ESPE

UNIVERSIDAD DE LAS FUERZAS ARMADAS

INNOVACIÓN PARA LA EXCELENCIA

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.

NOMBRE: DAVID ALEXANDER DIAZ VILLACIS

TUTOR: ELEANA JEREZ

FECHA: SANGOLQUI, 05 DE FEBRERO DEL 2024

ÍNDICE

ANTECEDENTES

Twitter (X) es vital en la comunicación global, influyendo en opiniones. La urgente necesidad de herramientas eficientes para la recolección de datos en tiempo real destaca. La generación de nodos de relacionamiento en la plataforma es esencial para anticipar preferencias políticas y sociales en un entorno donde la información impacta las decisiones.

PROBLEMA

La difusión de información no verificada en Twitter (X) sobre temas sociales, políticos y económicos es común y afecta el comportamiento colectivo. A pesar de estudios existentes con la API de Twitter (X), cambios recientes y limitaciones destacan la necesidad de identificar herramientas más accesibles para recopilar datos para modelar las acciones que realizan los usuarios al propagar intencionalmente información. .

JUSTIFICACIÓN

Considerando las limitaciones actuales y la importancia de adaptarse a las dinámicas cambiantes, utilizar herramientas o tecnicas que nos permitan modelar a los usuarios es fundamental en el proceso de difusión de información y asi encontrar alternativas para modelar a los usuarios en el proceso de propagación de información e identificar características similares de su comportamiento en la red utilizando técnicas de clasificación.

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

01 Realizar la revisión de literatura de las herramientas para la recolección de metadatos de los perfiles de usuario y tweets.

02 Definir la herramienta para recolectar metadatos de los perfiles de usuario y tweets.

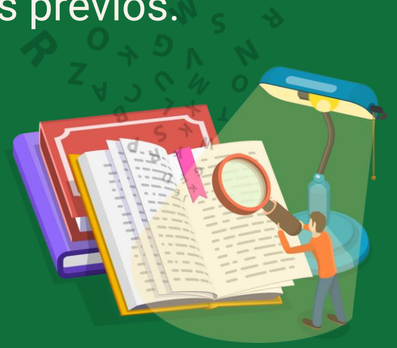
03 Recolectar metadatos de los perfiles de usuario y tweets en Twitter (X).

04 Recolectar metadatos de los perfiles de usuario y tweets en Twitter (X).

MARCO TEÓRICO

Revisión de la literatura

Nos enfocamos en realizar una revisión de la literatura, empleando una metodología orientada a la recopilación y análisis de estudios previos.



PREGUNTAS DE INVESTIGACIÓN

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

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

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

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

05 ¿Cuáles son los tipos de contribución?

RESULTADOS DE LA REVISION DE LITERATURA

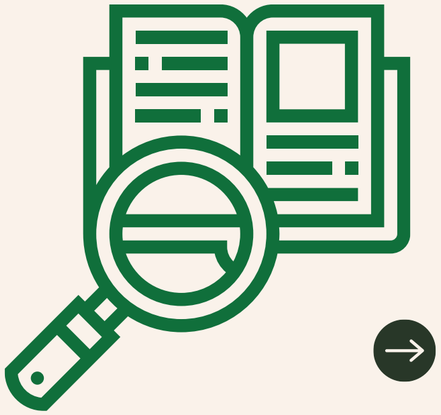
01 Aplicando los criterios de inclusión y exclusión, definidos obtuvimos un total de 13 artículos primarios (trabajos relacionados),

02 Verificando en los trabajos relacionados, se pudo obtener que los datasets mas utilizados son de la plataforma Kaggle.

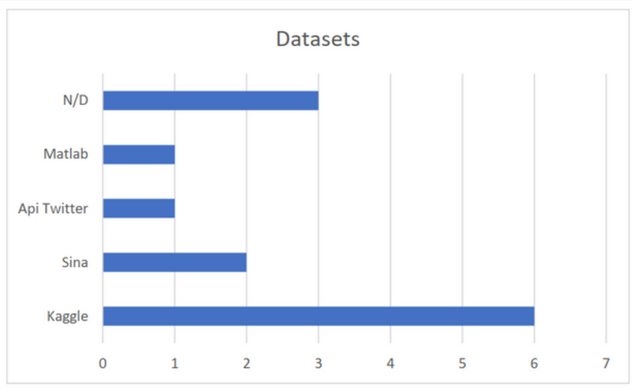
03 Pudimos obtener mediante los trabajos relacionados que las técnicas o herramientas utilizadas se enfocan en Python.

04 La técnica de clasificación que mas se ocupa dentro de los trabajos relacionados fue el aprendizaje no supervisado.

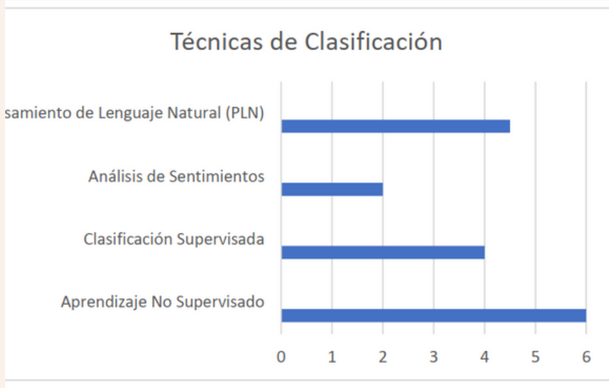
05 Siguiendo la clasificación propuesta en (Wieringa, 2006), el cual en nuestro trabajo hace evidencia que las categorías “Análisis”,"Método" y "Modelo" representan la mayor cantidad de contribuciones.



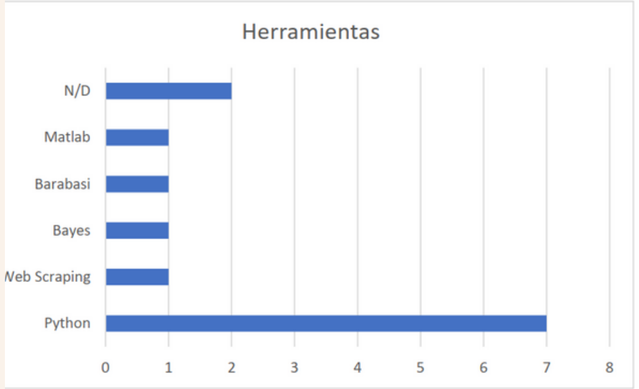
02


Datasets
N/D
Matlab
Api Twitter
Sina
Kaggle
0
1
2
3
4
5
6
7

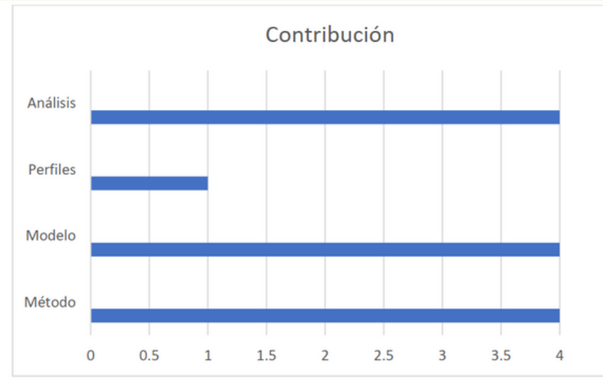
04


Técnicas de Clasificación
samiento de Lenguaje Natural (PLN)
Análisis de Sentimientos
Clasificación Supervisada
Aprendizaje No Supervisado
0
1
2
3
4
5
6

03


Herramientas
N/D
Matlab
Barabasi
Bayes
Web Scraping
Python
0
1
2
3
4
5
6
7
8

05


Contribución
Análisis
Perfiles
Modelo
Método
0
0.5
1
1.5
2
2.5
3
3.5
4

METODOLOGÍA DE LA SOLUCIÓN PROPUESTA

METODOLOGÍA DE LA SOLUCIÓN PROPUESTA

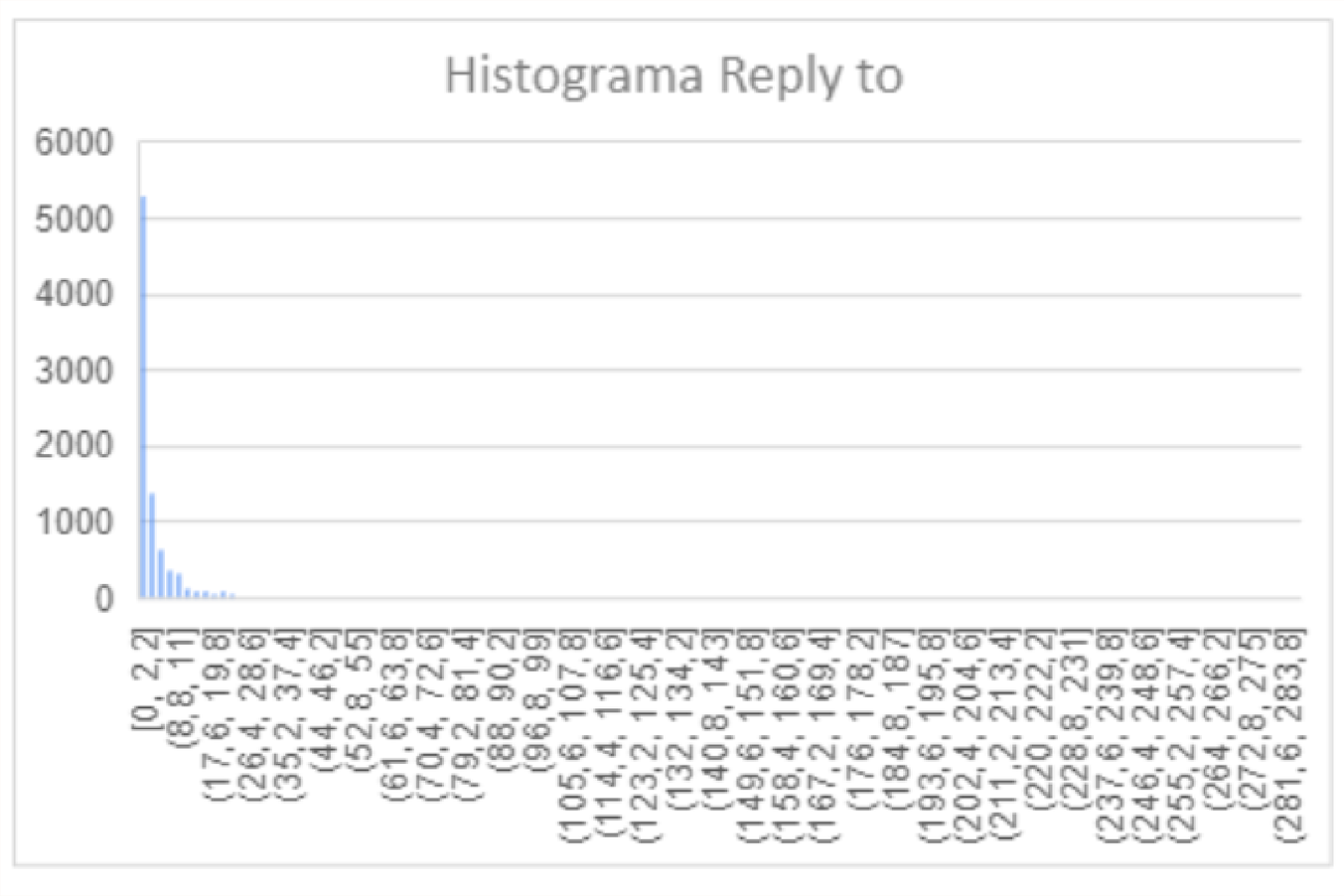
RECOPILACIÓN DE DATOS

SELECCIÓN DEL CONJUNTO DE DATOS

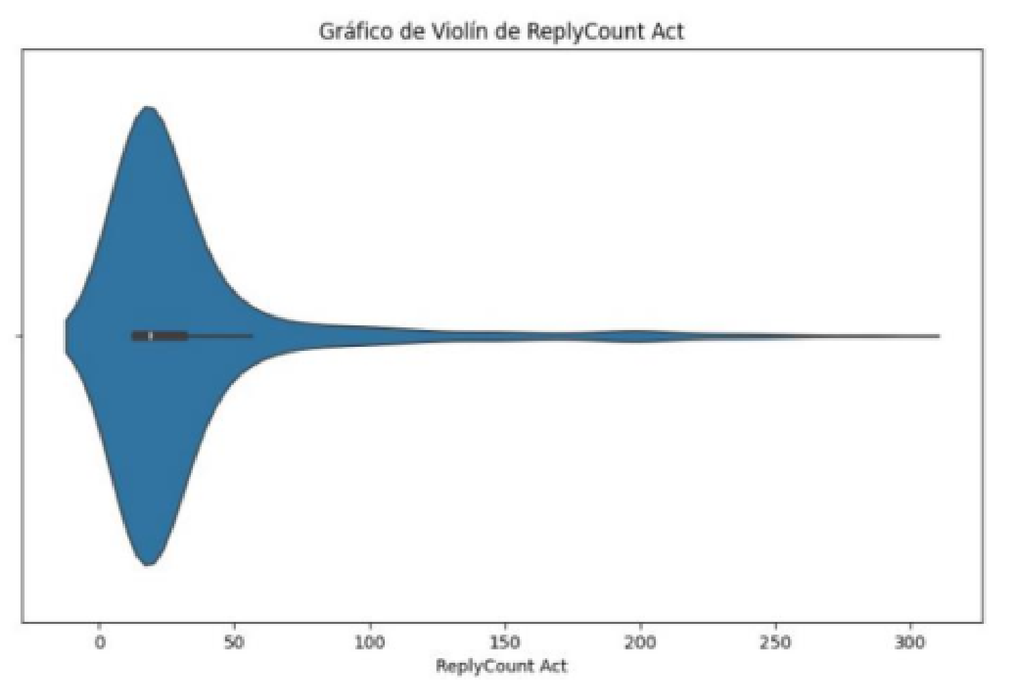
LIMPIEZA DEL DATASET

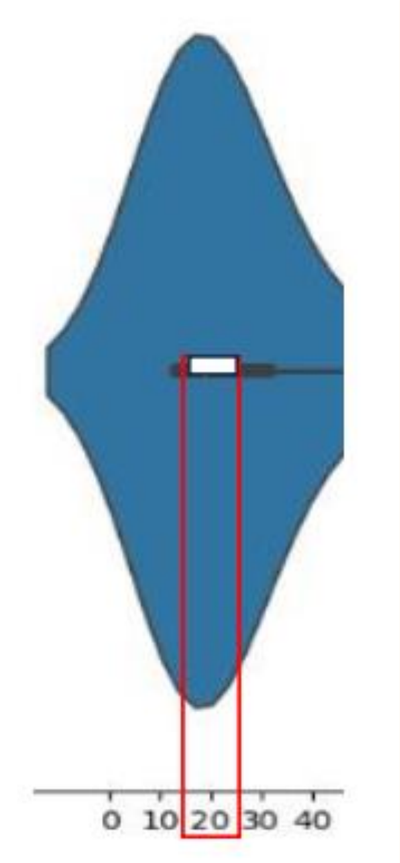
- Identificación de entidades: 200.
- Identificación de nodos de relacionamientos automatizados (robots): 122.
- Identificación de nodos de relacionamientos humanos: 8004.
- Inclusión exclusivamente tweets redactados en inglés.

EXTRACCIÓN DE LA MUESTRA



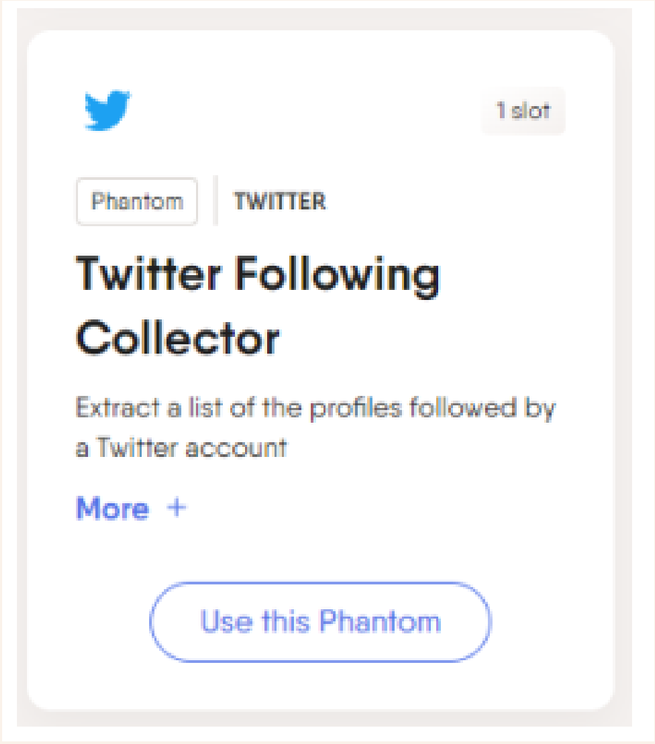
EXTRACCIÓN DE LA MUESTRA





EXTRACCIÓN DE FOLLOWINGS

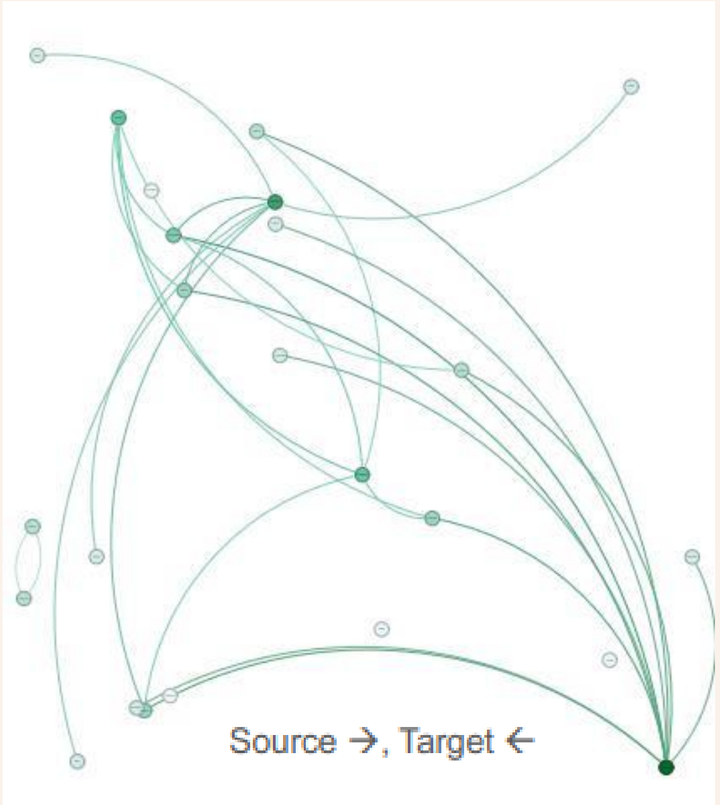
Procedimos a extraer los seguidos (followings) de cada usuario, este procedimiento se realiza a través de un panel predeterminado de la herramienta denominado "tweets following" .



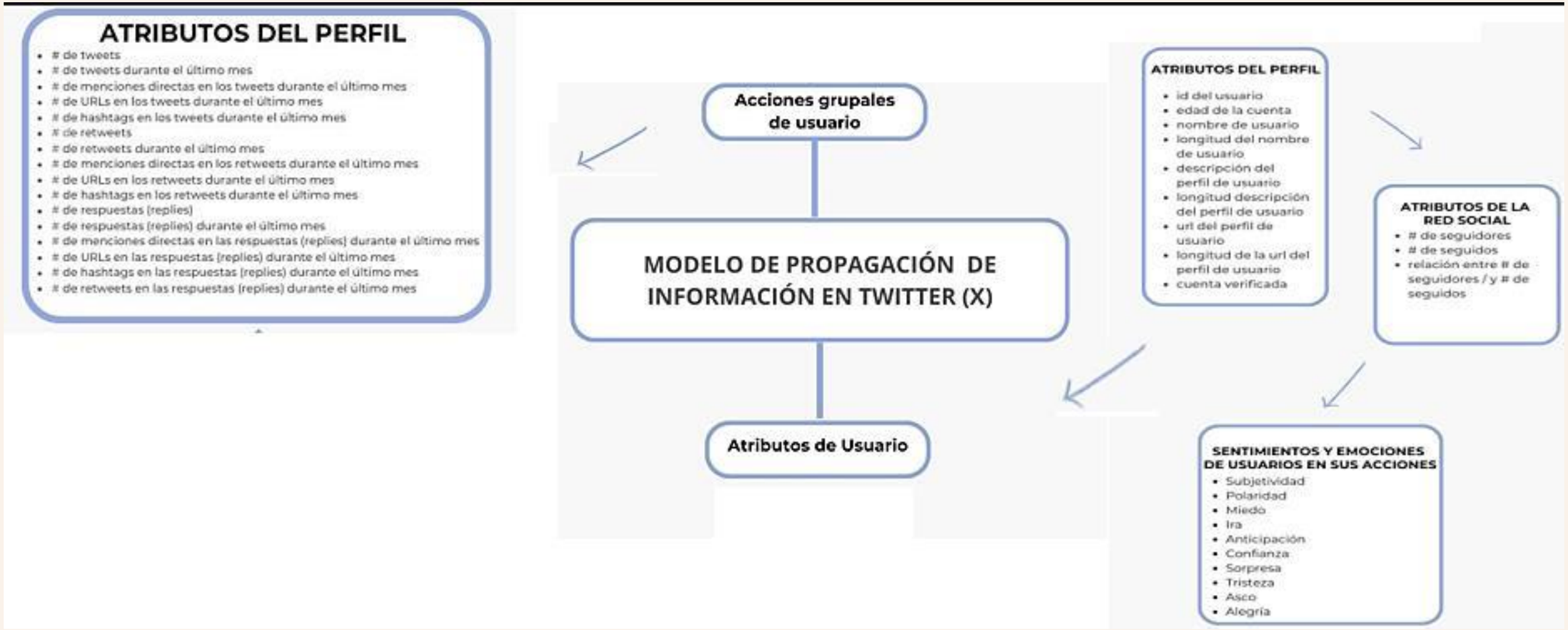
GENERACIÓN DE REDES

EXTRACCIÓN DE FOLLOWINGS

• 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

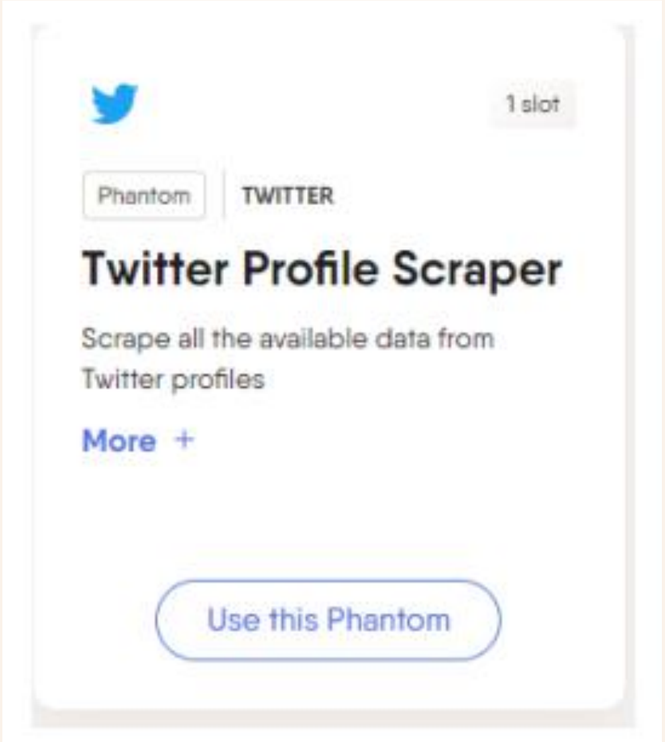


MODELO DE USUARIO

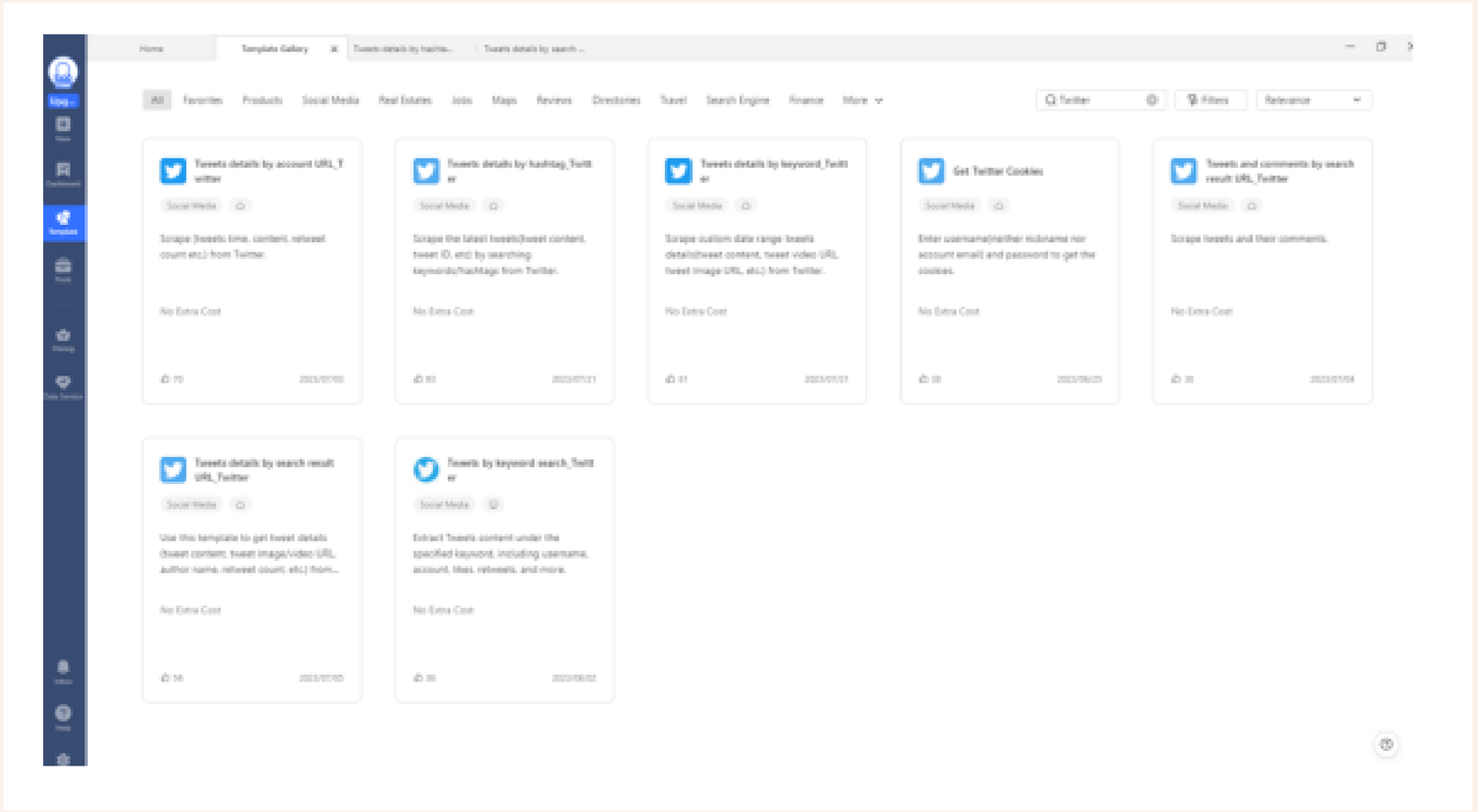


EXTRACCIÓN DE METADATOS DEL PERFIL DE USUARIO

Utilizamos Phatombuster ya que en este caso, se emplea para extraer datos detallados de los perfiles de usuario y de la actividad en tweets y retweets. Para ello utilizamos el dashboard “Twitter Profile Scraper”.



EXTRACCIÓN DE METADATOS DE LOS TWEETS



EXTRACCIÓN DE METADATOS DE LOS TWEETS

Analizamos 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)
- Eliminamos hashtags (#)
- Eliminamos la etiqueta de retweet (RT) y espacios asociados
- Eliminamos enlaces web
- Eliminamos dos puntos seguidos de espacios
- Eliminamos comillas simples
- Eliminamos puntos suspensivos
- Eliminamos emoticones comunes

EXTRACCIÓN DE METADATOS DE LOS TWEETS

Subjetividad y polaridad

Los afectos emocionales medidos incluyen lo siguiente:
miedo, enojo, anticipación, confianza , sorpresa, positivo, negativo, tristeza, asco, alegría

PREPROCESAMIENTO

Generación de vectores

La matriz de vectores está compuesta por diversos campos que representan el modelo de usuario propuesto.

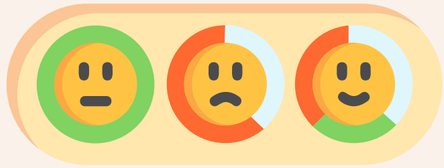
[matrix]

GENERACIÓN DE VECTORES

Datos de Usuario

Metadatos Derivados de Tweets y Retweets

Análisis de Sentimientos y Emociones (Entrada y Salida)



NORMALIZACIÓN DE VECTORES

Se aplica la normalización para ajustar los valores y asegurar una escala similar, evitando sesgos en el análisis.

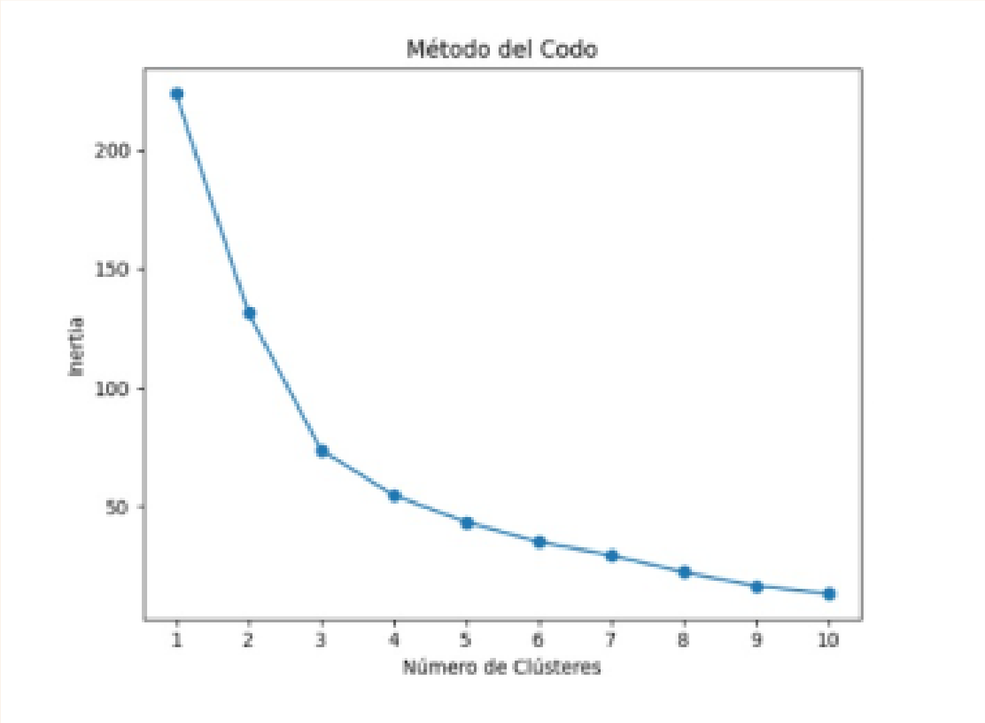
PROCESAMIENTO

CLASIFICACIÓN DE NODOS

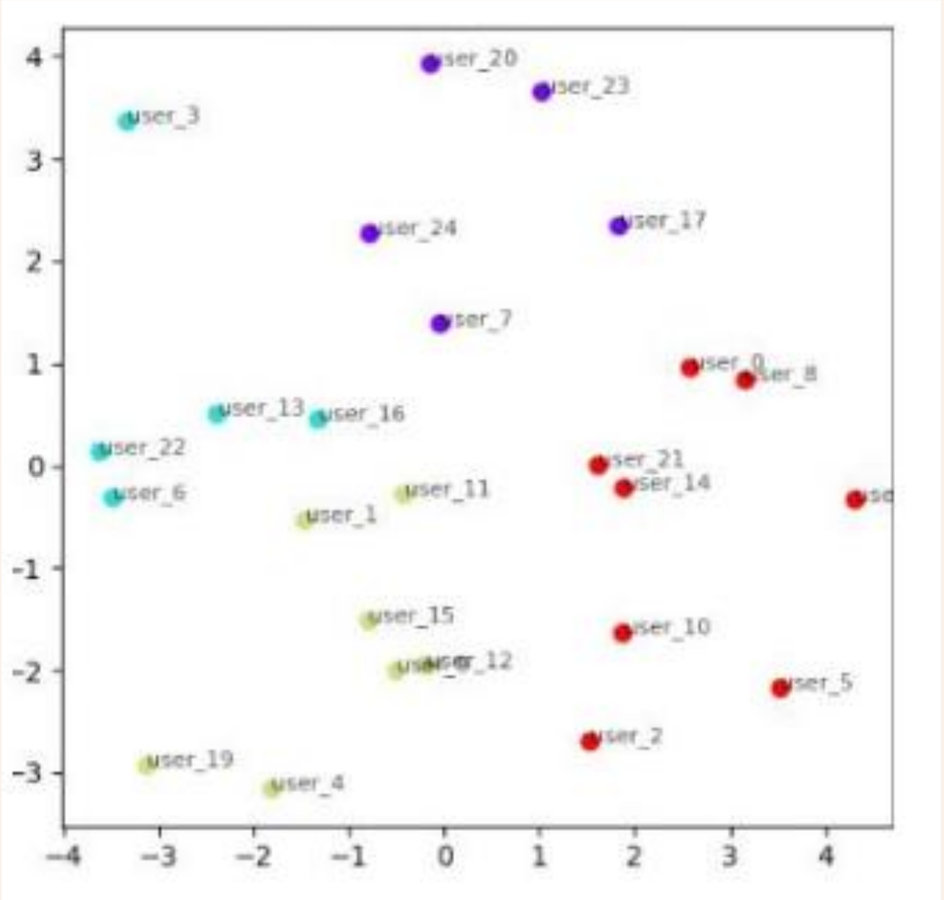
El algoritmo K-means es una técnica popular de agrupación (clustering) que se utiliza para dividir un conjunto de datos en k grupos distintos basándose en características similares (Grant RW, 2020).

CLASIFICACIÓN DE NODOS

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).



CLASIFICACIÓN DE NODOS



ANÁLISIS DE RESULTADOS

DEFINICIÓN DE PERFILES DE USUARIO

Agrupamiento	Caracteristica 1	Caracteristica 2	Caracteristica 3
Cluster 0	out_sadness: 0.75	out_surprise: 0.71	length_desc_userprofile: 0.66
Cluster 1	verified_account: 1.00	in_subjectivity: 0.87	out_anger: 0.82
Cluster 2	in_fear: 0.76	in_anticip: 0.72	in_anger: 0.72
Cluster 3	#URLstweet: 0.88	in_disgust: 0.87	#URLsretweet: 0.85

DEFINICIÓN DE PERFILES DE USUARIO

Influenciador Verificado:
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.

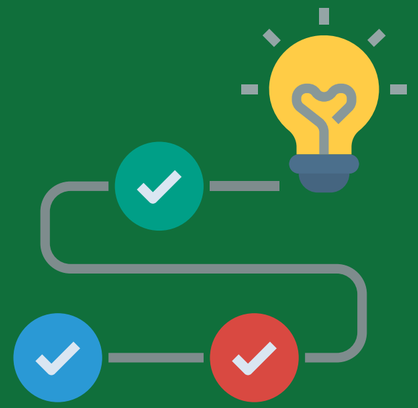
Optimista Verificado:
Este nombre destaca la autenticidad de la cuenta, indicada por la verificación, y enfatiza la predisposición de la persona a compartir contenidos que reflejan felicidad y optimismo en Twitter.

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.

DEFINICIÓN DE PERFILES DE USUARIO

Finalmente, en el proceso de generación de perfiles de usuario mediante el método de Kmeans 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.

CONCLUSIONES Y RECOMENDACIONES



CONCLUSIONES

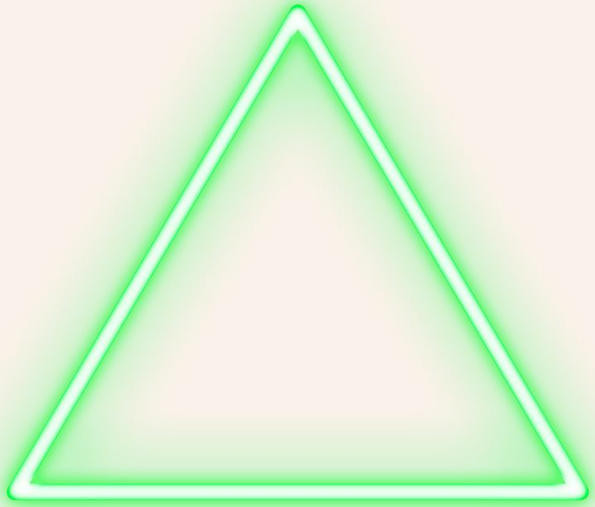
Importancia de la Detección de Información No Verificada

Rol Estratégico de los Nodos de Relacionamiento

Necesidad de Herramientas Alternativas para la Recolección de Datos

RECOMENDACIONES

Exploración de Herramientas Alternativas



Énfasis en la Seguridad de Datos

Desarrollo de Herramienta Personalizada

TRABAJOS FUTUROS

- **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.

- **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).

- **Desarrollo de Herramientas Avanzadas de Análisis:** Se podría explorar la creación de herramientas más avanzadas de análisis de nodos de relacionamiento en Twitter (X).

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.

BIBLIOGRAFÍA

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

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

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

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

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

16.PhantomBuster. (2024, 02 19). PhantomBuster. Retrieved from PhantomBuster: https://phantombuster.com/

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

18.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.

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

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

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

Nombre: David Diaz
Correo: david.diaz@softwareone .com
Tlfn: 0963648329
Quito, Ecuador

GRACIAS

“ESPE”