Aplicación de una herramienta externa a la API de Twitter (X) para recolectar metadatos que permitan modelar las acciones que realizan los usuarios al propagar intencionalmente información.

Díaz Villacis David Alexander

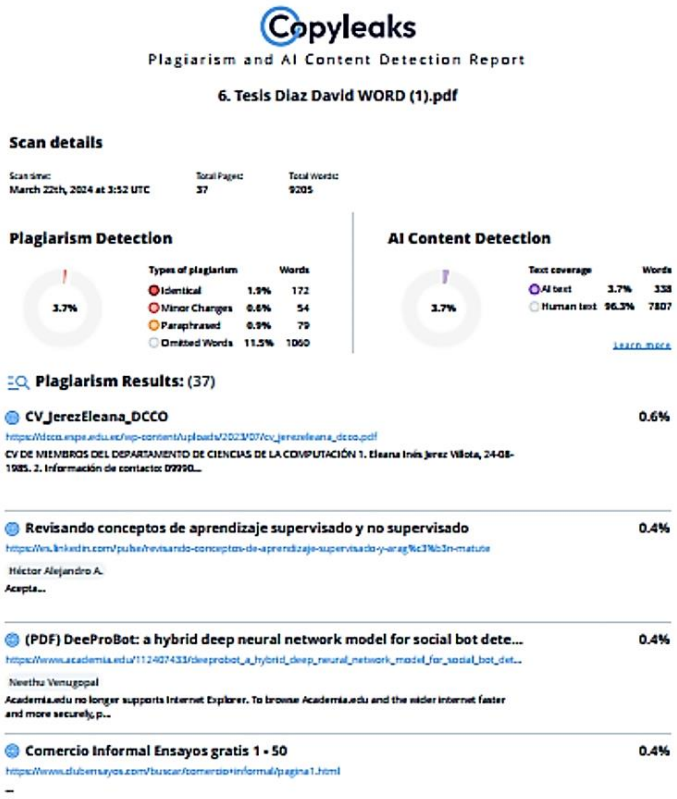
Departamento de Ciencias de la Computación

Carrera de Tecnologías de la Información

Trabajo de integración curricular, previo a la obtención del título de Ingeniero en Tecnologías de la Información

Ing. Jerez Villota Eleana Inés, MSc.

Sangolquí, 23 de febrero del 2024



Copyleaks

Plagiarism and AI Content Detection Report

6. Tesis Diaz David WORD (1).pdf

Scan details

Scan time	Total Pages	Total Words
March 22th, 2024 at 3:52 UTC	37	9205

Plagiarism Detection

Types of plagiarism		Words
Identical	1.9%	172
Minor Changes	0.6%	54
Paraphrased	0.9%	79
Omitted Words	11.5%	1060

3.7%

AI Content Detection

Text coverage		Words
AI text	3.7%	338
Human text	96.3%	7807

Learn more

Plagiarism Results: (37)

CV_JerezEleana_DCCO 0.6%

https://dcco.espe.edu.ec/wp-content/uploads/2023/07/cv_jerezeleana_dcco.pdf

CV DE MIEMBROS DEL DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN 1. Eleana Inés Jerez Villota, 24-08-1985. 2. Información de contacto 09990...

Revisando conceptos de aprendizaje supervisado y no supervisado 0.4%

https://es.linkedin.com/pulse/revisando-conceptos-de-aprendizaje-supervisado-y-arag%c3%b3n-matute

Héctor Alejandro A.

Acepta...

(PDF) DeeProBot: a hybrid deep neural network model for social bot dete... 0.4%

https://www.academia.edu/112407433/deeprobot_a_hybrid_deep_neural_network_model_for_social_bot_det...

Neethu Venugopal

Academia.edu no longer supports Internet Explorer. To browse Academia.edu and the wider internet faster and more securely, p...

Comercio Informal Ensayos gratis 1 - 50 0.4%

https://www.clubensayos.com/buscar/comercio+informal/pagina1.html

–

Firma:

Firmado electrónicamente por: ELEANA INES JEREZ VILLOTA

Ing. Eleana Inés Jerez Villota, MSc.

C. C: 1717225039

Departamento de Ciencias de la Computación

Carrera de Tecnologías de la Información

Certificación

Certifico que el trabajo de integración curricular: "**Aplicación de una herramienta externa a la API de Twitter (X) para recolectar metadatos que permitan modelar las acciones que realizan los usuarios al propagar intencionalmente información**" fue realizado por el señor Diaz Villacis David Alexander; el mismo que cumple con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, además fue revisado y analizado en su totalidad por la herramienta de prevención y/o verificación de similitud de contenidos; razón por la cual me permito acreditar y autorizar para que se lo sustente públicamente.

Sangolquí, 22 de marzo de 2024

ELEANA INES JEREZ VILLOTA

.....

Ing. Eleana Inés Jerez Villota, MSc.

C. C: 1717225039

Departamento de Ciencias de la Computación

Carrera de Tecnologías de la Información

Responsabilidad de Autoría

Yo, **Diaz Villacis David Alexander**; con cédula de ciudadanía N. 1718327347 declaro que el contenido, ideas y criterios del trabajo de integración curricular "**Aplicación de una herramienta externa a la API de Twitter (X) para recolectar metadatos que permitan modelar las acciones que realizan los usuarios al propagar intencionalmente información.**", es de mi autoría y responsabilidad, cumpliendo con los requisitos legales, teóricos, científicos, técnicos, y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Sangolquí, 22 de marzo de 2024

Diaz Villacis David Alexander

C.C. 1718327347

Departamento de Ciencias de la Computación

Carrera de Tecnologías de la Información

Autorización de Publicación

Yo, **Diaz Villacis David Alexander**; con cédula de ciudadanía N. 1718327347 autorizó a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de integración curricular "**Aplicación de una herramienta externa a la API de Twitter (X) para recolectar metadatos que permitan modelar las acciones que realizan los usuarios al propagar intencionalmente información.**", en el Repositorio Institucional, cuyo contenido, ideas y criterios son de mi responsabilidad.

Sangolquí, 22 de marzo de 2024

Diaz Villacis David Alexander

C.C. 1718327347

Dedicatoria

En el culminar de esta etapa, quiero dedicar este logro a aquellos que han sido pilares fundamentales en mi vida:

A Dios, fuente inagotable de fortaleza y guía, a quien agradezco por ser guía en cada paso de mi camino, a mis padres Nora y César, cuyo amor incondicional y sacrificio han sido la base sobre la cual he construido mis sueños. Su apoyo inquebrantable me ha llevado a alcanzar esta meta, a mi abuelito Carlos Villacis que en paz descanse, quien con su sabiduría y cariño dejo una huella imborrable en mi vida, a mi abuelita Pilar que toda la vida me ha inspirado a perseguir mis objetivos con determinación. A mis tías Isabel y Sandra, cuyo cariño y consejos han enriquecido mi vida.

A mis amigos Marcelo Arias, Jerry Tovar y Jean Pierre Toapanta, compañeros a quienes les dedico gratitud por ser parte de este trayecto y por su amistad leal que ha sido un sostén invaluable.

Este logro no solo es nuestro, sino de todos aquellos que han sido parte de este camino. ¡Gracias por ser parte de este capítulo tan importante en mi vida!

David Diaz

Agradecimiento

A mi familia en general, que me enseñaron a perseverar para alcanzar todas mis metas propuestas, les agradezco por impulsarme a conseguir aquellas metas propuestas sin importar las dificultades en el camino.

A los docentes de la Universidad de las Fuerzas Armadas ESPE por compartir sus conocimientos y por enfocarnos en que cada día debemos ser mejores profesionales y seres humanos, y de manera especial a la Ingeniera Eleana Jerez quién me oriento con toda la disponibilidad de tiempo para poder alcanzar el máximo de mi capacidad y poder concluir con éxito mi formación académica.

Quedo totalmente agradecido con todas las personas que participaron en este proceso, que ha sido parte de mi vida y que me llevará a ser una gran persona tanto en el ámbito laboral como personal.

David Diaz

Índice de Contenidos

Índice de tablas

Índice de figuras

Resumen

En la red social digital X (anteriormente conocida como Twitter), la difusión de información no verificada es común, especialmente en temas relacionados con asuntos sociales, políticos y económicos, lo que puede influir significativamente en el comportamiento de las personas y, en ocasiones, llevar al caos y al pánico colectivo. Aunque se han realizado muchos estudios sobre el análisis de información en esta plataforma, la mayoría de ellos dependen de la API de Twitter para recopilar datos. Sin embargo, con los recientes cambios en la plataforma, incluido el cambio de nombre a X, y las limitaciones de la API, es crucial encontrar otras herramientas que brinden un mayor alcance y facilidad para recopilar datos. Identificar estas alternativas no solo ayuda a comprender mejor cómo los usuarios difunden información intencionalmente, sino que también ayuda a entender y abordar los problemas asociados con la desinformación y sus efectos en la sociedad.

Además, al identificar estas herramientas alternativas, será posible modelar las acciones que realizan los usuarios al propagar información intencionalmente, lo que permitirá clasificarlos en grupos de comportamiento similar en la red social. Este enfoque facilitará el diseño de estrategias más efectivas para contrarrestar la desinformación y promover una comunicación más precisa y responsable en las plataformas digitales.

Palabras claves: Recopilación, Twitter (X), SNScrape, Tweepy, Nodos, Relacionamiento.

Abstract

In the digital social of X (formerly Twitter), the propagation of unverified information is rampant, particularly on topics spanning social, political, and economic domains, potentially exerting profound influence on individual behaviors and occasionally precipitating collective chaos and panic. While numerous studies have scrutinized information dynamics on this platform, the predominant reliance on Twitter's API for data aggregation presents limitations, especially amidst recent platform shifts, including the transition to X. Consequently, there arises a pressing need to explore alternative data collection tools that offer broader coverage and streamlined data gathering processes.

Identifying such alternatives not only enriches our understanding of deliberate information dissemination by users but also serves as a crucial step towards addressing the multifaceted challenges associated with misinformation and its societal ramifications. Moreover, the identification of these alternative tools holds the promise of enabling the modeling of user behaviors in intentional information propagation, facilitating their categorization into behaviorally akin clusters within the social network.

This nuanced approach not only fosters the development of more robust strategies to counter misinformation but also fosters a climate conducive to more accurate and responsible communication across digital platforms. In essence, it lays the groundwork for a more resilient and discerning online community, better equipped to navigate the complexities of information dissemination and consumption in the digital age.

Keywords: Data, Twitter (X), SNScrape, Tweepy, Nodes, Relationship Mapping.

Capítulo I

Introducción

En este capítulo se abordan los antecedentes, planteamiento del problema, la justificación, el alcance, el objetivo general, los objetivos específicos y la estructura del trabajo.

Antecedentes

En la era digital, las redes sociales desempeñan un papel fundamental en la comunicación, siendo Twitter (X) un claro ejemplo. Estas plataformas no solo conectan a personas de diversas partes del mundo, sino que también ejercen influencia al formar opiniones, difundir tendencias e información, ya sea verificada o no. El hecho esta en comprender cómo los usuarios propagan información ya sea verificada o no en Twitter (X) se vuelve esencial para comprender sus opiniones y cómo estas afectan las decisiones individuales y colectivas.

En Twitter (X) millones de usuarios comparten sus pensamientos en tiempo real, la urgente necesidad de contar con herramientas eficientes para la recolección y análisis de datos en esta plataforma es esencial (Vosoughi, 2017). La velocidad y amplitud con la que se comparten ideas hacen que el proceso de comprensión y clasificación de la información sea un desafío constante.

La generación de nodos de relacionamiento se convierte en una parte fundamental del trabajo ya que nos ayuda en la extracción de detalles como nombre, edad, sexo, ubicación e intereses desde las redes sociales, específicamente de Twitter (X). Estos nodos de relacionamiento permiten no solo verifcar el comportamiento en línea del usuario, incluyendo tweets y retweets, sino también anticipar su comportamiento futuro.

El análisis de nodos de relacionamiento va más allá de las interacciones evidentes en la plataforma, extendiéndose a la predicción de preferencias políticas y causas sociales que un usuario podría respaldar (Castillo, 2011). Este enfoque predictivo se vuelve crucial en un entorno donde la información y las opiniones influyen directamente en las decisiones y acciones de las personas.

Planteamiento del problema

La difusión de información no verificada en la red social digital Twitter(X) (antes conocida como Twitter) se ha generalizado, especialmente sobre temas relacionados con cuestiones sociales, políticas y económicas. Esta condición puede afectar significativamente el comportamiento de las personas, posiblemente provocando caos y pánico masivo.

Aunque se han realizado muchos estudios sobre la minería de datos de Twitter(X), la mayoría de ellos se basan en la API de Twitter(X) para recopilar datos. Sin embargo, debido a los cambios recientes en la plataforma, incluidos cambios de nombre y restricciones de la API, es importante encontrar otras herramientas que ofrezcan mayor alcance y facilidad de recopilación, más datos de uso y configuración. Por lo tanto, este trabajo identifica alternativas a la extracción de datos de Twitter(X) para modelar a los usuarios en el proceso de difusión de información e identificar características similares en su comportamiento en línea y que adoptan métodos de clasificación.

Justificación

Las plataformas de redes sociales como Twitter(X) desempeñan un papel importante en la difusión de noticias, opiniones y eventos relacionados en tiempo real. Sin embargo, la facilidad con la que se difunde la información en estas plataformas también genera problemas relacionados con la confiabilidad y eficacia de la información proporcionada. La difusión de información no verificada puede tener un impacto significativo

en cuestiones sociales, políticas y económicas, y posiblemente conducir al caos y al pánico masivo. Esta investigación se centrará en la necesidad de abordar los desafíos en la difusión de información en las redes sociales, especialmente en Twitter (X), y la importancia de explorar alternativas a la API de Twitter (X), para extraer datos que nos ayuden a modelar a los usuarios en la difusión de información, procesamiento y gestión, e identificar grupos de usuarios con comportamiento similar.

Objetivos

Objetivo General

Modelar las acciones que realizan los usuarios al propagar intencionalmente información, mediante una herramienta que no use la API de Twitter (X) para recolectar metadatos de los perfiles de usuario y de los tweets.

Objetivos Específicos

- Realizar la revisión de literatura de las herramientas para la recolección de metadatos de los perfiles de usuario y tweets.
- Definir la herramienta para recolectar metadatos de los perfiles de usuario y tweets.
- Recolectar metadatos de los perfiles de usuario y tweets en Twitter (X).
- Modelar las acciones que realizan los usuarios al propagar intencionalmente información.

Capitulo II

Revisión de la Literatura

Este capítulo se enfoca en realizar una revisión de la literatura, empleando una metodología orientada a la recopilación y análisis de estudios previos.

Preguntas de Investigación

En la etapa inicial de la revisión literaria se elaboraron preguntas de investigación (PI) que desempeñan el papel de orientación para la búsqueda de información en los artículos pertinentes. A continuación, se detallan las preguntas de investigación formuladas:

- PI1: ¿Cuál es el estado del arte acerca de la recolección de metadatos en redes sociales digitales para modelar usuarios?

Evaluar la literatura existente para comprender la evolución y las prácticas predominantes en la recolección de metadatos para la modelación de usuarios en el contexto de las redes sociales digitales.

- PI2: ¿Cuáles son los recursos (datasets) que se utilizan en los trabajos para la experimentación?

Identificar y examinar conjuntos de datos destinados a la modelación de usuarios en redes sociales digitales, para analizar su utilidad, alcance y aplicabilidad.

PI3: ¿Cuáles son las técnicas o herramientas para recolectar metadatos en las redes sociales digitales?

Investigar las diversas técnicas o herramientas utilizadas para la recolección de metadatos en el ámbito de las redes sociales digitales, con el objetivo de comprender las metodologías actuales y sus implicaciones prácticas.

- PI4: ¿Cuáles son las técnicas o métodos para la clasificación de nodos dentro de las redes sociales digitales?

Explorar las técnicas o métodos empleados para clasificar nodos en las redes sociales digitales.

PI5: ¿Cuáles son los tipos de contribución?

Analizar las contribuciones presentes en la literatura sobre la modelación de usuarios en redes sociales digitales, con el fin de identificar patrones, enfoques exitosos y áreas de innovación en el campo de estudio.

Cadena de búsqueda

Las palabras clave seleccionadas se emplearon para construir la cadena de búsqueda óptima, la cual se implementó en bases de datos científicas, ajustándola según las necesidades específicas de cada plataforma, con el propósito de localizar artículos pertinentes. La cadena de búsqueda resultante se presenta a continuación:

(gathering OR collection) AND ((user OR node) AND profile) AND (social AND (network OR media))

Criterios de inclusión y exclusión

Utilizando la cadena de búsqueda predefinida, se llevó a cabo una exploración exhaustiva en las bases de datos electrónicas IEEEXplorer y Scopus, aplicando criterios específicos de inclusión y exclusión.

Los criterios de inclusión adoptados fueron los siguientes:

- Se incluyeron los artículos escritos en inglés para garantizar la coherencia del análisis.
- Se incluyeron los artículos de revistas académicas, asegurando así la calidad y rigurosidad de los documentos seleccionados.
- Se focalizó la investigación en el campo de las Ciencias de la Computación para mantener la cohesión temática.
- Se incluyeron los artículos publicados entre 2019 y 2023, asegurando la relevancia de la información recopilada.

Exclusión de trabajos:

- Se excluyeron los trabajos que en el título y resumen no reportaban acerca de la recolección de datos en redes sociales digitales.
- Se excluyeron los trabajos que en el título y resumen no reportaban acerca de modelos de usuarios en redes sociales digitales.

La búsqueda arrojó los siguientes resultados detallados por cada base de datos:

En la base de datos IEEEXplorer, se hallaron un total de 12 artículos y en la base de datos Scopus, se identificaron 40 artículos conforme a los parámetros de búsqueda predefinidos. En conjunto, se obtuvieron 52 artículos. No obstante, se observó que 8 de estos documentos estaban duplicados, presentándose en ambas bases de datos.

Luego aplicamos los criterios de inclusión y exclusión, definidos anteriormente obtuvimos un total de 13 artículos primarios.

Resultados de la revisión de literatura

Tabla 1

Aplicación de los criterios de inclusión y exclusión, la Tabla 1 corresponde a la PI1, muestra los trabajos relacionados obtenidos del estado del arte de la investigación.

Id	Titulo	Fuente	Dataset	Herramientas	Redes sociales	Método	Contribución
1	Factor Graph Model Across Social Networks	IEEE	Sina Microblog, RenRen	N/D	Facebook, Twitter (X), MySpace	Aprendizaje No Supervisado	Método PIFGM, Algoritmo de aprendizaje distribuido basado en MPI.
2	Heterogeneous Social Media Analysis for Efficient Deep Learning	IEEE	Kaggle, Zenodo, Statista	Python(TensorFlow 2.4 y Keras)	Facebook, Twitter (X)	Procesamiento de Lenguaje Natural (PLN)	Modelo de aprendizaje de transferencia profunda, Análisis de datos multimodales.

	Fake-Profile Identification						
3	Ontology-Driven Digital Profiling for Identification and Linking Evidence Across Social Media Platform	IEEE	Matlab	Matlab	Facebook, Twitter (X), Tik Tok	Análisis de Sentimientos	Perfil de usuario basado en artefactos digitales.
4	Classification of social media users with generalized functional data análisis	SCOPUS	Kaggle	Clasificación de Bayes para observaciones funcionales	Twitter (X), Facebook, Reddit	Análisis de Sentimientos	Clasificación basada en el análisis de componentes principales funcionales a observaciones funcionales de valores binarios.
5	Rhythms in Twitter (X)	SCOPUS	Twitter (X) API	Técnicas de comportamiento de ráfagas predicho por Barabasi.	Twitter (X)	Aprendizaje No Supervisado	Análisis del ritmo de publicaciones en Twitter (X).
Id	**Titulo**	**Fuente**	**Dataset**	**Herramientas**	**Redes sociales**	**Método**	**Contribución**
6	Deep learning based topic and sentiment analysis: COVID19 information seeking on social media	SCOPUS	Kaggle, Api Twitter (X).	Python(TensorFlow 2.4 y Keras)	Twitter (X)	Clasificación Supervisada	Modelo de Red Neural Informada (INN) para identificar el sentimiento de cada tweet.
7	DeeProBot: a hybrid deep neural network model for social bot detection based on	SCOPUS	Kaggle	Python	Twitter (X)	Clasificación Supervisada	Modelo híbrido de red neuronal profunda para la detección de bots sociales basado en datos de perfil de usuario.

	user profile data						
8	Artificial Intelligence Model for the Identification of the Personality of Twitter (X) Users through the Analysis of Their Behavior in the Social Network	SCOPUS	Sina	Python	Twitter (X)	Clasificación Supervisada	Análisis de datos con algoritmos de aprendizaje automático.
9	Prediction of brand stories spreading on social networks	SCOPUS	N/D	Python	Twitter (X)	Aprendizaje No Supervisado	Modelos para predecir la difusión de historias de marca en las redes sociales, tanto en términos de capacidad de difusión como de nivel de difusión.
10	Social recruiting: an application of social network analysis.	SCOPUS	N/D	Web Scraping	Git Hub	Aprendizaje No Supervisado	Desarrollo de un enfoque metodológico que permita la preselección de candidatos mediante el análisis de redes sociales.
Id	**Titulo**	**Fuente**	**Dataset**	**Herramientas**	**Redes sociales**	**Método**	**Contribución**
11	Analyzing the Quality of Twitter (X) Data Streams	SCOPUS	Kaggle	Herramienta de software que permite capturar flujos de datos de Twitter (X).	Twitter (X)	Aprendizaje No Supervisado	Desarrollo de dimensiones y métricas de calidad, para capturar las características de los flujos de datos de Twitter (X).
12	Spam detection on profile and social media network using principal component	SCOPUS	Kaggle	Python	Twitter (X)	Aprendizaje No Supervisado	Análisis de Componentes Principales (PCA) y la agrupación de K-medias con distancia de Mahalanobis como método para detectar una colección de

	analysis| (PCA) and K-means clustering						usuarios que tienen propiedades similares para determinar el spam
13	Factor Graph Model Based User Profile Matching across Social Networks	SCOPUS	N/D	N/D	Twitter (X)	Clasificación Supervisada	PIFGM (Pairwise Identical Factor Graph Model), un Un novedoso modelo basado en un gráfico de factores, para abordar este problema considerando tanto los atributos del usuario como las relaciones de amigos a través de las redes.

Nota. Esta tabla muestra los trabajos relacionados que tiene el presente trabajo de titulación.

PI2:

Para responder a la PI2, elaboramos la Tabla 2 que muestra las redes sociales digitales utilizadas en el estado del arte.

Tabla 2

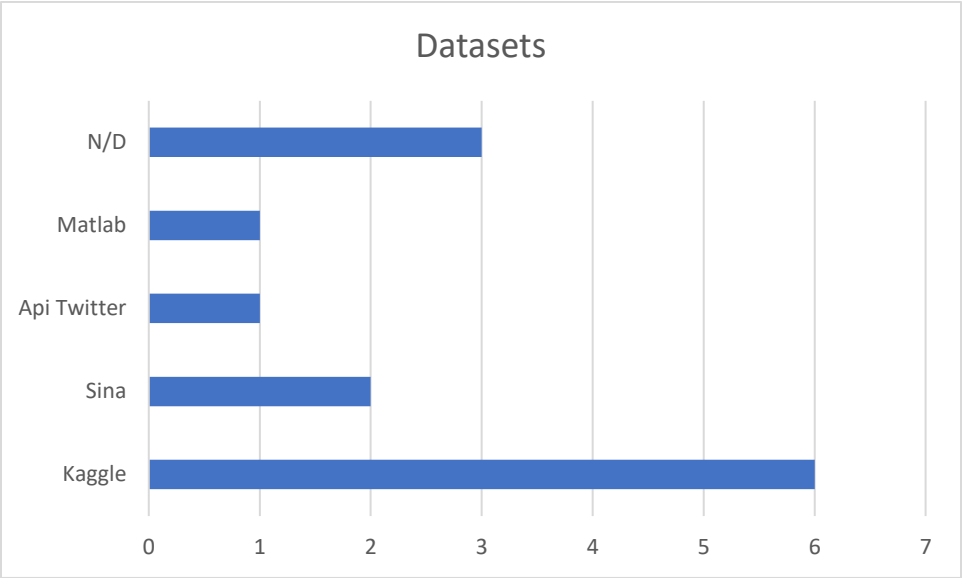
Datasets usados en los trabajos relacionados.

Datasets	Estudios
Kaggle	{2},{4},{6},{7},{11},{12}
Sina	{1},{8}
Twitter (X)	{5}
Datasets de la Herramienta Matlab	{3}
N/D	{9},{10},{13}

Nota. Esta tabla muestra los datasets encontrados en los trabajos relacionados.

Figura 1

Datasets usados en los trabajos relacionados.



Nota. La figura representa los datasets encontrados dentro de los trabajos relacionados.

Por lo tanto, estos datasets proporcionan información relevante como se puede observar en la Figura 1, ya que nos permitirá abordar la pregunta de investigación PI2 sobre los conjuntos de datos centrados en redes sociales digitales para la modelación de usuarios.

Ahora bien, enfocándonos en las redes sociales pudimos determinar las siguientes Tabla 3:

Tabla 3

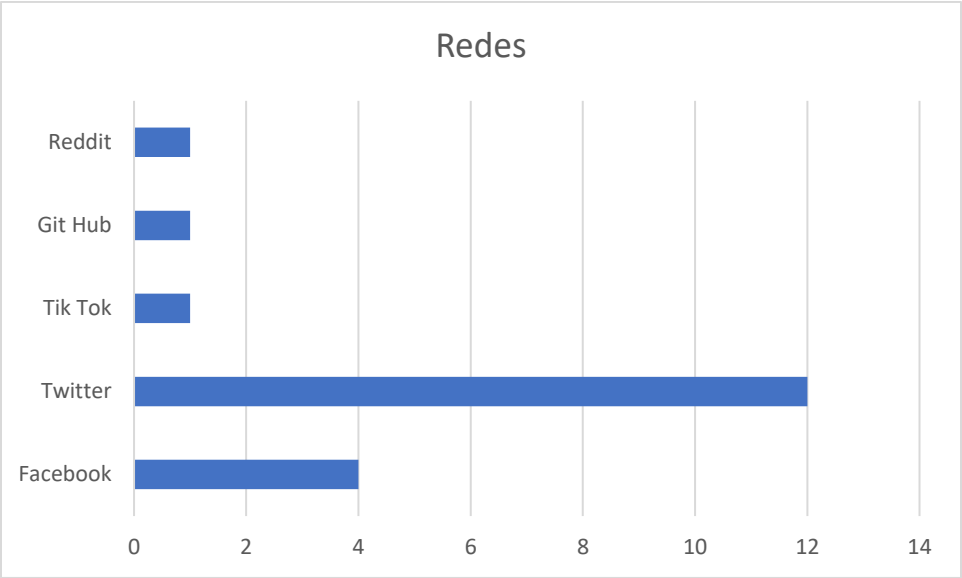
Redes Sociales usados en los trabajos de relacionados.

Redes Sociales	Estudios
Facebook	{1},{2},{3},{4}
Twitter (X)	{1},{2},{3},{4},{5},{6},{7},{8},{9},{11},{12},{13}
Tik Tok	{3}
Git Hub	{10}
Reddit	{4}

Nota. Esta tabla muestra las redes sociales encontrados en los trabajos relacionados.

Figura 2

Redes sociales usadas en los trabajos relacionados.



Nota. La figura representa las redes sociales que fueron encontradas dentro de los trabajos relacionados.

Mediante el análisis gráfico en la Figura 2, se puede inferir que Twitter (X) emerge como la plataforma más usada para la modelación de usuarios en el contexto de redes sociales digitales. Esta conclusión se sustenta en la observación de que Twitter (X) exhibe la mayor cantidad de estudios relacionados en comparación con otras plataformas. (Š. Grigaliūnas, 2023).

PI3:

Las técnicas o herramientas utilizadas para recolectar metadatos en redes sociales digitales, para dar respuesta a la pregunta de investigación PI3, realizamos la siguiente Tabla 4:

Tabla 4

Herramientas usadas en los trabajos de relacionados.

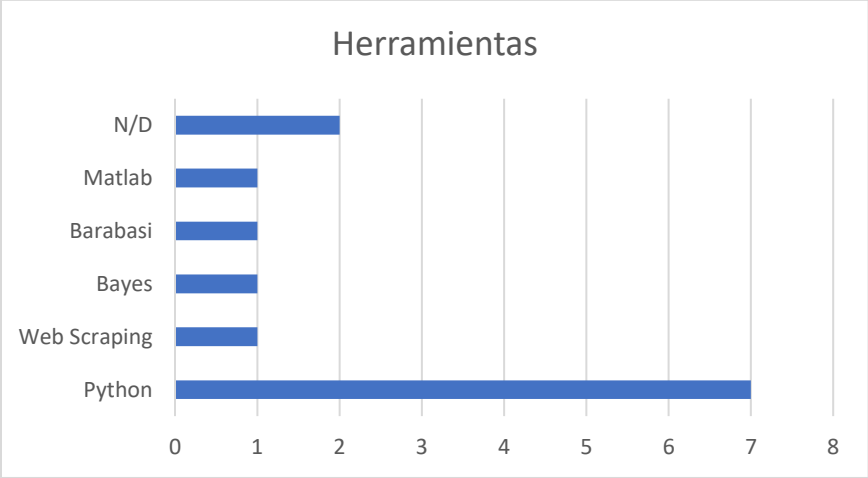
Herramientas	Estudios
Python	{2},{6},{7},{8},{9},{11},{12}
Web Scraping	{10}
Bayes	{4}

Modelo Barabasi	{5}
Matlab	{3}
N/D	{1},{13}

Nota. Esta tabla muestra las herramientas encontradas en los trabajos relacionados.

Figura 3

Herramientas usadas en los trabajos relacionados.



Nota. La figura representa las herramientas que fueron encontradas dentro de los trabajos relacionados.

En consecuencia, las herramientas y técnicas mencionadas, como Python, Web Scraping, Bayes, Método de Barabasi, y Matlab, son empleadas para la recolección de metadatos como se puede apreciar en la Figura 3, en el contexto de las investigaciones sobre redes sociales digitales. La falta de información detallada en algunos casos (N/D) subraya la diversidad de enfoques utilizados en la investigación en este campo.

PI4:

Las técnicas de clasificación utilizadas para clasificar los nodos de relacionamiento en redes sociales digitales para dar respuesta a la pregunta de investigación PI4, realizamos la siguiente Tabla 5:

Tabla 5

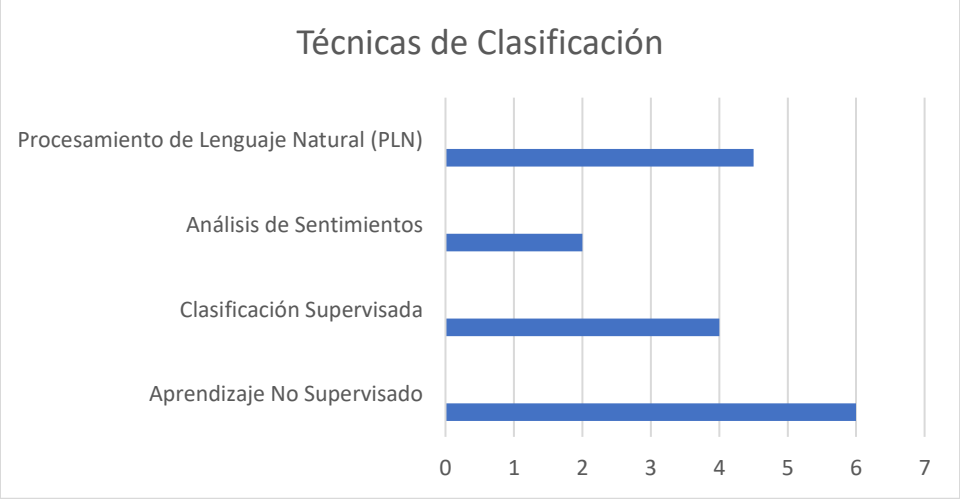
Técnicas de clasificación usadas en los trabajos de relacionados.

	Técnicas de clasificación	Estudios
	Aprendizaje No Supervisado	{1},{5},{9},{10},{11},{12}
	Clasificación Supervisada	{6},{7},{8},{13}
	Análisis de Sentimientos	{3},{4}
	Procesamiento de Lenguaje Natural (PLN)	{2}

Nota. Esta tabla muestra las técnicas de clasificación encontradas en los trabajos relacionados.

Figura 4

Técnicas de clasificación usadas en los trabajos relacionados.



Nota. Este grafico muestra las técnicas de clasificación encontradas en los trabajos relacionados.

Al analizar la Tabla 5, se observa que la técnica de clasificación que prevalece en los estudios revisados es el "Aprendizaje No Supervisado", evidenciado por la mayor cantidad de estudios asociados. Este hallazgo sugiere que, dentro del ámbito abordado, la comunidad científica ha mostrado una preferencia por la implementación de enfoques supervisados para la clasificación de textos. En contraste podemos observar en la Figura 4

que las técnicas de "Clasificación Supervisado", "Análisis de Sentimientos" y "Procesamiento de Lenguaje Natural (PLN)" están representadas con un número inferior de estudios, indicando una menor frecuencia de aplicación o investigación en comparación con la clasificación supervisada.

PI5:

Los tipos de contribución encontrados dentro de los trabajos relacionados, para dar respuesta a la pregunta de investigación PI5, hemos realizado la siguiente Tabla 6:

Tabla 6

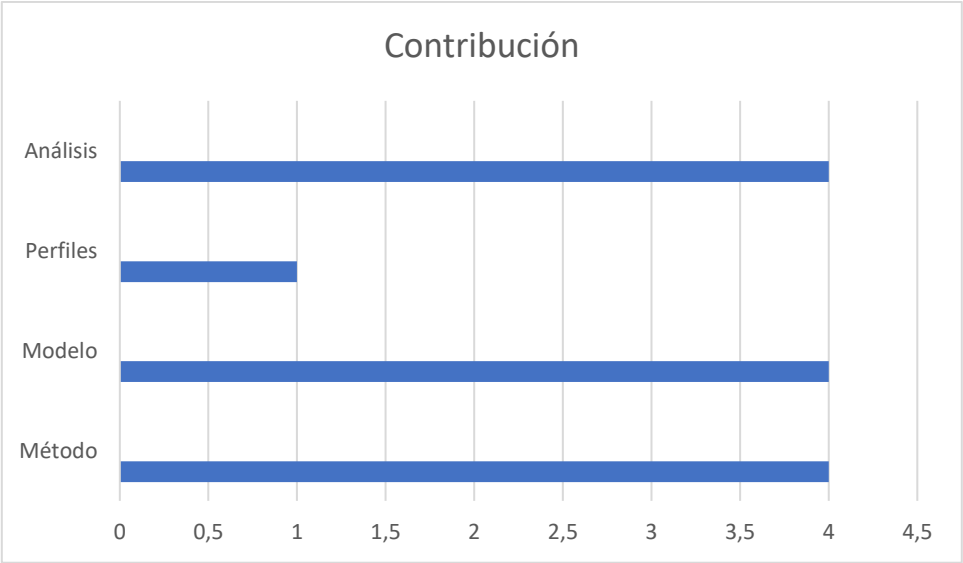
Tipos de contribución de los trabajos relacionados.

Contribución	Estudios
Método	{1},{10},{11},{13}
Modelo	{2},{6},{7},{9}
Validación	{3}
Experimento	{4},{5},{8},{12}

Nota. Esta tabla muestra las contribuciones encontradas en los trabajos relacionados.

Figura 5

Tipos de contribución de los trabajos relacionados.



Nota. Este gráfico nos muestra las contribuciones de los trabajos relacionados.

Para entender mejor las contribuciones en este tema, revisamos detalladamente algunos estudios y resumimos lo encontrado en una tabla, organizamos las ideas según diferentes tipos, siguiendo la clasificación propuesta en (Wieringa, 2006), el cual hace evidencia que las categorías "Método" y "Modelo" representan la mayor cantidad de contribuciones.

Capítulo III

Metodología de solución propuesta

A lo largo de este capítulo se describirá la metodología que se siguió para cumplir el objetivo general de este trabajo que consiste principalmente en las fases que comprenden el “Proceso de Análisis de datos”, estas fases incluyen:

La recopilación de datos se basará en el empleo especializado de técnicas de web scraping en Twitter (X). Se proporcionará una descripción exhaustiva del funcionamiento de

las herramientas seleccionadas, respaldando la elección de estas en el contexto único de este proyecto.

El preprocesamiento de datos se presentará como una fase crucial para garantizar la calidad y coherencia de la información recopilada. Se describirán las técnicas aplicadas para limpiar y estructurar los datos antes de avanzar hacia su procesamiento.

La fase de procesamiento se sumergirá en la implementación del modelo de usuario propuesto. Se destacarán los algoritmos y técnicas específicos utilizados para procesar la información recopilada.

La metodología finalmente abordará el análisis de datos, para descubrir patrones de comportamiento de los nodos de relacionamiento en el proceso de propagación de información con el fin de clasificarlos en grupos de características similares.

Recopilación de datos

Selección del conjunto de datos

De acuerdo con la revisión de literatura Figura 1, observamos que la mayoría de los trabajos utilizan datasets tomados de la plataforma Kaggle ya que es una plataforma en línea que se centra en la ciencia de datos y la minería de datos estos han obtenido buenos resultados (Kaggle, 2024). Por ello creemos que es una buena opción para nuestro trabajo.

Seleccionamos un dataset que en la época contemporánea se ha vuelto un tema en tendencia siendo este el uso de inteligencia artificial enfocado en el chat gpt, este tema nos parece ideal debido a que no es un tema polémico y no involucra temas políticos, este dataset está conformado por 50000 tweets que involucran información como: Datetime, Tweet Id, Text, Username, Permalink, User, Outlinks, CountLinks, ReplyCount, RetweetCount, LikeCount, QuoteCount, ConversationId, Language, Source, Media, QuotedTweet, MentionedUsers, hashtag|, hashtag_counts.

Limpieza del Dataset

Para el proceso de limpieza del conjunto de datos, se procedió a la extracción de tweets que presumiblemente fueron redactados por individuos. El propósito de esta acción es obtener nodos que representen a usuarios en el contexto de la propagación de información en las redes sociales. Las actividades realizadas incluyeron:

- Identificación de entidades: 200.
- Identificación de nodos de relacionamientos automatizados (robots): 122.
- Identificación de nodos de relacionamientos humanos: 8004.
- Inclusión exclusivamente tweets redactados en inglés.

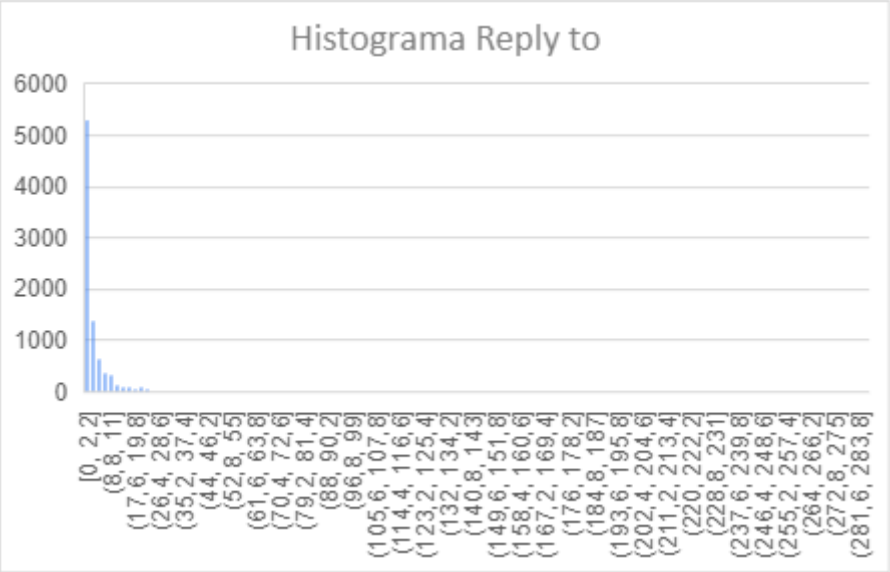
Como resultado de este proceso, se infiere que posiblemente 8004 tweets del conjunto de datos fueron redactados por usuarios humanos. Para validar esta inferencia, se reconoció la necesidad de actualizar el conjunto de datos. Con este fin, se contó con la colaboración de 12 participantes para llevar a cabo la actualización de los campos de respuesta durante 19 días.

Extracción de la muestra

Para llevar a cabo la extracción de la muestra, implementamos procedimientos destinados a obtener resultados más significativos. Adoptamos un criterio riguroso al seleccionar únicamente aquellos tweets que cuentan con más de 12 respuestas. Esta elección se basa en datos gráficos como histogramas Figura 6 que nos demuestran que los tweets con respuestas >= a 12 podría indicar una probabilidad más elevada de relaciones entre los participantes.

Figura 6

Histograma para la selección de la muestra



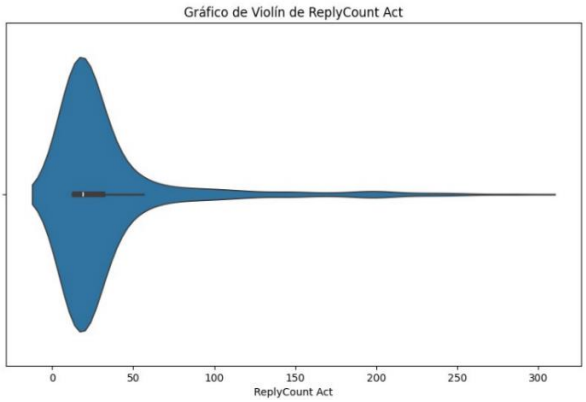
Nota. Este gráfico nos permite observar la distribución de los datos mediante un histograma. Fuente propia

Para respaldar nuestra decisión, recurrimos a una herramienta visual: el gráfico del violín. Este gráfico nos permite no solo observar la distribución de los datos, sino también obtener información detallada sobre la simetría y concentración de valores en nuestra muestra.

La elección de incorporar el gráfico nos ayuda a tener un enfoque integral y una validación certera sobre la muestra extraída.

Figura 7

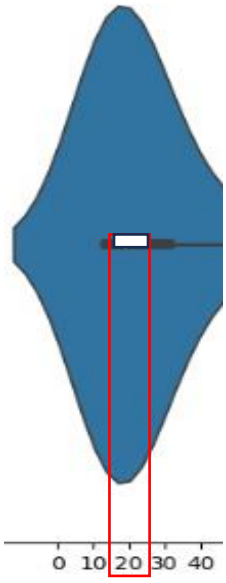
Gráfico de violín para la selección de la muestra



Nota. Este gráfico nos permite observar la distribución de los datos. Fuente propia

Figura 8

Especificación de la muestra en el grafico de violín



Nota. Este grafico nos muestra los datos individuales repetitivos que se encuentran dentro de nuestro dataset. Fuente propia

Como podemos observar en la Figura 8, dentro del gráfico de violín, los puntos, ya sean blancos o negros, generalmente representan datos individuales repetitivos o valores

atípicos en la distribución. En nuestro caso, observamos que la mayoría de los puntos blancos se encuentran entre 12 y 25 que corresponden a los tweets que forman parte de la muestra final de 342 tweets, indicando que la muestra utilizada para el estudio se concentra principalmente en este rango de respuestas.

Puntos Negros: Nos indican valores atípicos o datos individuales que están alejados de la mayoría, pero en este caso, notamos que la mayor densidad de puntos se encuentra entre 10 hasta 50 respuestas.

Punto Blanco: Representa el centro de la distribución de datos y su posición sugiere la concentración principal de valores. En este contexto, refuerza la observación de que la muestra se centra alrededor de 12 respuestas.

Extracción de followings

Con el objetivo de inferir las relaciones que tienen los usuarios en la red a partir de la información que se propaga, realizamos la extracción de los seguidos (followings) de cada usuario que publicó un tweet de la muestra. Para ello, primero realizamos una comparación de herramientas que normalmente son empleadas por los investigadores para este fin. A continuación, se presenta una tabla comparativa de las herramientas Tabla 7:

Tabla 7

Tabla comparativa herramientas para web scraping.

Herramienta	Características Principales	Funcionalidad Destacada
Octoparse	- Extracción visual de datos.	- Permite realizar scraping mediante una interfaz gráfica sin necesidad de programación.
ParseHub	- Capacidad de navegación web.	- Facilita la extracción de datos a partir de páginas web con su interfaz gráfica.
Import.io	- Extracción de datos en páginas web complejas.	- Ofrece funcionalidades avanzadas para extraer datos de sitios web complejos.

Diffbot	- Extracción de datos estructurados de páginas web.	- Proporciona una API para obtener datos estructurados de páginas web.
Phantombuster	- Automatización de tareas repetitivas en la web.	- Permite la automatización de diversas acciones en la web con su interfaz gráfica.
Content Grabber	- Extracción avanzada de datos y navegación automatizada.	- Ofrece capacidades avanzadas para extraer datos y automatizar flujos de trabajo.
WebHarvy	- Extracción de datos en modo punto y clic.	- Permite la extracción de datos de manera intuitiva mediante clics en la interfaz.

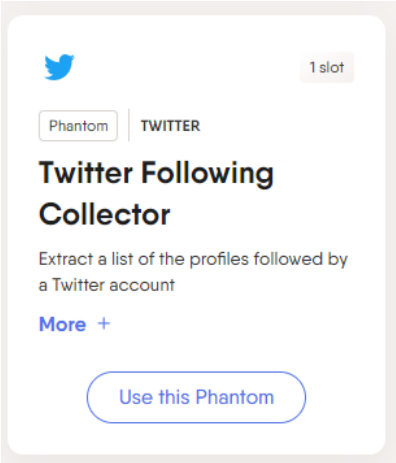
Nota. Esta tabla muestra la comparativa entre herramientas de web scraping y su funcionalidad.

Una vez realizada la comparación optamos por utilizar PhantomBuster para la extracción de seguidos (followings).

Con PhantomBuster que es una herramienta externa a la API de Twitter (X) permite a los usuarios crear "dashboards" para realizar diversas acciones en la plataforma, como la extracción de datos, la interacción con usuarios, el seguimiento y el envío de mensajes automáticos en Twitter (X) (PhantomBuster, 2024), con ello procedimos a extraer los seguidos (followings) de cada usuario, este procedimiento se realiza a través de un panel predeterminado de la herramienta denominado "tweets following". De esta manera, procedemos a extraer los seguidos (followings) tanto de los nodos que generaron el tweet, como de los nodos, que participaron en el hilo de respuestas al tweet.

Figura 9

Dashboard del aplicativo phantombuster “tweets following collector”.



Nota. Este panel de control nos ayuda a extraer los tweets tando del nodo principal como de los secundarios.

Mediante este panel de control Figura 9, ingresamos el enlace de Twitter (X) de cada usuario, y la herramienta extrae los seguidores, almacenándolos en un archivo .csv. Este archivo posteriormente nos servirá como medio para validar la existencia de relaciones entre los participantes del hilo, a continuación, se presenta el proceso a seguir:

- Configuración del Script: Se utiliza Phantombuster para crear un script personalizado que establece los parámetros de la extracción. Estos parámetros pueden incluir la URL del perfil de Twitter (X), el número máximo de seguidos(followings) a extraer, y otros criterios relevantes.
- Autenticación: Es posible que se requiera autenticación a través de credenciales de Twitter (X) para acceder a ciertos nodos y recopilar información sobre los seguidos(followings). Phantombuster puede gestionar este proceso de autenticación según las configuraciones proporcionadas.
- Ejecución del Script: Una vez configurado, el script se ejecuta en la plataforma de Phantombuster. Durante la ejecución, el script navega automáticamente por la

página del perfil de Twitter (X), recopilando información sobre los usuarios seguidos por el perfil en cuestión.

- Extracción de Datos: Phantombuster extrae datos relevantes, como nombres de usuario, nombres completos, biografías y otros detalles de los seguidos(followings) del perfil.
- Almacenamiento de Datos: Los datos recopilados se almacenan en un formato estructurado, como un archivo .csv en nuestro caso.

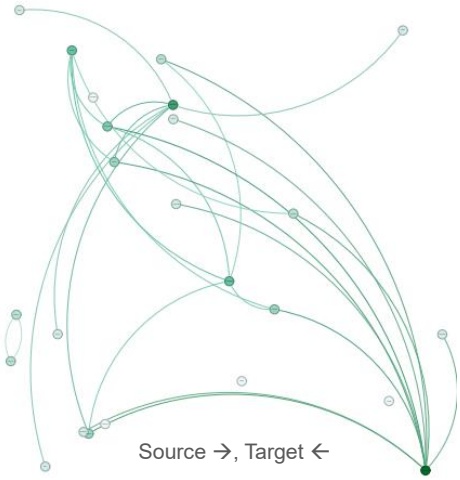
Generación de redes

Iniciamos el proceso de validación del archivo .csv extraído, el cual alberga información relativa a los seguidos de cada usuario. Para ello utilizamos la librería Pandas en Python, su principal estructura de datos, el DataFrame, facilita la manipulación y exploración de información (Pandas, 2024). Con Pandas, podemos cargar datos desde archivos .csv, Excel u otras fuentes, y realizar operaciones eficientes como filtrar, agrupar y combinar datos (Pandas, 2024). Es una herramienta versátil para trabajar con conjuntos de datos estructurados y una elección destacada en análisis de datos en Python. Después utilizamos la librería os que nos ofrece funcionalidades para interactuar con el sistema operativo. Con ella, podemos realizar operaciones relacionadas con archivos y directorios, como listar archivos en un directorio específico, unir rutas de archivos y verificar la existencia de archivos. os es esencial para gestionar archivos y directorios proporcionándonos control sobre la ubicación y acceso a tus datos (Os, 2024).

Mediante la aplicación de estas librerías, hemos identificado tres atributos para la generación de las redes: "source" (es el nodo remitente de la información), "target" (es el nodo receptor de la información) y "type" (indica si la relación es directa o indirecta). En la Figura 10, mostramos una de las redes generadas.

Figura 10

Ejemplo ilustrativo de una de las redes generadas.



Nota. Red generada a través de las librerías de Python en donde se puede observar los atributos identificados como source y target.

La relación entre dos nodos que en las redes generadas representan a usuarios proporciona al usuario seguidor la capacidad de recibir las actualizaciones y publicaciones del usuario seguido en tiempo real, en este sentido definimos lo siguiente:

- Un nodo es source si tiene conexiones salientes. Esto implica que el nodo está emitiendo información.
- Un nodo es source y target cuando tiene conexiones salientes, pero también tiene conexiones entrantes. En otras palabras, este nodo emite y recibe información.
- Un nodo es target si tiene conexiones entrantes. Estos implican que el nodo está recibiendo información.

La aplicación de estos criterios nos permite clasificar los nodos en la red según su papel en la transmisión de información.

Este enfoque nos proporciona una visión detallada de cómo los nodos participan en la propagación de información en la red. Identificar nodos según estos criterios es esencial para comprender la dinámica de la difusión de contenido y las relaciones entre los usuarios en Twitter (X).

Modelo de usuario

El modelo de usuario propuesto en este trabajo se basa en el modelo definido por (Jerez-Villota, Jurado, & Moreno-Llorena, 2023). Cuyo objetivo es entender el rol del usuario en el proceso de propagación de información y como modelo ilustrativo utiliza Twitter (X). El modelo en que nos basamos está dividido en cuatro categorías las cuales describimos a continuación:

- Atributos del Perfil: Estos atributos proporcionan información sobre las características y detalles del perfil de un usuario, como la biografía, la ubicación, la cantidad de seguidores, entre otros. Extraer estos atributos permite construir nodos de relacionamiento detallados que facilitan la comprensión de quiénes son los usuarios y qué aspectos de su perfil pueden influir en su comportamiento en la plataforma.
- Atributos de la Red Social: Estos atributos incluyen información sobre la estructura y las conexiones en la red social. Al extraer estos atributos, se puede analizar la topología de la red, identificar nodos influyentes, comprender la dinámica de las conexiones y evaluar la importancia relativa de los usuarios en la propagación de información.
- Atributos de Acciones del Usuario: Capturan las acciones que los usuarios realizan en la plataforma, como retweets, me gusta, respuestas, entre otras. Estos atributos permiten entender cómo interactúan los usuarios con el contenido, qué tipo de

información encuentran relevante y cómo participan en la difusión de contenido dentro de la red social.

- Sentimientos y Emociones en Acciones del Usuario: La extracción de sentimientos y emociones asociadas con las acciones de los usuarios proporciona una capa adicional de comprensión. Permite identificar cómo reaccionan los usuarios ante ciertos tipos de contenido, qué emociones predominan en sus interacciones y cómo estos aspectos afectan la percepción general de la información compartida.

El modelo de usuario que proponemos se enfoca exclusivamente en la extracción de tweets y retweets, ya que las nuevas políticas de Twitter (X) han limitado la capacidad de extraer respuestas directas. Al ajustarnos a estas restricciones, buscamos adaptar nuestro enfoque de recopilación de datos a las reglas actuales de la plataforma.

Las razones detrás de esta modificación incluyen:

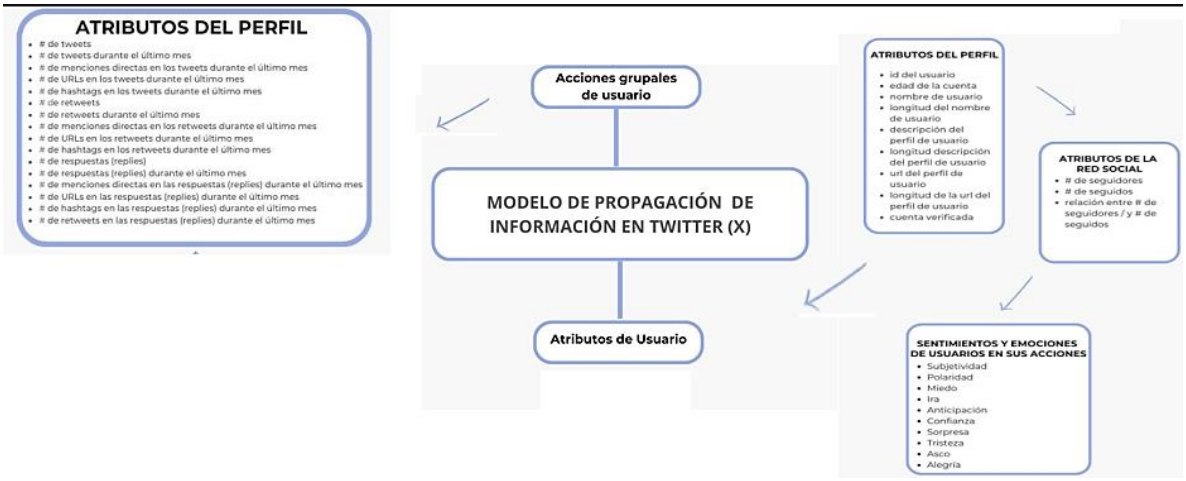
- Cambios en las Políticas de Twitter (X): Las recientes actualizaciones en las políticas de Twitter (X) han restringido la extracción de respuestas directas, lo que hace necesario ajustar nuestro modelo para cumplir con estas nuevas regulaciones (Twitter, 2023).
- Privacidad y Cumplimiento Normativo: El enfoque exclusivo en tweets y retweets se alinea con las preocupaciones de privacidad y las normativas actuales, evitando posibles conflictos éticos o legales asociados con la recopilación de datos sensibles (Twitter, 2023).
- Simplicidad y Eficiencia en el Análisis: Al prescindir de respuestas directas, simplificamos el proceso de análisis, enfocándonos en interacciones más directas y fáciles de interpretar, como tweets y retweets.

- Adaptación a las Condiciones Actuales: Este ajuste garantiza que nuestro modelo sea compatible con las condiciones cambiantes de Twitter (X), permitiéndonos seguir obteniendo información valiosa de la plataforma mientras cumplimos con las normativas vigentes.

Con esta propuesta, buscamos no solo optimizar nuestro modelo en función de la calidad de los datos, sino también asegurar la conformidad con las nuevas políticas establecidas por Twitter (X).

Figura 11

Modelo de usuario propuesto



Nota. Modelo de usuario propuesto.

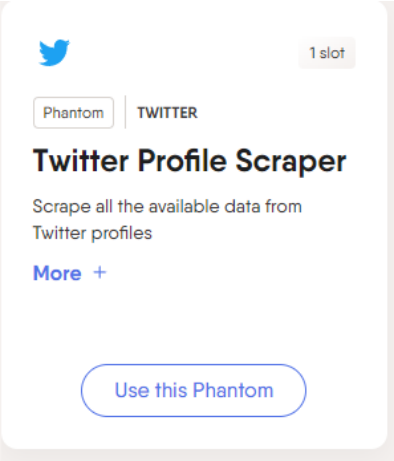
Extracción de metadatos del perfil de usuario

La extracción de metadatos del perfil de usuario es una práctica clave para comprender y analizar la dinámica de la red social. Los datos extraídos con herramientas como Phantombuster nos ayuda a entender el uso de scripts personalizables para automatizar acciones repetitivas y recuperar información específica (PhantomBuster, 2024).

En este caso, se emplea para extraer datos detallados de los nodos y de la actividad en tweets y retweets.

Figura 12

Dashboard "Twitter Profile Scraper"



Nota. Dashboard dentro del aplicativo Phantombuster para extraer los datos del perfil de usuario.

Mediante este panel de control, ingresamos el enlace de Twitter (X) de cada usuario, y la herramienta extraerá las características, almacenándolos en un archivo .csv, que proporcionan información valiosa sobre los usuarios y sus interacciones en la plataforma. A continuación, se explica por qué se extraen estos datos y cómo se utilizan:

Datos del Perfil de Usuario (User Data):

- user_name: El nombre de usuario es fundamental para identificar y etiquetar a cada usuario de manera única en la red.
- age_account: La antigüedad de la cuenta refleja la duración desde la creación del perfil, lo que puede ser indicativo de la experiencia y la consistencia del usuario en la plataforma.

- lenght_username: La longitud del nombre de usuario puede proporcionar información sobre la brevedad o complejidad del identificador.
- length_desc_userprofile: La longitud de la descripción del perfil puede indicar la profundidad de la información que un usuario comparte sobre sí mismo.
- length_url_userprofile: La longitud de la URL del perfil puede revelar la complejidad o simplicidad de los enlaces compartidos.
- verified_account: Indica si la cuenta está verificada, lo cual es importante para establecer la autenticidad del usuario.
- #_followers, #_followings: El número de seguidores y seguidos proporciona información sobre la popularidad y la red de conexiones del usuario.
- #_followers/#_followings: Esta métrica da una relación entre seguidores y seguidos, ofreciendo un panorama de la influencia relativa del usuario en la plataforma.

Estos datos de perfil son esenciales para entender la audiencia de un usuario, su credibilidad y su comportamiento en la plataforma.

Datos de Actividad en Tweets y Retweets:

- #tweets, #duringlastmonth: El número total y el número durante el último mes de tweets del usuario.
- #directmentionstweet, #duringlastmonth: Cantidad de menciones directas en los tweets, tanto en total como en el último mes.
- #URLstweet, #duringlastmonth: Número de URLs compartidas en tweets, tanto en total como en el último mes.
- #hashtagstweet, #duringlastmonth: Cantidad de hashtags utilizados en tweets, tanto en total como en el último mes.

- #retweets, #duringlastmonth: Número total y durante el último mes de retweets realizados por el usuario.
- #directmentionsretweet, #duringlastmonth: Cantidad de menciones directas en los retweets, tanto en total como en el último mes.
- #URLsretweet, #duringlastmonth: Número de URLs compartidas en retweets, tanto en total como en el último mes.
- #hashtagsretweet, #duringlastmonth: Cantidad de hashtags utilizados en retweets, tanto en total como en el último mes.

Estos datos de actividad revelan patrones de comportamiento del usuario, su nivel de participación, el alcance de sus mensajes y la naturaleza de sus interacciones:

- Input: es todo el contenido que los usuarios podrían leer desde sus perfiles, es decir, es todo el contenido que publican sus seguidos (followings), como tweets retweets y replies.
- Output: es todo el contenido que los usuarios publican como tweets, retweets y replies.

Extracción de metadatos de los tweets

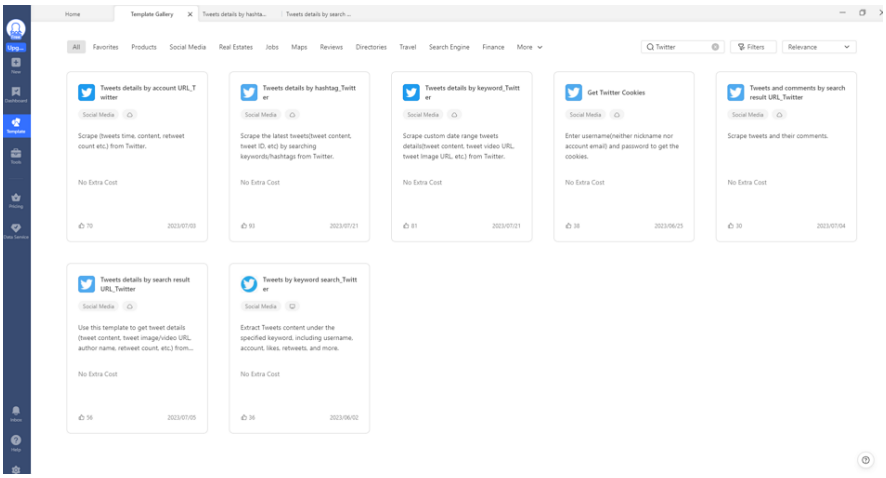
Una vez extraídos todos los nodos de relacionamiento de cada usuario en la plataforma, procedemos a generar archivos de texto plano individualizados para cada hilo de tweet. Este método eficaz nos brinda la capacidad de identificar fácilmente el nodo principal, el origen de la información, y hacia dónde se dirige. Este enfoque no solo simplifica la comprensión de la estructura de la red, sino que también aporta claridad al representar y analizar la información transmitida a través de la plataforma.

La obtención de estos metadatos se realiza utilizando Octoparse, que es una herramienta de scraping visual que facilita la extracción de datos de la web sin la necesidad

de programación. Octoparse nos ayuda a recopilar la información necesaria para generar los archivos de texto plano .txt de cada usuario (Octoparse, 2024).

Figura 13

Dashboard "Twitter from Octoparse"



Nota. Dashboard dentro del aplicativo Octoparse para extraer los metadatos de los tweets.

La generación de estos archivos de texto plano se lleva a cabo para cada usuario como se explica anteriormente, permitiendo una mejor organización y categorización de los datos extraídos. Esto nos facilita la interpretación, y también nos posibilita la aplicación de técnicas analíticas basadas en el modelo de usuario propuesto.

Una vez obtenidos estos datos, procedemos a analizar los archivos .txt sin antes limpiarlos para ello debemos estandarizar y simplificar el texto extraído con esto hemos tomado en cuenta las siguientes estandarizaciones:

- Eliminamos menciones (@usuario) para centrarnos en el contenido principal del tweet y no en nombres de usuario.

- Eliminamos hashtags (#) para simplificar el texto y centrarnos en el contenido del tweet sin distracciones.
- Eliminamos la etiqueta de retweet (RT) y espacios asociados para eliminar información redundante y centrarnos en el contenido original.
- Eliminamos enlaces web para evitar que afecten el análisis de sentimientos o emociones y mantener el foco en el texto principal.
- Eliminamos dos puntos seguidos de espacios para mejorar la legibilidad del texto y eliminar caracteres innecesarios.
- Eliminamos comillas simples seguidas de espacios para mantener la coherencia en el formato del texto.
- Eliminamos puntos suspensivos repetidos para reducir la redundancia y mejorar la comprensión del texto.
- Eliminamos emoticones comunes para reducir el ruido en el análisis de sentimientos y emociones.

Estas acciones de limpieza tienen como objetivo simplificar y estandarizar el texto, eliminando elementos que podrían introducir ruido o afectar la precisión al extraer metadatos derivados de tweets y retweets.

Luego utilizamos las bibliotecas de Python: TextBlob, NLTK y NRClex para el análisis de sentimientos y emociones.

TextBlob: Esta biblioteca de procesamiento de lenguaje natural (NLP) realiza análisis de sentimientos, proporcionando métricas de polaridad y subjetividad. La polaridad indica la positividad o negatividad del texto, mientras que la subjetividad mide cuán subjetivo u objetivo es el contenido (TextBlob, 2023).

NLTK (Natural Language Toolkit) y NRClex: Utilizadas para procesamiento de lenguaje natural, ofrecen herramientas para analizar el texto en términos de emociones. Utilizando métricas específicas, nos permiten identificar y clasificar expresiones emocionales en el contenido de los tweets (NLTK-NRClex, 2023).

NLTK y NRClex utilizan modelos de análisis de emociones para clasificar las expresiones emocionales presentes en el texto en diversas categorías.

La combinación de estas herramientas en el análisis de los archivos .txt no solo nos proporciona una visión más profunda de la información transmitida, sino que también nos permite explorar y cuantificar los aspectos emocionales y sentimentales asociados a dicha información.

Cálculo de metadatos derivados

Las métricas específicas que utilizan las bibliotecas de Python TextBlob, NLTK y NRClex en el análisis de los archivos .txt son:

TextBlob:

Polaridad:

Rango: [-1, 1]

Indica la positividad o negatividad del texto. Un valor positivo sugiere un tono positivo, mientras que un valor negativo indica un tono negativo.

Subjetividad:

Rango: [0, 1]

Mide cuán subjetivo u objetivo es el contenido. Un valor cercano a 0 indica objetividad, mientras que un valor cercano a 1 sugiere subjetividad.

NLTK y NRCLex:

Emociones:

Rango: [0, 1]

Utiliza modelos de análisis de emociones para clasificar las expresiones emocionales presentes en el texto con categorías específicas estas son: Feliz, Triste, Enojado, Sorprendido, Miedo, Disgustado, Neutral, entre otras.

Es importante destacar que estas categorías son resultados de los modelos y enfoques específicos de cada biblioteca. Además, la interpretación precisa de las categorías puede variar según el contexto y la naturaleza del texto analizado.

Preprocesamiento

Generación de vectores

La matriz de vectores está compuesta por diversos campos que representan el modelo de usuario propuesto. A continuación, realizaremos una breve descripción de como estaría compuesta la matriz de vectores:

- Datos de Usuario:

  Nombre de usuario, antigüedad de la cuenta, longitud del nombre de usuario, longitud de la descripción del perfil, longitud de la URL en el perfil, indicador de cuenta verificada, número de seguidores, número de cuentas seguidas, y la relación de seguidores respecto a cuentas seguidas.
- Metadatos Derivados de Tweets y Retweets:

  Número total de tweets, número de tweets durante el último mes, número de menciones directas en tweets, número de URLs en tweets, número de hashtags en tweets, número total de retweets, número de menciones directas en retweets, número de URLs en retweets, y número de hashtags en retweets.

- Análisis de Sentimientos y Emociones (Entrada y Salida):

  Métricas de análisis de sentimientos y emociones para los tweets y retweets del usuario (entrada), así como métricas para los tweets y retweets dirigidos al usuario (salida) estas métricas serian: subjetividad, polaridad, miedo, enojo, anticipación, confianza, sorpresa, tristeza, repugnancia, alegría.

Con esta matriz, podríamos identificar y analizar patrones, comportamientos y características específicas de los usuarios en la red social Twitter (X).

Normalización de vectores.

En el contexto de nuestra investigación, la normalización de datos se llevará a cabo como parte del proceso analítico para ello ocupamos la librería de Python (Scikit-learn, Scikit-learn, 2024) proporciona herramientas para normalizar datos mediante su módulo preprocessing. Este procedimiento se centra en ajustar las magnitudes de las características en la matriz de datos, específicamente aquellas que contienen información numérica.

En el proceso de selección de características relevantes para la investigación, se identifican columnas con datos numéricos pertinentes, como recuentos de interacciones y números de seguidores. Luego, se aplica la normalización para ajustar los valores y asegurar una escala similar, evitando sesgos en el análisis. La matriz resultante refleja las actualizaciones, permitiendo la comparación equitativa entre nodos de relacionamiento de usuario. La normalización juega un papel crucial al mejorar la precisión de los modelos y facilitar la interpretación coherente de resultados, contribuyendo a una evaluación más precisa de los nodos en redes sociales.

Procesamiento

Clasificación de nodos

Después analizar los trabajos relacionados, procedemos a efectuar una comparación detallada de las técnicas más recurrentemente empleadas por los autores en sus respectivas investigaciones. A continuación, se presenta una tabla comparativa que abarca las técnicas más utilizadas Tabla 9:

Tabla 8

Tabla comparativa de las técnicas de clasificación.

Técnicas	Descripción	Funcionalidad	Libreria
KMeans	Algoritmo de agrupamiento basado en centroides	Agrupa datos en k clústeres según la similitud de sus características.	scikit-learn
Cuantitativo y Coseno	Métricas cuantitativas y medida de similitud coseno	Proporciona medidas numéricas y similitud coseno para análisis de datos.	NumPy, Pandas, SciPy
Método del Codo	Técnica de determinación óptima del número de clústeres	Ayuda a identificar el número óptimo de clústeres en un conjunto de datos usando la variabilidad intraclúster.	scikit-learn

Nota. Esta tabla muestra la comparativa entre las técnicas de clasificación para identificar los nodos.

Después de analizar las diversas librerías y técnicas para la clasificación de los nodos de relacionamiento encontrados, hemos concluido que, con base en sus funcionalidades, optaremos por emplear Kmeans y el método del codo.

El algoritmo K-means es una técnica popular de agrupación (clustering) que se utiliza para dividir un conjunto de datos en k grupos distintos basándose en características similares (Grant RW, 2020). El método del codo (elbow method) es una técnica utilizada para encontrar el número óptimo de clusters (k) en un conjunto de datos (Syakur, 2018).

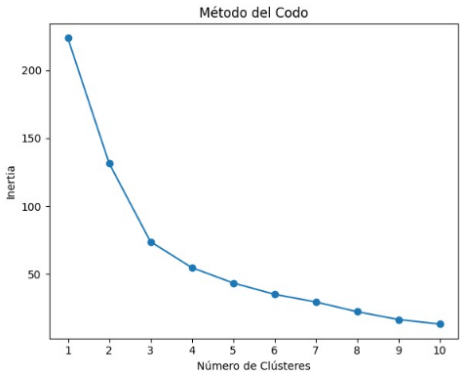
Como parte inicial aplicaremos el método del codo ya que nos permitirá saber el valor exacto de k para la clasificación de los nodos de relacionamiento, para cada valor de

k se calcula la suma de los cuadrados de las distancias intra-cluster (WCSS, por sus siglas en inglés). El WCSS representa qué tan compactos son los grupos (clusters) (Samuel & Kridanto, 2019).

Se grafica los valores de k en el eje X y los correspondientes valores de WCSS en el eje Y. En la Figura 14 se busca el "codo" o el punto donde la disminución en WCSS comienza a desacelerarse significativamente. Este punto indica el número óptimo de clusters.

Figura 14

Método gráfico del codo.



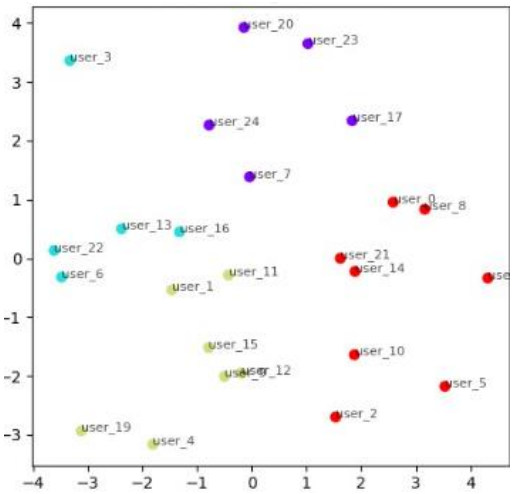
Nota. Gráfica representativa utilizando el método del codo mediante Python.

Después utilizaremos el algoritmo K-means que nos ayudará agrupar a los usuarios en k clusters. Este valor de k se determinará una vez se aplique el método del codo, para esto, utilizaremos este algoritmo mediante Python con la librería scikit-learn esta es una biblioteca de aprendizaje automático de código abierto que proporciona implementaciones eficientes y fáciles de usar de varios algoritmos de aprendizaje automático, incluido K-means (Scikit-learn, 2024).

Introducimos los vectores normalizados para ejecutar este algoritmo dentro de los datos normalizados. Al momento de ejecutar nuestro algoritmo, el modelo ajusta los centroides de los clústeres para minimizar la suma de las distancias al cuadrado entre los nodos de relación y el centroide de su clúster asignado. Después de ajustar el modelo, podemos obtener un gráfico que representa la clasificación de los nodos dentro de la red, como se muestra en la Figura 15, que representa la agrupación de los nodos de relación en cada centroide generado.

Figura 15

Gráfico de clasificación



Nota. Grafica de clasificación mediante Python utilizando la librería scikit-learn.

Análisis de datos

Definición de perfiles de usuario

En esta fase nos centraremos en descubrir características similares en el comportamiento de los nodos en el proceso de propagación de información lo que nos dará como resultado la definición de perfiles de usuario. Para ello nos basamos en los resultados

obtenidos en el proceso de clasificación ver Anexo 1. Por lo tanto, observamos las características importantes de cada grupo de relacionamiento como en la Figura 16.

Figura 16

Clasificación de los nodos

Agrupamiento	Caracteristica 1	Caracteristica 2	Caracteristica 3
Cluster 0	out_sadness: 0.75	out_surprise: 0.71	length_desc_userprofile: 0.6
Cluster 1	verified_account: 1.00	in_subjectivity: 0.97	out_anger: 0.82
Cluster 2	in_fear: 0.76	in_anticip: 0.72	in_anger: 0.72
Cluster 3	#URLstweet: 0.98	in_disgust: 0.87	#URLsretweet: 0.85

Nota. Gráfica de los nodos de relacionamiento según sus características.

Mediante el análisis de los gráficos generados por cada red observamos características similares en ellas y extrajimos la frecuencia con la que se repiten obteniendo así los siguientes perfiles de usuario:

- Influenciador Verificado: Este nombre refleja la autenticidad de la cuenta, indicada por la verificación, y destaca la capacidad de esta persona para influir en la audiencia de Twitter. La designación "verificado" añade credibilidad a su presencia en la plataforma, mientras que el término "influenciador" sugiere un impacto significativo en el contenido y las interacciones de la comunidad de Twitter. Este perfil probablemente se caracterizaría por una actividad frecuente de publicación de tweets que resuena en la audiencia, consolidando así su posición como un influenciador reconocido en la plataforma.
- Optimista Verificado: Este nombre destaca la autenticidad de la cuenta, indicada por la verificación, y enfatiza la predisposición de la persona a compartir contenidos que reflejan felicidad y optimismo en Twitter. La actividad notable de retuitear sugiere una participación en la difusión de mensajes positivos durante un mes. "Optimista

Verificado" evoca una imagen de una presencia genuina y alegre en la plataforma, contribuyendo a la creación de un espacio positivo y amigable en la red social.

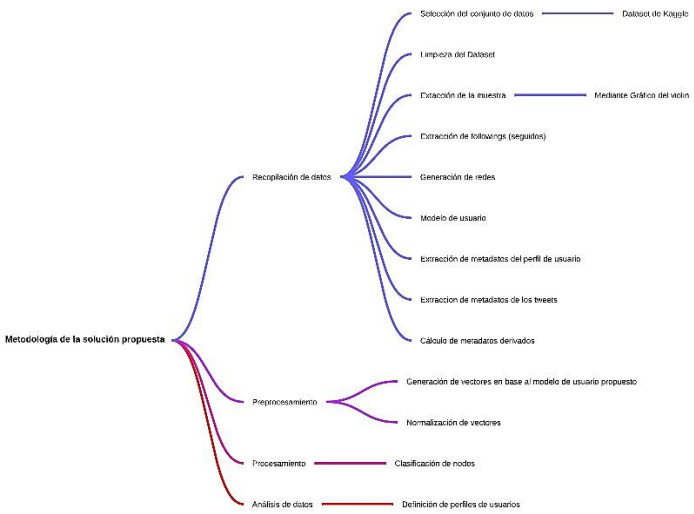
- Influenciador Feliz: Este nombre refleja la felicidad de la persona al leer las opiniones de sus seguidores, transmitiendo una conexión positiva y emocional con su audiencia. La gran cantidad de seguidores sugiere una presencia significativa y una red extensa, lo que refuerza la idea de que esta persona disfruta de una comunidad considerable. "Influenciador Feliz" comunica la imagen de un perfil que se enriquece emocionalmente al interactuar con su audiencia, creando un ambiente positivo y conectado en la plataforma.

Finalmente, en el proceso de generación de perfiles de usuario mediante el método de K-means y la identificación del codo, se logró categorizar eficazmente los nodos en clusters distintos. La selección de características relevantes, la normalización de datos y la aplicación de K-means permitieron definir perfiles claros.

La interpretación de los resultados reveló perfiles diferenciados, cada uno con sus propias características distintivas. Los nombres de los perfiles se asignaron considerando las características comunes dentro de cada grupo, lo que facilita la identificación y comprensión de las tendencias específicas. La metodología empleada ver Figura 17 permitió una clasificación efectiva de los nodos, proporcionando perfiles significativos que reflejan la diversidad de comportamientos dentro de la red social Twitter (X).

Figura 17

Metodología de la solución propuesta



Nota. Este gráfico, nos permite observa la metodología que implementaremos para nuestro trabajo de investigación.

Capítulo IV

Conclusiones y recomendaciones

Conclusiones

Tras la culminación del presente trabajo de investigación se puede concluir, de manera general, que el principal objetivo de esta fue cumplido, de forma específica y referenciando a los objetivos específicos de este proyecto, se puede también obtener las siguientes conclusiones:

- Importancia de la Detección de Información No Verificada: El fenómeno de la difusión de información no verificada en Twitter (X) destaca la importancia de desarrollar herramientas eficientes para la recolección y análisis de datos. La propagación de noticias no verificadas puede tener impactos significativos en la sociedad, afectando la toma de decisiones individuales y colectivas. Por lo tanto, es

crucial abordar este problema mediante la identificación de alternativas a la API de Twitter (X) para recopilar datos de manera efectiva.

- Necesidad de Herramientas Alternativas: La limitación de la API de Twitter (X) y los cambios recientes en la plataforma resaltan la necesidad de buscar alternativas para la recolección de datos. La investigación subraya la importancia de explorar y desarrollar herramientas que ofrezcan un mayor alcance y facilidad de uso e instalación, permitiendo así un análisis más profundo de las interacciones y acciones de los usuarios en Twitter (X).
- Rol Estratégico de los Nodos de Relacionamiento: La generación de nodos de relacionamiento se revela como una estrategia valiosa para comprender el comportamiento en línea de los usuarios en Twitter (X). Estos nodos no solo permiten rastrear las interacciones evidentes, como tweets y retweets, sino que también son esenciales para prever el comportamiento futuro de los usuarios, incluyendo preferencias políticas y respaldo a causas sociales.
- Predicción de Tendencias y Comportamientos: La capacidad de anticipar el comportamiento futuro de los usuarios, basada en la generación de perfiles detallados, se convierte en un activo estratégico. Esta predicción no solo facilita la adaptación proactiva a las tendencias emergentes, sino que también permite a las organizaciones y empresas anticipar las necesidades y preferencias de sus audiencias, generando ventajas competitivas en un entorno digital dinámico.
- Contribución a la Investigación Social y Científica: La generación de perfiles de usuario, al proporcionar una visión detallada de las interacciones en línea, se convierte en una valiosa fuente de datos para la investigación social y científica. La información recopilada puede ser utilizada para comprender patrones de comportamiento, cambios en las opiniones públicas y dinámicas sociales, contribuyendo así al avance del conocimiento en diversos campos.

Recomendaciones

Con toda la información obtenida de la solución propuesta se han establecido las siguientes recomendaciones:

- Exploración de Herramientas Alternativas: Dada la dependencia de la API de Twitter (X) y sus limitaciones, se recomienda realizar una exhaustiva exploración de herramientas alternativas para la recolección de datos. Buscar soluciones que ofrezcan mayor alcance, facilidad de uso e instalación, y que se adapten a los cambios dinámicos en la plataforma.
- Desarrollo de Herramienta Personalizada: Considerando el objetivo de modelar las acciones de los usuarios, se recomienda evaluar la posibilidad de desarrollar una herramienta personalizada que se ajuste específicamente a los requisitos del estudio. Esto podría implicar la creación de un sistema propio para la recolección de metadatos de nodos de relacionamiento y tweets sin depender completamente de la API de Twitter (X).
- Énfasis en la Seguridad de Datos: Dada la sensibilidad de la información recopilada de los usuarios, es crucial incorporar medidas sólidas de seguridad y privacidad en el proceso de recolección y almacenamiento de datos. Garantizar la confidencialidad y anonimato de la información recolectada es esencial para cumplir con estándares de las nuevas políticas de Twitter (X).

Trabajos Futuros

El presente trabajo de investigación y su solución propuesta dejan algunas líneas de investigación que pueden ser estudiadas por trabajos futuros. A continuación, se presentan algunos temas que no fueron abordados por exceder el alcance planteado en este proyecto, pero que podrían ser abordados en trabajos a futuro.

- Optimización de Herramientas de Generación de Perfiles: Investigar y desarrollar herramientas más eficientes y precisas para la generación de perfiles de usuarios en redes sociales. Esto podría incluir técnicas avanzadas de procesamiento de lenguaje natural y aprendizaje automático para extraer información detallada de las interacciones en línea.
- Evaluación Continua de Alternativas Tecnológicas: Dada la naturaleza cambiante de las plataformas sociales, se recomienda realizar evaluaciones continuas de alternativas tecnológicas para la recolección de datos en Twitter (X). Mantenerse al tanto de nuevas API, herramientas de análisis y enfoques metodológicos permitirá mantener la relevancia y eficacia en la investigación.
- Desarrollo de Herramientas Avanzadas de Análisis: Se podría explorar la creación de herramientas más avanzadas de análisis de nodos de relacionamiento en Twitter (X). Esto incluiría la implementación de técnicas de inteligencia artificial y aprendizaje automático para identificar patrones de comportamiento más complejos, lo que podría mejorar la precisión en la predicción de preferencias y la propagación de información no verificada.

Bibliografía

Arolfo, F., Rodriguez, K., & Vaisman, A. (2022). Analyzing the Quality of Twitter Data Streams. *Inf Syst Front*, 349–369.

Bashar, M. A. (2022). Deep learning based topic and sentiment analysis: COVID19 information seeking on social media. *Soc. Netw. Anual*, 12-90.

Castillo, M. &. (2011). Information Credibility on Twitter. *The Web Conference*, 1-10.

Chalmers, D. (2011). Rhythms in Twitter. *10.1109/PASSAT/SocialCom.2011.226*, 1409-1414.

Grant RW, M. J. (2020). Use of Latent Class Analysis and k-Means Clustering to Identify Complex Patient Profiles. *JAMA Netw Open*, 10-1001.

Hayawi, K., Mathew, S., & Venugopal, N. (2022). DeeProBot: a hybrid deep neural network model for social bot detection based on user profile data. *Soc. Netw. Anual*, 12-43.

Hoang, T. (2022). Prediction of brand stories spreading on social networks. *Data Anal Classif* , 559–591.

Jerez-Villota, E., Jurado, F., & Moreno-Llorena, J. (2023). Understanding the Role of the User in Information Propagation on Online Social Networks: A Literature Review and Proposed User Model. In International Conference on Ubiquitous Computing and Ambient Intelligence. *Springer Nature Switzerland*, 304-315.

Kaggle. (2024, 02 19). *Kaggle*. Retrieved from Kaggle: https://www.kaggle.com/

method, E. (2024).

Milovanović, S. (2021). Social recruiting: an application of social network analysis for preselection of candidates. *Data Technologies and Applications ahead-of-print*, 192-198.

Mohanty, B. L. (2023). Heterogenous Social Media Analysis for Efficient Deep Learning Fake-Profile Identification. *IEEE Access*, 99339-99351.

NLTK-NRClex. (2023, 10 18). *NLTK-NRClex*. Retrieved from NLTK-NRClex: https://www.nltk.org/

Octoparse. (2024, 01 12). *Octoparse*. Retrieved from Octoparse: https://www.octoparse.com/blog/how-to-extract-data-from-twitte

Os. (2024, 02 19). *Os*. Retrieved from Os: https://docs.python.org/es/3.10/library/os.html

Pandas. (2024, 02 19). *Pandas*. Retrieved from Pandas: https://pandas.pydata.org/

PhantomBuster. (2024, 02 19). *PhantomBuster*. Retrieved from PhantomBuster: https://phantombuster.com/?gs2id=CjsaCQiAqsitBhDlARIsAGMR1RixghUQjuWARa8LoLw9WNnE2i9HzhshG2-dyV_IS3ok6UH048kqGIMaAlbLEALw_wcB&ource=vaneadssgaolpo&edium=vasasdneraukilop2&ampaign=ganikolopare3da&ontent=ikoloete&deal=mikhail77&erm=bananaterjikulop&gad_sour

Š. Grigaliūnas, R. B. (2023). Ontology-Driven Digital Profiling for Identification and Linking Evidence Across Social Media Platform. *IEEE Access*, 111672-111691.

Samuel, A., & Kridanto, S. (2019). Spam Detection on Profile and Social Media Network using Principal Component Analysis (PCA) and K-means Clustering. *Int. J. Advance Soft Compu*, 2074-8523.

Sanjaya, S. (2019). Spam Detection on Profile and Social Media Network . *Int. J. Advance Soft Compu*, 2074-8523.

Scikit-learn. (2024, 20 02). Retrieved from https://scikit-learn.org/stable/

Scikit-learn. (2024, 01 23). *Scikit-learn*. Retrieved from Scikit-learn: https://scikit-learn.org/

Syakur, M. A. (2018). Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster. *IOP Conf*, 1088-1757.

TextBlob. (2023, 07 23). *TextBlob*. Retrieved from TextBlob: https://textblob.readthedocs.io/en/dev/

Twitter, P. (2023, 06 12). *Politicas Twitter*. Retrieved from Politicas Twitter: https://help.twitter.com/es/rules-and-policies

Villegas, W. (2022). Artificial Intelligence Model for the Identification of the Personality of Twitter Users through the Analysis of Their Behavior in the Social Network. *Electronics 2022*, 11-22.

Vosoughi, R. &. (2017). The spread of true and false news online. *American Association for the Advancement of Science*, 1146 - 1151.

Wang, L. (2019). Factor Graph Model Based User Profile Matching Across Social Networks. *IEEE Access*, 52429-152442.

Weishampel, A. (2023). Classification of social media users with generalized functional data analysis. *Computational Statistics & Data Analysis*, 1120–1129.

Wieringa, R. M. (2006). Requirements engineering paper classification and evaluation criteria: A proposal and a discussion. *Springer-Verlag London Limited 2005*, 102-107.

Apéndices