tagged Twitter data**. *2022 30th Signal Processing and Communications Applications Conference (SIU)*, 15-18 May 2022, Safranbolu, Turkey. IEEE. https://ieeexplore.ieee.org/document/9864702

[10] J. I. R. Molano, D. G. Muñoz and J. D. C. Bernal, "Machine learning applications in tourism - case study from Colombia and its post-conflict areas," 2023 18th Iberian Conference on Information Systems and Technologies (CISTI), Aveiro, Portugal, 2023, pp. 1-6, doi: 10.23919/CISTI58278.2023.10211549. https://ieeexplore.ieee.org/document/10211549

[11] E. Puertas, J. C. Martinez-Santos and P. Andrés Pertuz-Duran, "Presidential preferences in Colombia through Sentiment Analysis," 2022 IEEE ANDESCON, Barranquilla, Colombia, 2022, pp. 1-5, doi: 10.1109/ANDESCON56260.2022.9989700. https://ieeexplore.ieee.org/document/9989700

[12] J. F. Sossa Rojo, L. A. Fletscher Bocanegra, J. F. Botero Vega and N. G. Gomez, "Crime Prediction Using Support Vector Machine and Extracted Twitter Features," 2023 IEEE Colombian Conference on Communications and Computing (COLCOM), Bogota, Colombia, 2023, pp. 1-5, doi: 10.1109/COLCOM59909.2023.10334281. https://ieeexplore.ieee.org/document/10334281

[13] O. Bustos and A. Pomares, "Analysis of social networks publications for stock market movement prediction: methodology and case study in a Spanish-speaking country," 2022 Congreso Internacional de Innovación y Tendencias en Ingeniería (CONIITI), Bogota, Colombia, 2022, pp. 1-5, doi: 10.1109/CONIITI57704.2022.9953669. https://ieeexplore.ieee.org/document/9953669

[14] C. Johnathan Izquierdo Ortiz, "Analysis of Economic Indicators Through News and Twitter Using Text Mining, Machine Learning and Multiagent Systems," 2023 IEEE Colombian Conference on Applications of Computational

Intelligence (ColCACI), Bogotá D.C., Colombia, 2023, pp. 1-6, doi: 10.1109/ColCACI59285.2023.10225755. https://ieeexplore.ieee.org/document/10225755
[15] J. C. Martinez-Santos, J. Vásquez and E. Puertas, "Researcher Profile: An Automated Solution for Searching and Gathering People's Profiles," 2023 IEEE Colombian Caribbean Conference (C3), Barranquilla, Colombia, 2023, pp. 1-5, doi: 10.1109/C358072.2023.10436220. https://ieeexplore.ieee.org/document/10436220
[16] C. Johnathan Izquierdo Ortiz, "Analysis of Economic Indicators Through News and Twitter Using Text Mining, Machine Learning and Multiagent Systems," 2023 IEEE Colombian Conference on Applications of Computational Intelligence (ColCACI), Bogotá D.C., Colombia, 2023, pp. 1-6, doi: 10.1109/ColCACI59285.2023.10225755. https://ieeexplore.ieee.org/document/10225755
[17] T. L. Ben, N. Ravikumar R, P. C. R. Alla, G. Komala and K. Mishra, "Detecting Sentiment Polarities with Comparative Analysis of Machine Learning and Deep Learning Algorithms," 2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT), Gharuan, India, 2023, pp. 186-190, doi: 10.1109/InCACCT57535.2023.10141741. https://ieeexplore.ieee.org/document/10141741
[18] Said-Hung, E., Arce-García, S., & Mottareale-Calvanese, D. "Sentimental polarization on Twitter during the 2021 National Strike in Colombia." https://teologiayvida.uc.cl/index.php/cdi/article/download/50483/50837/183565
[19] Plata, D., Torres, D., Solano, E. " Data analysis y clustering para el análisis de crímenes de alto impacto en Barranquilla". https://manglar.uninorte.edu.co/handle/10584/11883
[20] Atencia, W., Rambal, J., Bustillo, J. "Analizador de tweets asociados a la política y polarización Colombiana". https://manglar.uninorte.edu.co/handle/10584/9280
[21] Severyn, A., & Moschitti, A. (2015). Twitter Sentiment Analysis with Deep Convolutional Neural Networks. *ACM*. https://doi.org/10.1145/2766462.2767830
[22] A. M. Naira and I. Benelallam, "Evaluating ESG Impacts in African Cities through Topic-Level Sentiment Analysis," 2023 10th International Conference on Wireless Networks and Mobile Communications (WINCOM), Istanbul, Turkiye, 2023, pp. 1-6, doi: 10.1109/WINCOM59760.2023.10322894. https://ieeexplore.ieee.org/document/10322894
[23] V. Ramanathan and T. Meyyappan, "Twitter Text Mining for Sentiment Analysis on People's Feedback about Oman Tourism," *2019 4th MEC International Conference on Big Data and Smart City (ICBDSC)*, Muscat, Oman, 2019, pp. 1-5, doi: 10.1109/ICBDSC.2019.8645596. https://ieeexplore.ieee.org/document/8645596
[24] GeeksforGeeks. (s.f.). What is Sentiment Analysis? GeeksforGeeks. Recuperado de https://www.geeksforgeeks.org/what-is-sentiment-analysis/
[25] Shetty, A. M., Aljunid, M. F., Manjaiah, D. H., & Shaik Afzal, A. M. (2023, July). Hyperparameter Optimization of Machine Learning Models Using Grid Search for Amazon Review Sentiment Analysis. In International Conference on Data Science and Applications (pp. 451-474). Singapore: Springer Nature Singapore. https://link.springer.com/chapter/10.1007/978-981-99-7814-4_36