UACM

Universidad Autónoma
de la Ciudad de México

NADA HUMANO ME ES AJENO

COLEGIO DE CIENCIA Y TECNOLOGÍA

LICENCIATURA EN MODELACIÓN MATEMÁTICA

Construcción y Análisis de una red de retweets a partir de un tema de tendencia (Trending Topic).

T E S I S

QUE PARA OPTAR POR EL TÍTULO DE

LICENCIADO EN MODELACIÓN MATEMÁTICA

P R E S E N T A:

IVÁN RAIR PONCE AVILA

D I R E C T O R:
MTRO. HÉCTOR RUIZ SORIA

Ciudad de México, Noviembre 2024.

UACM
Universidad Autónoma de la Ciudad de México
Nada humano me es ajeno

CA Coordinación Académica

SISTEMA BIBLIOTECARIO DE INFORMACIÓN Y DOCUMENTACIÓN

UNIVERSIDAD AUTÓNOMA DE LA CIUDAD DE MÉXICO
COORDINACIÓN ACADÉMICA

RESTRICCIONES DE USO PARA LAS TESIS DIGITALES

DERECHOS RESERVADOS©

Agradecimientos

Con profunda gratitud y cariño, dedico el presente trabajo a quienes fueron parte esencial en este largo camino de licenciatura y en la realización de esta tesis. A mis padres, Maricela Ávila Mendoza y Juan Maximiliano Ponce Vázquez, cuyo amor y apoyo incondicional me sostuvieron en cada paso; a mi hermano, Alan Jafet Ponce Ávila, siempre presente, y a mi bebé que descansa en el cielo. Aunque no esté físicamente, su memoria me acompañó en las noches de desvelo y en los días de incertidumbre. Descansa en paz, pequeño; tu familia nunca te olvida, siempre estás en nuestros pensamientos y en nuestros corazones.
Este, mi núcleo familiar fueron mi refugio en los momentos más difíciles, cuando las dudas me hicieron considerar abandonar esta vocación y emprender otro camino. Su apoyo constante fue el recordatorio de propósito que hoy guía mi profesión y carrera como docente.

Agradezco profundamente a mi director, Héctor Ruiz Soria, cuya excelencia como docente y profesional ha sido para mí un ejemplo invaluable. Su dedicación a su labor y a cada proyecto inspira y me ha guiado en este trayecto. Gracias, profesor.

A mi alma máter, la Universidad Autónoma de la Ciudad de México, que me formó como el profesional que hoy ejerce con orgullo esta carrera que tanto esfuerzo me costó aprender.

Y, finalmente, a ti, Sofía Julieta. Con tu amor, comprensión y paciencia, me diste la fortaleza para perseverar en esta etapa tan ardua. A ti te debo tanto, aunque estas palabras no sean suficientes para expresar mi gratitud y amor hacia ti, cada página, cada esfuerzo llevan un pedazo de lo que juntos hemos recorrido. Gracias por acompañarme con tu apoyo constante y silencioso.

Contenido

Agradecimientos III

1 Introducción. 7
1.1 Preguntas de investigación. . . . 9

2 Twitter y base de datos. 11
2.1 Twitter y RStudio. . . . 11
2.1.1 Solicitud de acceso como desarrollador en Twitter. . . . 11
2.1.2 Recopilación de datos. . . . 17
2.1.3 Limpieza y preparación de datos. . . . 18
2.1.4 Preparación de datos y bases de datos. . . . 19
2.2 Variables de estudio. . . . 23
2.2.1 Análisis de datos en RStudio. . . . 26

3 Integración de datos y minería en análisis de bases de datos. 27
3.1 Introducción. . . . 27
3.1.1 Grafo: los puentes de Königsberg. . . . 28
3.1.2 Grafo en una red social. . . . 29
3.2 Aspectos matemáticos de un grafo en una red social. . . . 31
3.2.1 Modularidad. . . . 31
3.2.2 Tipos de comunidades 32
3.2.3 Louvain. . . . 33
3.2.4 Vector propio (eigenvector) y Centralidad. . . . 35
3.3 Comunidades en la red social Twitter. . . . 35
3.4 Twitter y R: Herramientas para el análisis de redes sociales. . . . 36
3.4.1 Extracción de datos en Twitter. . . . 37
3.5 R para el análisis de datos. . . . 38
3.5.1 Extracción de la información por el API. . . . 38

4 Construcción y análisis de la red de retweets. 41
4.1 Introducción. . . . 41
4.2 Construcción de las redes de retweets. . . . 41
4.3 Ranking. . . . 42
4.3.1 Red de retweets, base general. . . . 46
4.3.2 Red base de datos verificada. . . . 49
4.3.3 Ranking para la base de datos de Tweets verificados. . . . 51
4.3.4 Red base de datos no verificada. . . . 53
4.3.5 Red de retweets, base no verificado. . . . 56

5 Conclusiones. **59**

A Apéndices **63**

Apéndices **63**

A.1 Tabla con la descripción con las variables que genera cada tweet 63

A.2 Terminología de la minería y análisis de datos. 78

A.3 Terminología de Twitter. 79

A.4 Terminología de RStudio. 80

Bibliografía **81**

Índice de tablas

2.1 Variables de estudio y count 25
2.2 Variables con información extra del usuario que no son _count 26

4.1 Resumen estadístico de la variable retweet. 41
4.2 Tabla resumen de los 5 valores (counts). 44
4.3 Tabla con los valores generales (counts), las funciones de distribución acumulativa empírica (*ecdf*) y el puntaje para el clasificado número 1, así como para el usuario con el mayor número de seguidores. 46
4.4 Tabla con los valores (counts), las funciones de distribución acumulativa empírica (*ecdf*) y el puntaje para las cinco variables de los diez primeros usuarios en el ranking general. 47
4.5 Ranking y top_score que ocupan en la base general de los más influyentes por comunidad. 48
4.6 Tabla con los valores (counts) verificados, las funciones de distribución acumulativa empírica (*ecdf*) y el puntaje para el clasificado número 1 y el usuario con el mayor número de seguidores. 50
4.7 Tabla con los valores (counts), las funciones de distribución acumulativa empírica (*ecdf*) y el puntaje para las cinco variables de los diez primeros usuarios en el ranking de cuentas verificadas. 51
4.8 Ranking y top_score que ocupan los usuarios mas influyentes en la base verificados por comunidad y en la base general. 53
4.9 Ranking de la base no verificados de la función de distribución acumulativa empírica (*ecdf*) para los primeros lugares para cada una de las 5 variables. 56
4.10 Tabla con los valores (counts), las funciones de distribución acumulativa empírica (*ecdf*) y el puntaje para las cinco variables de los diez primeros usuarios en el ranking de la base de datos no verificados. 56
4.11 Ranking y top_score que ocupan los usuarios más influyentes por comunidad para la base de datos no verificados y en la base general. 58

5.1 Usuarios más influyentes para la base general, verificados, no verificados y sus respectivas comunidades 60

A.1 Descripción de las 91 variables generadas por cada tweet. 77
A.2 Terminología de la minería y Análisis de datos. 78

A.3 Terminología empleada en Twitter. 79

A.4 Terminología empleada en RStudio. 80

Índice de figuras

2.1 Liga del portal para solicitar cuenta de desarrollador. https://developer.twitter.com/en . . . 12
2.2 Obtención de la primera llave de acceso (PROJECT APP) . . . 12
2.3 Obtención de la segunda llave de acceso. . . . 13
2.4 Credenciales para la extracción de información, (Página de usuario) https://developer.twitter.com/en . . . 13
2.5 Estructura del token con las llaves de acceso ejecutadas desde RStudio. . . . 14
2.6 Ejecución del token en RStudio. . . . 16
2.7 Secuencia de comandos en R para extraer y almacenar en una base de datos los tweets del tema de tendencia de interés. . . . 16
2.8 Librerías empleadas en R para el análisis de la base de datos. . . . 17
2.9 Trozo de código para importar base general y convertirla en tabla con el comando tibble. . . . 19
2.10 Trozo de código para importar Base Verificados y convertirla en tabla con el comando tibble. . . . 20
2.11 Trozo de código para importar base no verificados y convertirla en tabla con el comando tibble. . . . 20
2.12 (Bases generadas.), [Formato.csv]. RStudio. . . . 21
2.13 Diagrama del proceso para la obtención de bases, rankings y comunidades. . . . 22

3.1 Representación de los puentes de Königsberg y su trazo con teoría de grafos. https://is.gd/ZRmrwb . . . 28
3.2 Calvo, J. (n.d.). Teoría de grafos para los puentes de Königsberg. Diagrama. Problema simplificado. . . . 29
3.3 Representación de grafo dirigido y no dirigido para la interacción unidereccional y bidireccional de la información. . . . 30
3.4 Representación centralizada y descentralizada. (2021, November 8). https://is.gd/Q21xRR . . . 32
3.5 Grafo aleatorio. Creación propia. https://graphonline . . . 33
3.6 Modularidad. Creación propia. Grafo. https://gephi.org/users/download/. . . . 34
3.7 Componentes de una red de grafos con centralidad. https://bit.ly/acortado . . . 35
3.8 Pastel del estudio de la red social Twitter. (Análisis de Redes Sociales, Mapeo de Conexiones y Redes Mediante Social Media Analytics - FasterCapital, s. f.) . . . 36

3.9 Interfaz y partes del Twitter. (2012, December 18). Interfaz. Un Enchufing de Social Media a la sangre. 37

3.10 (Pantalla principal de RStudio) 38

3.11 Logo de Twitter y R. `https://is.gd/uXLJeN` 39

4.1 Boxplot de retweet. 42

4.2 *ecdf* de una distribución normal estándar 43

4.3 Funciones de distribución acumulativa empírica (*ecdf*()) para las seis variables de los tweets de la base general. 45

4.4 Ranking vs top score para los 10 primeros clasificados en la base general. 47

4.5 Red de retweets base general (Algoritmo Louvain). 48

4.6 Funciones de distribución acumulativa empírica (*ecdf*()) para las cinco variables de los tweets verificados. 50

4.7 Diagrama de dispersión del ranking versus top score para los 10 primeros clasificados en la base de usuarios verificados. 52

4.8 Red de retweets base verificados (algoritmo Louvain) 53

4.9 Funciones *ecdf*() para las seis variables de los tweets base no verificados. 55

4.10 Diagrama de dispersión del ranking versus top score para los 10 primeros clasificados en la Base de usuarios no verificados. 57

4.11 Red de retweets base no verificados (algoritmo Louvain) 58

Objetivos

Los objetivos que pretendemos alcanzar son: uno en general y cuatro particulares.

Objetivo general

El objetivo general es realizar un análisis de una red social creada a partir de una serie de retweets a un tema de tendencia central (trending topic o hashtag), para identificar quiénes fueron los participantes que mayor influyeron (influencers) en la percepción pública del tópico. Además, se busca identificar las posibles comunidades en la discusión del tema.

Objetivos particulares

- Ranking. Construcción de una métrica (*ranking*) empleando la función de distribución acumulativa empírica ("*ecdf*()" , por sus siglas en inglés), involucrando algunas de las variables cuantitativas que se generan por cada tweet en el tema de tendencia.
- Generar la red de retweets para la base general, así como la identificación de comunidades y del participante con mayor influencia a través del ranking en la red y en las comunidades.
- Generar la red de retweets para la base de las cuentas **no verificadas**, así como la identificación de comunidades y del participante con mayor influencia a través del ranking en la red y en las comunidades.
- Generar la red de retweets para la base de cuentas **verificadas**, así como la identificación de comunidades y del participante con mayor influencia a través del ranking en la red y en las comunidades.

Capítulo 1

Introducción.

> Felder y Silverman (1988) afirman que los estudiantes aprenden de muchas maneras: viendo y escuchando, reflexionando y actuando, razonando lógica e intuitivamente, memorizando y visualizando, construyendo analogías y modelos matemáticos. También los métodos de enseñanza son variados. [...] Cuanto aprenda un estudiante en una clase dependerá de la habilidad innata y de su preparación previa, pero además de la compatibilidad entre su estilo de aprendizaje y el estilo de enseñanza de su instructor. Minería de datos para descubrir estilos de aprendizaje, s. f. [1]

Históricamente las matemáticas han jugado una importante labor en las disciplinas del modelado, motivada por la revolución tecnológica y la globalización económica. En la última década se ha presenciado una enorme demanda social por la creación y desarrollo de sistemas que ayudan en la toma de decisiones que pudieran extraer información y generar conocimiento a partir de las extensas bases de datos que se generan a cada instante. Así, por ejemplo, las ciencias como la Genómica, las Finanzas, la Medicina, la Informática y Telecomunicaciones, entre otras generan nuevos retos por el manejo y análisis de una gran cantidad de información, que para poder abordarlos ha sido necesario desarrollar nuevas herramientas tanto en los aspectos de modelado como en algoritmos. A raíz de estos nuevos retos es como toma auge la llamada minería de datos.

En el presente trabajo, como ya se mencionó, se realizó un análisis de una red social generada a partir de una serie de retweets de un tema de tendencia (trending topic o hashtag). En particular se emplea la información generada en la problemática que se desencadenó en septiembre del año 2019, en el cual se vío involucrado el Dr. Antonio Lazcano[1] el cual fue removido de la Comisión

[1] Científico especializado en Biología Evolutiva y divulgador de la ciencia. Investigador nacional nivel III por parte del CONACYT y dirige en forma honoraria el Centro Lynn Margulis de Biología Evolutiva en las Islas Galápagos,

Dictaminadora del Área 2, de Biología y Química del Sistema Nacional de Investigadores. Lo que generó un debate público que se hizo viral en varias redes sociales, en particular en Twitter, que al paso de los días se generaron opiniones divergentes tanto en redes sociales como en la comunidad científica nacional e internacional. La situación provocó que analistas políticos periodistas y la comunidad científica tuvieran un gran intercambio de publicaciones a través de Twitter (tweets), a tal grado que hashtag $\#CONACYT$ llegó a ser trending topic (tema de tendencia)[2].

El análisis que se realizará forma parte de la llamada minería de datos, en particular se analizará la red social que se genera a partir de la base de datos extraída de Twitter del tema de tendencia, usando la teoría de grafos y empleando en lenguaje de programación RStudio.

El trabajo se encuentra organizado en 5 capítulos más anexos y referencias, cuyo contenido se describe brevemente a continuación.

En el capítulo 1, se describe el tema, el problema y la metodología que se empleará para explicar el mismo, así como el estado del arte.

Uno de los enfoques más destacados en el análisis de datos de Twitter la detección y análisis de comunidades. Este enfoque busca identificar grupos de usuarios que interactúan entre sí con mayor frecuencia que con el resto de la red. La identificación de comunidades proporciona información valiosa sobre la estructura subyacente de la red social, la difusión de información y la formación de opiniones.

Investigaciones recientes, como el estudio de Smith et al. (2021)[59], quienes aplicaron técnicas de detección de comunidades en redes de Twitter para identificar grupos de usuarios con intereses similares, opiniones compartidas o comportamientos colectivos. Estos estudios han demostrado la utilidad del análisis de comunidades para comprender la dinámica de las conversaciones en Twitter y para identificar comunidades influyentes y temas emergentes.

Además del análisis de comunidades, el análisis de sentimientos y la detección de temas comunes son áreas complementarias que han ganado atención en la investigación de Twitter empleando R. Estos enfoques permiten comprender la actitud y emociones expresadas en los tweets, así como identificar los temas dominantes en las conversaciones en curso.

También, este análisis tiene un fin y es uno de los más populares (Shmueli et al., 2017)[58], incluye librerías para manejos estadísticos y de aprendizaje automático para la predicción, clasificación, visualización, reducción de dimensiones, sistemas de recomendación, clustering, minería de texto y análisis de redes. Sotelo, F. (2023) [57].

A pesar de los avances significativos en el análisis de datos de Twitter con R, persisten desafíos importantes, como la identificación de sesgos y la interpretación de los resultados en el contexto de la red social en constante evolución.

En el capítulo 2 se describe la teoría de grafos y cómo es empleada como una herramienta de

Ecuador.

[2]Los temas de tendencia indican de que tema o tópico se está hablando masivamente en Twitter en ese momento.

análisis para una red social, y cómo puede ayudar a la construcción y manejo de los componentes que conforman una base de datos. Se explica cómo dicha teoría puede y tiene la facilidad de crear, describir redes complejas y su comportamiento además, arrojar información valiosa como lo son las comunidades y los participantes más influyentes de cada comunidad.

En el capítulo 3 se presenta y describen las dos herramientas para el análisis, RStudio y Twitter, y cómo se pueden emplear para trabajar en conjunto. Para la solicitud de información en Twitter respecto a un tema en tendencia, primero se solicitan los permisos (las API.[3]) Una vez obtenidos los permisos se hace una búsqueda de los tweets que contengan los temas de interés. En nuestro caso una vez limpiada la base de datos se obtuvo una base de datos que consta de alrededor de 18,000 tweets con 90 variables recolectados en los dos primeros días de un periodo que fue tema de tendencia (21 al 22 de septiembre de 2019), cuando se alcanzó el trending topic con el hashtag #CONACYT[4]. Cabe resaltar que esta cantidad de información no es comparable a una base de datos empresarial o gubernamental.

En el capítulo 4 se realiza el análisis y la formación de la red de retweets así como el ranking general, por comunidad para usuarios verificados, no verificados y se identificara de acuerdo a la métrica definida (ranking) de los usuarios más influyentes. Posteriormente en el capítulo 5 se recopilan las conclusiones y algunas observaciones en general para futuros trabajos alrededor del mismo tema. Finalmente se agregan algunos apéndices y la bibliografía empleada en el desarrollo de la investigación.

1.1. Preguntas de investigación.

- Identificación de los usuarios más influyentes (mejores rankeados por la métrica empleada).
- Identificar las posibles comunidades que se generan en las distintas redes, base de datos General, Verificada y No Verificados.
- Medición del "poder o índice de penetración en la red social (red de retweets)"

[3]Es un conjunto de reglas, protocolos y herramientas que permiten a diferentes aplicaciones comunicarse entre sí.

[4]Oficialmente el nombre CONACYT cambio a CONAHCYT (Consejo Nacional de Humanidades, Ciencia y Tecnología)

Capítulo 2

Twitter y base de datos.

> Las bases de datos utilizan diferentes motores de búsqueda diseñados por las empresas proveedoras del acceso a las mismas. Un motor de búsqueda es una plataforma que permite recuperar archivos almacenados en un servidor de Internet. Para buscar a través de ellos generalmente sólo se requieren palabras claves, las cuales son cotejadas con las bases de datos de cada servidor. Posteriormente, los resultados son presentados en orden de relevancia o fecha, según las características específicas de cada uno. [2] (Castrillón et al., 2008. p.101)

La base de datos[1] que emplearemos en la aplicación se extrajo de twitter para el tema de tendencia elegido.

2.1. Twitter y RStudio.

Twitter desempeñó un papel fundamental en el trabajo de tesis, al proporcionar una plataforma rica en datos y conversaciones en tiempo real relacionadas con temas de tendencia. A continuación, se destacan algunas formas en que Twitter contribuyó a la investigación:

2.1.1. Solicitud de acceso como desarrollador en Twitter.

Una vez seleccionado el tema de tendencia o hashtag se solicita en Twitter acceso al API de Twitter. A continuación se describe de manera general los pasos llevados a cabo para solicitar

[1] La base de datos utilizada en la aplicación se refiere a un sistema de almacenamiento y gestión de datos diseñado para recolectar y analizar información extraída de Twitter. Esta base de datos se encarga de almacenar los tweets relacionados con el tema de tendencia elegido, permitiendo su posterior procesamiento y análisis para identificar patrones y tendencias.

acceso a datos de Twitter como cuenta de desarrollador *(tema de tendencia)*.

- Paso 1: Para la solicitud de información en Twitter, se crea una cuenta para generar las credenciales de acceso como desarrollador (2.1), a continuación se mostrara los pasos para generar las credenciales de acceso.

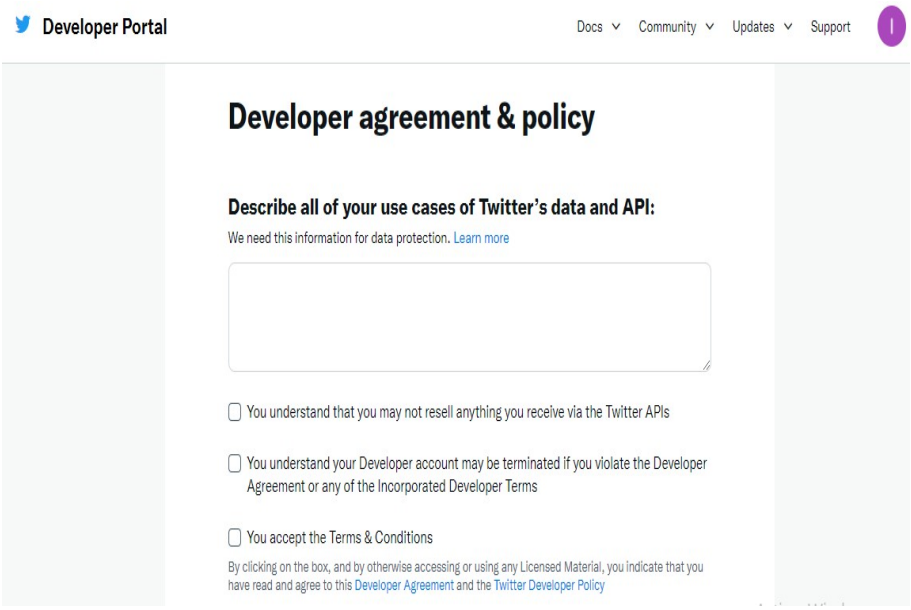


Figura 2.1: Liga del portal para solicitar cuenta de desarrollador. https://developer.twitter.com/en

Las "credenciales de acceso" en Twitter se refieren a la información de autenticación que un usuario o una aplicación necesita proporcionar para acceder a una cuenta de Twitter o para utilizar la API de Twitter de manera segura. Estas credenciales son esenciales para verificar la identidad y la autorización del usuario o la aplicación en la plataforma.

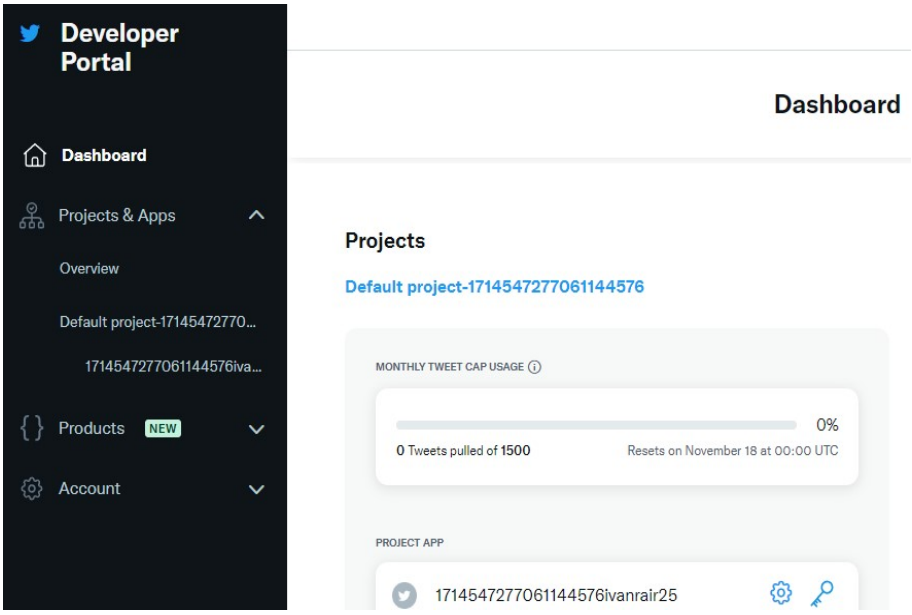


Figura 2.2: Obtención de la primera llave de acceso (PROJECT APP)

- Paso 2: Una vez que se tenga una cuenta de desarrollador, se crea una una aplicación en el panel de desarrollador de Twitter. Esto proporcionará las credenciales (Claves de acceso, figuras 2.2 y 2.3) necesarias para conectarse a la API de Twitter.

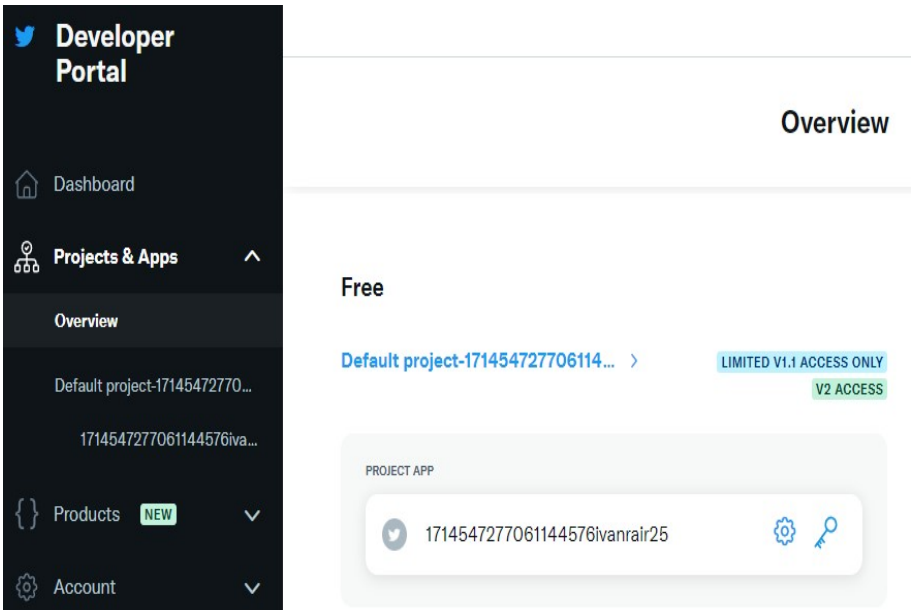


Figura 2.3: Obtención de la segunda llave de acceso.

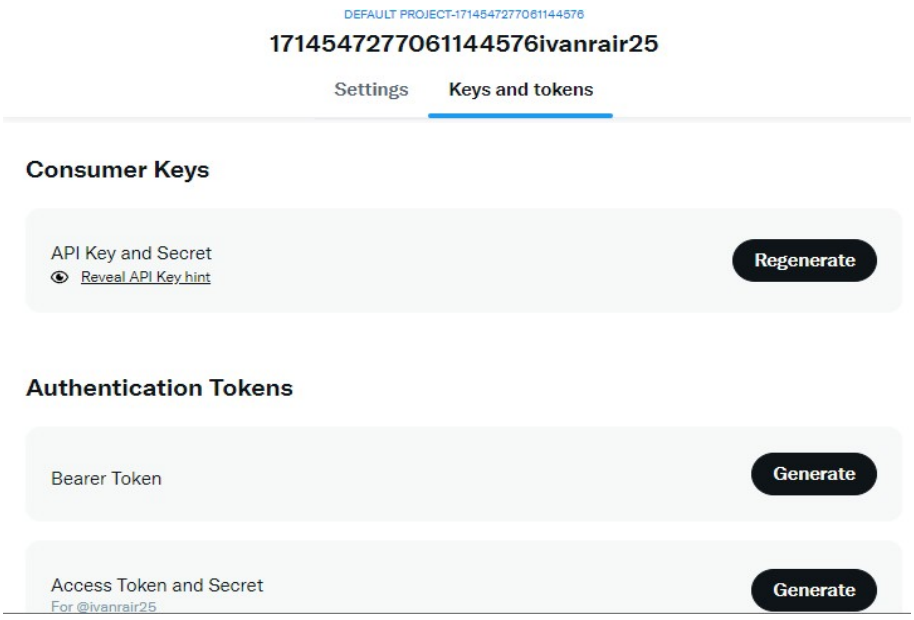


Figura 2.4: Credenciales para la extracción de información, (Página de usuario) https://developer.twitter.com/en

- Paso 3: Acceso al API. Iniciando una sesión en R y cargando la librería rtweet generamos un token [2] para insertar los permisos generados en el paso 2 (ver figura 2.5).

[2]Un "token" en el contexto de Twitter y R hace referencia a un conjunto de claves de acceso que se utilizan para autenticar y autorizar una aplicación o script de R para interactuar con la API de Twitter. La API de Twitter permite

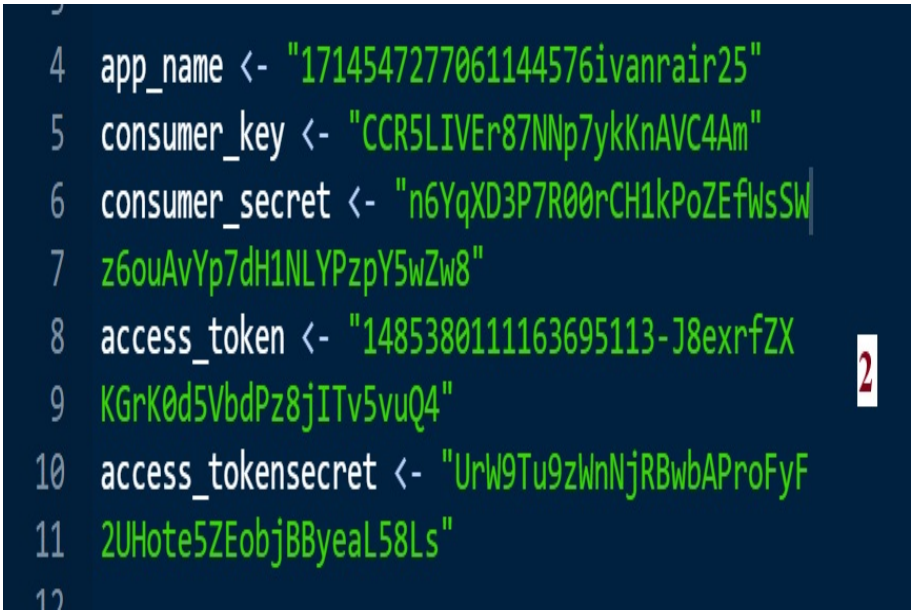


Figura 2.5: Estructura del token con las llaves de acceso ejecutadas desde RStudio.

- Llaves de acceso.

  - app_name:

    Esta clave generalmente se utiliza para proporcionar un nombre descriptivo para la aplicación o proyecto que se está conectando a la API de Twitter. Sirve para identificar la aplicación en Twitter y puede ser útil para gestionar la autenticación de la API.

  - consumer_key:

    Este es un identificador único asociado a la aplicación en la API de Twitter. Se debe proporcionar esta clave cuando registras tu aplicación en el portal de desarrolladores de Twitter, y se emplea para autentificar la aplicación y garantizar que se está autorizado para acceder a la API.

  - consumer_secret:

    Esta es una clave secreta que se utiliza junto con consumer_key para la autentificación de la aplicación, y debe mantenerse en secreto y no compartirse públicamente. En otras palabras, el "consumer_key" y el "consumer_secret" permiten que Twitter verifique la autenticidad de tu aplicación (es el token generado anteriormente).

  - access_token:

a los desarrolladores acceder a los datos y funcionalidades de Twitter, como la obtención de tweets, publicación de tweets, análisis de tendencias y más.

Este es un identificador único que se genera una vez que la aplicación ha sido autorizada por un usuario de Twitter para acceder a sus datos, es decir, es una cuenta de usuario y una aplicación. También se utiliza el "access_token" para autentificar la aplicación y para que se pueda acceder a la cuenta de Twitter del usuario.

- access_token_secret:

  Al igual que la llave *consumer_secret* es una llave de acceso que se utiliza en conjunto con el *token de acceso* para autentificar la aplicación y el usuario al acceder a los datos de Twitter. También debe mantenerse en secreto y no compartirse públicamente. En

  resumen, estas llaves de acceso permiten la autentificación del desarrollador y de un token generado que permiten la búsqueda de información en Twitter.

- Paso 4: Extracción de la información. A través de R ejecutamos el token generado en la

  (figura 2.6) con las llaves de acceso.

  Los parámetros de la función *create_token* son:

  - token <- create_token(...): Esta línea de código mediante la función *create_token*() genera un objeto llamado *token*, y le asigna al resultado de la función create_token(). La función se utiliza para generar un token de acceso que se utilizará para la autentificación en la API de Twitter.

  - app = app_name: Esta parte del código establece el nombre de la aplicación en el token de acceso.

  - consumer_key = consumer_Key: Este parámetro se utiliza para proporcionar la clave del consumidor (consumer key) que identifica la aplicación en la API de Twitter.

  - consumer_secret = consumer_Secret: Esta llave se utiliza para la autentificación del desarrollador.

  - access_token = access_Token: Una vez que la aplicación ha sido autorizada para acceder a la base de datos, esta llave permite vincular la cuenta y la aplicación en R.

  - access_tokensecret = access_TokenSecret: Al igual que en el paso anterior, es otra llave de acceso de seguridad para la autentificación.

  En resumen, la conexión a Twitter a través del API, se realiza mediante la creación de una función llamada Token entre las llaves de acceso para conectar R con Twitter.

  Se utilizó para recopilar los tweets que coincidan con tus criterios de búsqueda, almacenamos estos tweets en un archivo o en una estructura de datos adecuada para su análisis posterior.

Figura 2.6: Ejecución del token en RStudio.

- Paso 5: Extracción de la información. Una vez conectados Twitter y R a través del token, se extrae la información mediante el comando *search_twets* especificando el hashtag del tema de tendencia y la almacena en un objeto (base de datos) en R, posteriormente se puede importar en cualquier otro formato compatible con R, en particular se usa el formato csv (figura 2.7.).

Figura 2.7: Secuencia de comandos en R para extraer y almacenar en una base de datos los tweets del tema de tendencia de interés.

Paso 6: Análisis de Datos con RStudio: Importar los datos limpios en RStudio y utilizar las capacidades de análisis de R para explorar y entender los patrones de comportamiento dentro de los tweets que contienen el tema de tendencia de nuestro interes.

Paso 7: Visualización de Datos y Conclusiones: Cree representaciones gráficas de los resultados empleando las herramientas conocidas y disponibles en RStudio para interpretar los resultados y llegar a conclusiones.

2.1.2. Recopilación de datos.

Una vez obtenidos los permisos, se usa la API de Twitter y se recopilan datos. Cabe resaltar que se puede utilizar una variedad de lenguajes de programación y recopilar datos relacionados con el tema de interés. En este caso se utiliza R, y la libreria *rtweet* (Figura 2.8) para interactuar con la API.

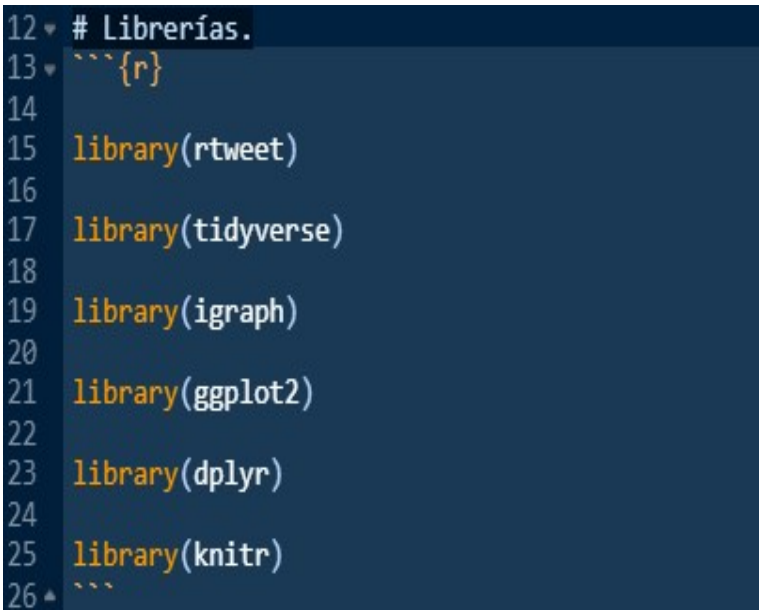


Figura 2.8: Librerías empleadas en R para el análisis de la base de datos.

Para el análisis y manipulación de la base obtenida con la API de Twitter se puede utilizar una variedad de lenguajes de programación y recopilar datos relacionados con el tema de tendencia. Como ya se mencionó en el presente trabajo empleamos en R las librerías que se menciona a continuación.

- #**rtweet** es una biblioteca específica para trabajar con Twitter en R. Proporciona funciones y herramientas para acceder a la API de Twitter, recopilar datos de tweets, usuarios, y realizar diversas operaciones relacionadas con Twitter, como buscar tweets, extraer datos de usuarios y más. Es una biblioteca fundamental para la recopilación y análisis de datos de Twitter en R.

- #**tidyverse** es una colección de paquetes de R que se utilizan comúnmente para manipulación y visualización de datos. Incluye bibliotecas como *dplyr*, *ggplot2*, y otras que proporcionan un conjunto coherente de herramientas para la limpieza, transformación y visualización de datos.

- **#igraph** es una biblioteca utilizada para el análisis de redes y grafos. En el estudio, se puede utilizar para crear y analizar redes de interacciones entre usuarios de Twitter, lo que es útil para detectar comunidades, calcular métricas de centralidad y visualizar la estructura de la red.

- **#ggplot2** es una biblioteca para crear gráficos y visualizaciones de datos. Se permite generar gráficos de alta calidad y personalizables para representar los hallazgos de una manera efectiva y comprensible. Se utiliza para mostrar gráficos relacionados con el análisis de datos de Twitter.

- **#dplyr** es una parte clave de tidyverse y proporciona una serie de funciones para realizar operaciones de manipulación y transformación de datos. Se usa para filtrar, resumir, agrupar y modificar los datos de Twitter de manera eficiente.

- **#knitr** es una biblioteca que facilita la generación de informes, presentaciones y documentos reproducibles en R. Se puede utilizar para combinar texto, código R y resultados de análisis.

2.1.3. Limpieza y preparación de datos.

Se realiza un análisis exploratorio para limpiar la base de datos, (por ejemplo, si existen ausencia de datos, identificar observaciones atípicas, caracterizar o clasificar las variables, entre otras). En el análisis de la base de datos es muy común encontrar:

- Eliminación de duplicados: Se realiza una prueba para detectar datos duplicados y eliminarlos, lo cual mejora el análisis.

- Corrección de errores tipográficos: Se revisan los tweets en búsqueda de posibles errores tipográficos, especialmente en los hashtags, menciones de usuarios y enlaces.

- Manejo de caracteres especiales: Los tweets pueden contener caracteres, emojis y símbolos que requieren un manejo adecuado. Se puede optar por eliminarlos o reemplazarlos según el enfoque de análisis.

- Tratamiento de links: Algunos tweets pueden contener enlaces. Para este estudio se decidió no extraer información de esos enlaces o simplemente eliminarlos.

- Normalización de texto: Se normalizo el texto en los tweets para garantizar que el análisis sea más coherente. Esto incluye la conversión de todo el texto a minúsculas, la eliminación de signos de puntuación y separación del token para el texto en palabras individuales, con el fin de evitar algunos símbolos raros los puede mal interpretar R al momento del análisis.

- Codificación de categorías: Se analizaron categorías, como el sentimiento de los tweets, codificamos categorías numéricas. Por ejemplo, se asignan valores numéricos a los sentimientos (positivo, negativo, neutral).

- Manejo de fechas y horas: Se utiliza el formato de fecha y hora para las variables cualitativas, ya que algunas de estas conversaciones se realizaron en diferentes países y esto implica diferentes zonas horarias.

- Eliminación de información irrelevante: Si hay atributos o columnas que no son relevantes para los objetivos de la investigación, se optó por eliminarlos para simplificar al conjunto de datos y no tomar en cuenta esos datos.

- Gestión de valores faltantes: R tiene la facilidad de analizar variables con valores faltantes en el análisis. También se optó por eliminar filas con valores faltantes o valores basados en ciertas estrategias, dependiendo del contexto.

2.1.4. Preparación de datos y bases de datos.

Una vez hecha la limpieza de la base de datos para extraer la información relevante se analizaron los atributos específicos de los tweets, por ejemplo: la fecha, el contenido, las menciones de usuarios, los retweets, los "me gusta", las imágenes, entre otras.

En la figura 2.13 se muestra la estructura de cómo se analizó la información para las tres bases de datos empleadas para identificar comunidades y usuarios más influyentes.

- Base general. En la figura 2.9 se muestra el trozo de código mediante el cual se filtró la base llamándola *General*. Esta base es el resultado de los tweets obtenidos y sujetos al proceso de limpieza mencionados anteriormente, y consta de tweets verificados y no verificados.

```
# Importar base de datos y convertida a tibble
```{r}
base <- read.csv("tuits-conacyt.csv")
# head(tuits)
base <- as_tibble(base)
#Tabla de frecuencia de screen_name
TT <- transform(table(base$screen_name))

```

Analisis exploratorio de las 5 variables
```{r}
variables<-data.frame(base$followers_count,
                      base$friends_count,
                      base$favourites_count,
                      base$statuses_count,
                      base$listed_count,
                      base$retweet_count)
summary(variables)

```
```

Figura 2.9: Trozo de código para importar base general y convertirla en tabla con el comando tibble.

- Base verificados. En la figura 2.10 se muestra el trozo de código para filtrar los tweets *verificados*. Los tweets verificados son aquellos que han sido autentificados oficialmente con legitimidad e identificados a entidades o individuos (Usuarios en general, figuras públicas, deportistas, políticos, organizaciones públicas y privadas, entre otras.)

```
# Importar base de datos y convertida a tibble

```{r}
baseV <- read.csv("verificados.csv")
# head(tuits)
baseV <- as_tibble(baseV)
```


Filtrado de usuarios
```{r}
UsuariosV <- baseV%>%select(screen_name,
                            verified,
                            followers_count,
                            friends_count,
                            favourites_count,
                            statuses_count,
                            listed_count,
                            retweet_count,
                            mentions_screen_name,
                            comunidad) %>%
  unique()
```

Figura 2.10: Trozo de código para importar Base Verificados y convertirla en tabla con el comando tibble.

- Base no verificados. En la figura 2.11 se muestra el trozo de código empleado para filtrar y obtener los tweets *no verificados*. Los tweets no verificados son usuarios que no están verificados por la plataforma, es decir, se refiere a cuentas de usuarios que no han sido sometidas a proceso de verificación.

```
# Importar base de datos y convertida a tibble

```{r}
baseNV <- read.csv("Nverificados.csv")
# head(tuits)
baseNV <- as_tibble(baseNV)
```

Filtrado de usuarios
```{r}
UsuariosNV <- baseNV%>%select(screen_name,
                              followers_count,
                              friends_count,
                              favourites_count,
                              statuses_count,
                              listed_count,
                              retweet_count,
                              mentions_screen_name) %>%
  unique()
```

Figura 2.11: Trozo de código para importar base no verificados y convertirla en tabla con el comando tibble.

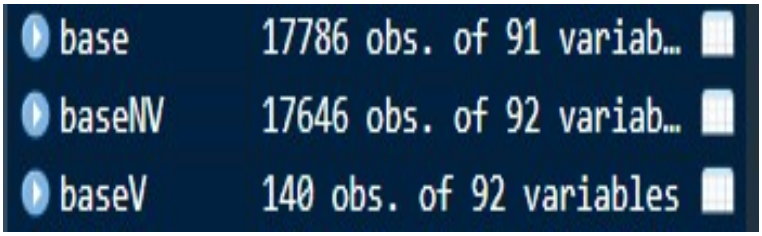


Figura 2.12: (Bases generadas.), [Formato.csv]. RStudio.

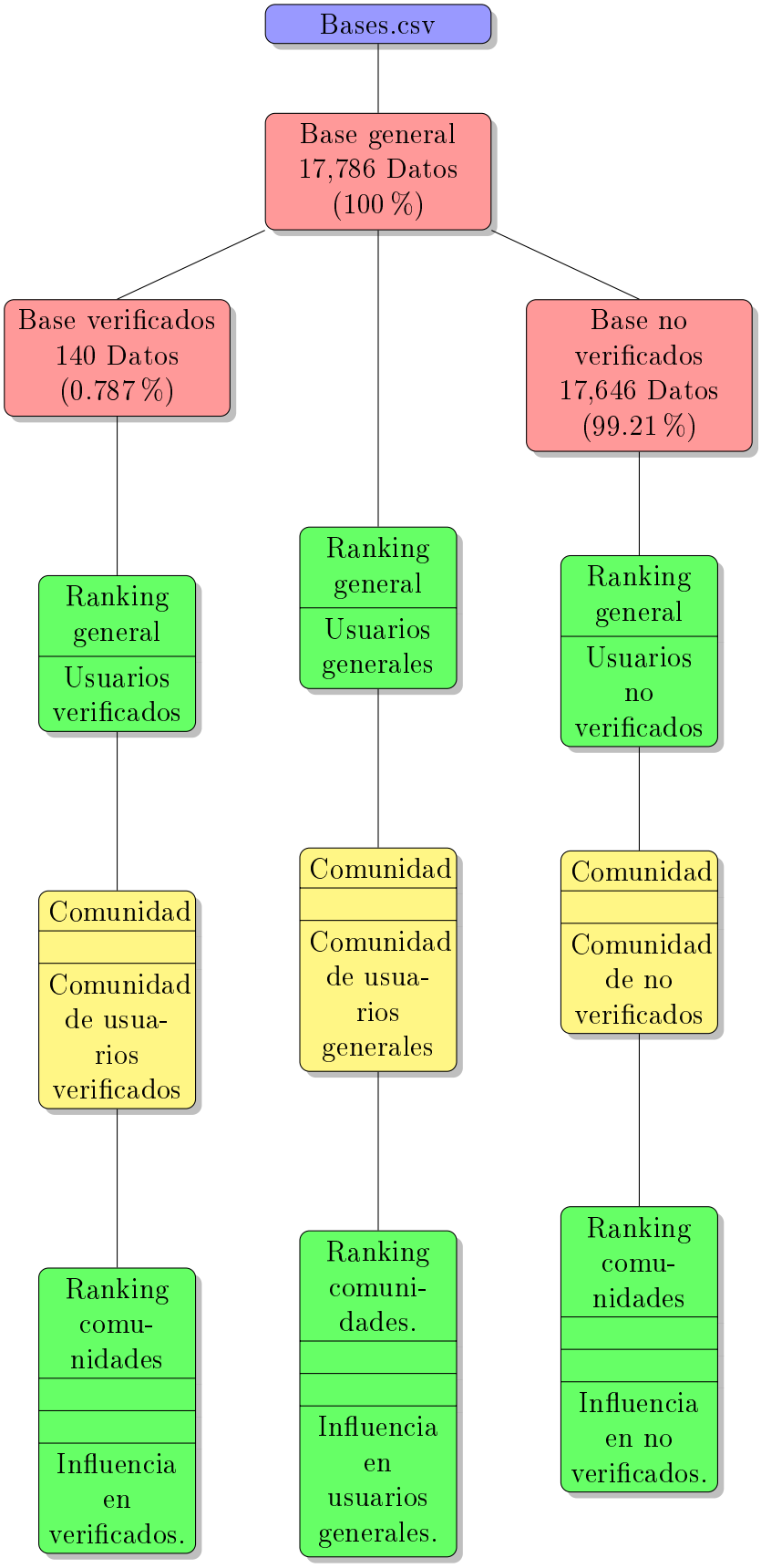


Figura 2.13: Diagrama del proceso para la obtención de bases, rankings y comunidades.

2.2. Variables de estudio.

Como ya se mencionó cada tweet contiene 90 variables (ver Anexo A), abarcando diversas características de los usuarios y datos relacionados entre ellos por ejemplo *seguidores*, *retweets*, *listas de cuentas*, etc.

Las variables de estudio desempeñan un papel fundamental en cualquier investigación, permitiendo un análisis profundo de fenómenos o la respuesta a preguntas específicas. En esta tesis sobre minería y análisis de datos de Twitter en relación con el problema de CONACYT, se han identificado 90 variables de estudio, abarcando diversas características y datos relacionados con tweets y usuarios involucrados en el tema. Presentar primero las variables más relevantes facilita la asimilación de información clave, mientras que colocar las variables adicionales en un anexo mejora la estructura general de las 82 variables restantes en esta sección específica se basa en la búsqueda de claridad, promoviendo la accesibilidad y eficiencia en la recuperación de detalles específicos para lectores interesados en investigaciones futuras. Este enfoque garantiza una presentación organizada y efectiva de la información.

Selección de variables.

Empleando la función de distribución acumulativa empírica, se utilizaron cinco variables cuantitativas de cada tweet en la base de datos: favourites_count, friends_count, followers_count, listed_count y retweet_count. Esta última fue clave para generar la red de comunidades de retweets. Además, se incluyeron otras tres variables para obtener más información sobre los usuarios y menciones: verified, screen_name y mentions_screen_name. Estas variables cuantitativas, (identificadas con el sufijo *x_ count*), fueron seleccionadas para generar una métrica que permitiera identificar comunidades y los usuarios más influyentes, como se detalla en la tabla adjunta 2.1.

No.	Utilidad	Nombre de la variable	Descripción
1	favourites_count:	Contador de favoritos	El número de tweets marcados como favoritos por un usuario refleja sus preferencias y podría indicar los temas que le interesan. Esto puede ser relevante para comprender los intereses de los usuarios involucrados en el debate de CONACYT.
2	followers_count:	Contador de seguidores	La cantidad de seguidores de una cuenta es un indicador de su alcance en la plataforma. Cuantos más seguidores tenga una cuenta, mayor será su influencia. Esta variable es relevante para identificar cuentas influyentes en el contexto de CONACYT.

3	friends_count	Contador de amigos	El número de cuentas que sigue un usuario puede dar una idea de sus intereses y conexiones en Twitter. El análisis de esta variable podría ayudar a identificar patrones de seguimiento y relaciones entre cuentas relacionadas con CONACYT.
4	listed_count	Listas	El número de listas en las que se incluye una cuenta puede proporcionar información sobre su relevancia en comunidades específicas. Esto puede ser útil para identificar cuentas que son consideradas importantes en el contexto de CONACYT.
5	statuses_count	Cuenta de estados	El número de tweets publicados por una cuenta puede indicar su nivel de actividad en la plataforma. Cuantos más tweets haya publicado una cuenta, mayor será su participación. Esta variable es relevante para identificar cuentas activas en el contexto de CONACYT.
6	retweet_count	Recuentos de retweets	La variable retweet_count se refiere al número de veces que un tweet específico ha sido retuiteado en la plataforma de Twitter. Un retweet es cuando un usuario comparte un tweet de otra persona en su propia línea de tiempo, lo que permite que el contenido llegue a un público más amplio.
7	mentions_screen_name	Menciona el nombre de usuario	Esta variable indica a qué cuentas se menciona en los tweets. El análisis de las menciones puede ayudar a identificar las interacciones y las relaciones entre usuarios en las conversaciones relacionadas con CONACYT.

8	screen_name	Nombre de usuario	El nombre de pantalla de un usuario es su identificador en Twitter. El análisis de los nombres de pantalla puede ayudar a identificar cuentas específicas que son relevantes en las conversaciones relacionadas con CONACYT. Esto podría ser útil para seguir la actividad de usuarios clave.
9	verified	Verificados	Esta variable indica si una cuenta está verificada por Twitter como auténtica. Las cuentas verificadas suelen ser de figuras públicas, celebridades o instituciones. El análisis de esta variable podría ayudar a identificar cuentas de alto perfil que están involucradas en el debate sobre CONACYT.

Tabla 2.1: Variables de estudio y count

El análisis principal, como se comentó se centró en las variables de conteo (_count) debido a su capacidad para cuantificar diversas interacciones en el tema de conversación. El resto de las variables que contiene cada tweet son variables de tipo cualitativo, que no se emplean para el propósito del trabajo. Por ejemplo en la tabla 2.2 de abajo se muestran dos de ellas.

No.	Utilidad	Nombre de la variable	Descripción
8	Reply_to_status_id.	Respuesta al id de estado	Si el tuit es una respuesta, este campo contendrá el ID del tuit original, permitiendo representar una cadena de respuestas.

63	place_name.	Nombre del lugar.	El dato place_name en Twitter almacena el nombre del lugar geográfico mencionado en un tuit. Este valor proporciona información sobre la ubicación asociada al tuit, como una ciudad, país, región u otra descripción geográfica. Es útil para entender el contexto geográfico de los tuits y realizar análisis basados en la ubicación, identificando patrones de contenido y tendencias relacionadas con lugares específicos. Es importante tener en cuenta que no todos los tuits incluyen información de lugar, por lo que la disponibilidad de datos en la columna place_name puede variar.

Tabla 2.2: Variables con información extra del usuario que no son _count

2.2.1. Análisis de datos en RStudio.

El ranking de usuarios más influyentes se realizó empleando la métrica [62]. Las distribuciones de probabilidad están relacionadas con las distribuciones de frecuencias. Una distribución de frecuencias teórica es una distribución de probabilidades que describe la forma en que se espera que varíen los resultados. Carmona, O. (2004). La cual se basa en la función de distribución acumulativa empírica (*ecdf*) entre las variables seleccionadas. Esto es, a cada una de las variables (columnas de la base), se calcula su función ***ecdf()*** asociada que se utilizará para calcular los percentiles de cada usuario para las variables seleccionadas que generaran el ranking para clasificarlos.

Para la creación del ranking tomamos las variables.

- $favourites\_count$: Número de usuarios que han marcado tus tweets como favoritos.
- $friends\_count$: Número de cuentas que un usuario sigue.
- $followers\_count$: Número de cuentas que siguen al usuario.
- $statuses\_count$: Número total de tweets publicados por un usuario.
- $listed\_count$: Número de listas de distribución en las que el usuario está incluido.
- $retweet\_count$: Número o medida directa de la popularidad y la difusión de un tweet en la plataforma. Cuantos más retweets tenga un tweet, más personas lo están compartiendo con sus seguidores, amplificando así su alcance y visibilidad dentro de la red social.

Capítulo 3

Integración de datos y minería en análisis de bases de datos.

> Las bases de datos utilizan diferentes motores de búsqueda diseñados por las empresas proveedoras del acceso a las mismas. Un motor de búsqueda es una plataforma que permite recuperar archivos almacenados en un servidor de Internet. Para buscar a través de ellos generalmente sólo se requieren palabras claves, las cuales son cotejadas con las bases de datos de cada servidor. Posteriormente, los resultados son presentados en orden de relevancia o fecha, según las características específicas de cada uno. [2] (Castrillón et al., 2008. p.101)

3.1. Introducción.

En este capítulo se presenta la teoría de grafos[1] como una herramienta clave para construir y analizar una red, en este caso de retweets. Se explica cómo se utiliza la teoría de grafos con herramientas como RStudio y Excel para comprender las interacciones en Twitter, incluyendo la metodología de identificación de temas de tendencia (trending topic[2]). Se realiza un análisis de minería de datos de la información recabada, en este caso se tiene una base de datos de 18,000 filas por 90 columnas, con representaciones numéricas y gráficas para enriquecer la comprensión

[1] La teoría de grafos es una rama de las matemáticas que estudia las relaciones entre objetos, representadas como nodos o vértices, conectados por enlaces llamados aristas. En esencia, la teoría de grafos proporciona un marco conceptual y herramientas matemáticas para comprender y resolver problemas que involucran relaciones entre entidades.

[2] hashtags que están siendo discutidos y compartidos en gran medida en un momento específico. Estos temas suelen aparecer en una lista de tendencias en la plataforma y pueden variar en función de la ubicación geográfica y las preferencias del usuario.

del tema de tendencia en la red, es decir se establecen las bases para comprender como la teoría de grafos y análisis de redes pueden ayudar a comprender eventos mediáticos como los temas de tendencia que se analizan.

3.1.1. Grafo: los puentes de Königsberg.

En el siglo XVIII, la ciudad de Königsberg se enfrentó a un enigma matemático intrigante: ¿se podía cruzar cada uno de sus siete puentes exactamente una vez?, ver figura 3.1. La solución a este problema se abordado por el matemático Leonhard Euler utilizando lo que ahora se conoce como la teoría de grafos. Este problema se convirtió en un hito en la historia de esta rama de las matemáticas, y también sentó las bases para futuros desarrollos en diversos campos.[43] S-a. (2004, febrero). En otras palabras, Euler representó cada isla como un nodo y cada puente como una arista en un grafo, y al analizarlo identificó cuatro nodos impares, lo que reveló una verdad matemática fundamental:

> **En un grafo con más de dos nodos impares, no se puede encontrar un camino que cruce cada arista exactamente una vez.**

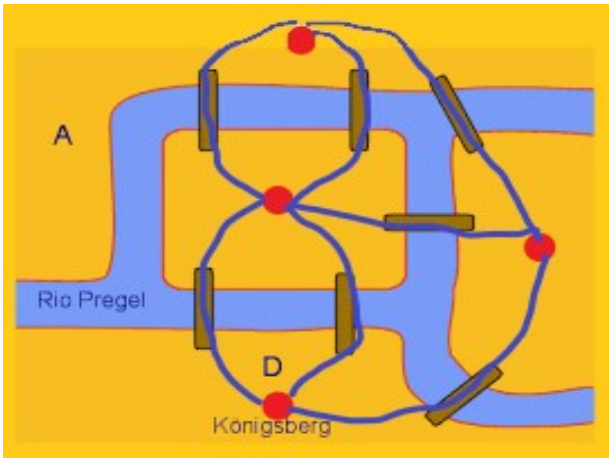


Figura 3.1: Representación de los puentes de Königsberg y su trazo con teoría de grafos. https://is.gd/ZRmrwb

Para comprender completamente la solución de Euler, es esencial comprender algunas definiciones básicas en la teoría de grafos:

- Orden: Es el número total de nodos en el grafo. Esta medida es fundamental ya que determina el tamaño del conjunto de elementos que se está tratando en el grafo.

- Tamaño: Es el número total de aristas en el grafo. Esta medida nos da una idea de la complejidad de las conexiones en el grafo y es esencial para comprender la densidad y la conectividad de este.

- Grado: Es el número de aristas incidentes en un nodo y es crucial para entender su importancia y su papel en la red. Los nodos con un grado alto suelen ser nodos importantes en términos de conexiones.

- Distancia: Es la longitud mínima de un camino entre dos nodos. Esta medida es fundamental para comprender la estructura de la red y la proximidad entre nodos en términos de conexión.

Regresando al problema de los puentes, es necesario entender la cantidad de elementos (orden) en el grafo, es decir, el número de islas (nodos) y puentes (aristas) involucrados, además el número total de conexiones 3.2 (tamaño) entre estas islas es crucial para evaluar la complejidad del problema y la viabilidad de encontrar una solución.

El grado de cada nodo, es decir, el número de puentes conectados a una isla específica desempeña un papel crucial en la determinación de su importancia y su contribución a la solución del problema. Además, la distancia entre nodos proporciona información valiosa sobre la estructura de la red y la proximidad entre islas en términos de conexión, lo que influye directamente en la búsqueda de un camino que cruce cada puente exactamente una vez.

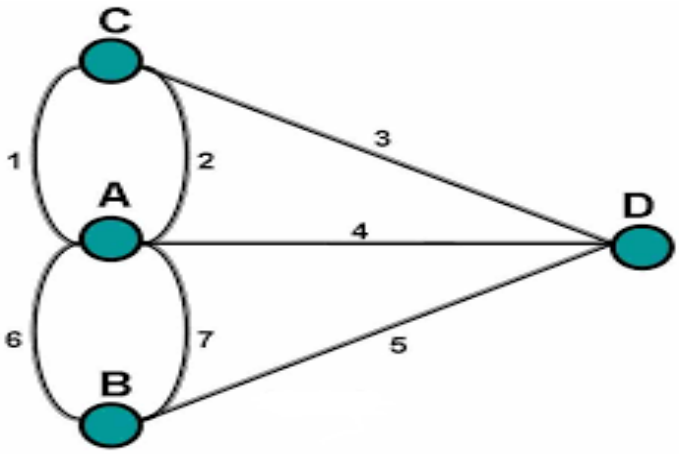


Figura 3.2: Calvo, J. (n.d.). Teoría de grafos para los puentes de Königsberg. Diagrama. Problema simplificado.

3.1.2. Grafo en una red social.

Una red social es una estructura social compuesta por actores (en este caso individuos o nodos) que están conectados por enlaces (aristas), que se pueden interpretar como lazos de amistad o "parentesco" [41] Iribarren y Moro, (2011). Este tipo de relaciones, tanto las individuales como la estructura global de la red, tienen influencia en los individuos. Es decir, bajo este contexto de grafos, una red social es una representación visual y matemáticas entre individuos o entidades. En donde, esta estructura, cada nodo o vértice representa a una persona o grupo dentro de la red, mientras que las aristas o conexiones entre los nodos simbolizan las relaciones o interacciones entre ellos.
Por ejemplo, un análisis de marketing basado en una red, se refiere a un conjunto de técnicas de marketing que aprovechan los vínculos entre los consumidores para aumentar las ventas. [42]

Christakis, N. A., y Fowler, J. H., (2008).

Los nodos, a menudo se representan con círculos, son los actores principales en una red de grafo social, cada uno de estos puede contener información adicional sobre el individuo o grupo que representa (nombre, intereses, ubicación, etc). Por otro lado, las aristas, representadas como líneas que conectan los nodos, indican las conexiones entre los individuos o grupos en la red (conexiones, amistades, interacciones en línea, retweets, etc).

En una red social (grafos social) las aristas pueden ser bidireccionales como se representa en la figura 3.3, en este caso de las amistades mutuas, o unidireccionales, como en el caso de las menciones en Twitter. En la red de retweets que se construirán las aristas que representan amistades mutuas o unidireccionales.

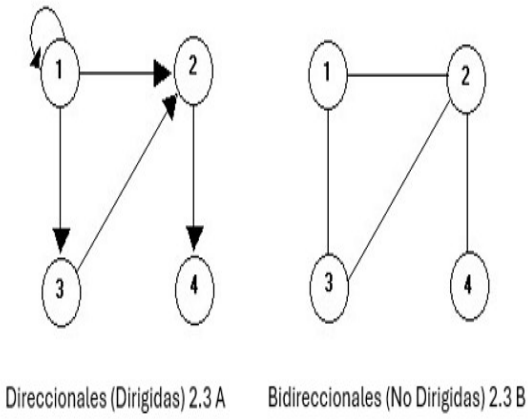


Figura 3.3: Representación de grafo dirigido y no dirigido para la interacción unidereccional y bidireccional de la información.

El flujo en una red de grafo social indica la dirección de los vértices. Por ejemplo, en una red social donde las interacciones son bidireccionales, como en Twitter, el flujo puede ser de ida y vuelta entre los nodos. Sin embargo, puede haber nodos que no tienen flujo entrante o saliente, lo que se puede representar como nodos sueltos dentro de la red.

La creación de una red de grafo para una red social implica recopilar datos sobre las interacciones entre los usuarios. Estos datos pueden incluir quién sigue a quién, quién menciona a quién, quién retuitea a quién, entre otros. Una vez recopilados estos datos, se pueden representar gráficamente como un grafo, con nodos para cada usuario y aristas que conectan los nodos según las interacciones registradas.

Las redes sociales digitales se clasifican en directas e indirectas. Las primeras son aquellas (de carácter generalista) en las que existe una colaboración entre los grupos de personas que comparten algunos intereses comunes y que interactúan en igualdad de condiciones a través de perfiles (con determinados grados de privacidad) mediante los cuales gestionan su información personal y la relación con los otros usuarios. Las redes indirectas (foros y comunidades virtuales), precursoras de las directas, son las que suelen disponer de un perfil identificable o reconocible por el resto de la comunidad, con una persona o grupo (moderador) que controla y dirige la información o las discusiones en torno a temas concretos. [45] (Cuadernosartesanos.org. September 12, 2023)

3.2. Aspectos matemáticos de un grafo en una red social.

Como menciona Alvarez (2013) [48], los grafos son estructuras matemáticas muy útiles sobre las cuales existe una amplia teoría desarrollada y estudiada desde la mismísima época de Euler.

1. **Definición:** . Un grafo es un par G = (V, E), donde V es un conjunto finito no vacío cuyos elementos se llaman **vértices o nodos**, y E es un conjunto cuyos elementos se llaman **aristas o ejes**:

   - Si las aristas son pares no ordenados de vértices de V, entonces diremos que el grafo G es no dirigido. En este caso, denotamos las aristas por $e = u, v$, indicando que la arista e une los vértices u y v.
   - Si las aristas son pares ordenados de vértices de V, entonces diremos que G es un grafo dirigido o dígrafo. En este caso, denotamos las aristas por $e = (u, v)$, indicando que la arista e sale del vértice u y termina en el vértice v.

2. **Grafo no dirigido:** Sea $G = (V, E)$ un grafo no dirigido (o dirigido indistintamente). Un grafo dice [49] Arenado (2023) que es conexo si para cualesquiera dos vértices existe un camino entre ellos. Un dígrafo decimos que es fuertemente conexo cuando desde cualquier vértice se puede trazar un camino orientado hasta cualquier otro vértice. Por lo tanto, se dice que G es un grafo ponderado si cada arista tiene asociado un valor de los reales R.

3.2.1. Modularidad.

Para generar la red de retweets sobre un tema de tendencia utilizando RStudio y detectar comunidades, se puede emplear el concepto de modularidad, como sugiere Newman (2006) [47]. El proceso comienza con la recolección de datos relevantes sobre el tema de investigación a través de la API[3] de Twitter. Para este fin, se utiliza el paquete "rtweet" en RStudio. Posteriormente, se procede a limpiar y depurar los datos, eliminando duplicados y tweets irrelevantes.

Una vez depurada la base de datos, se crea una red de retweets en la que los nodos representan a los usuarios y los bordes indican interacciones, como retweets y respuestas. Esta tarea se lleva a cabo con el paquete "igraph" en RStudio. El concepto de modularidad permite detectar y visualizar comunidades dentro de la red, utilizando herramientas de visualización como "ggplot2" o las funciones específicas del paquete "igraph", el cual asigna colores distintos a cada comunidad, facilitando su análisis. En la figura 3.4 se presentan dos ejemplos de modularidad que ilustran las configuraciones de centralidad y descentralidad.

En el caso que estamos analizando, una red centralizada se construye alrededor de un único participante en el tema de conversación (nodo central) que posee el mayor ranking dentro de esa comunidad, de ahí que los demás nodos o participantes se conectan o ubican alrededor de él. Esto es similar a lo que sucede en la mayoría de los servicios web (aplicaciones de teléfonos celulares, bancarias o servicios de comunicación, entre otros) que están coordinados o administrados por un propietario.

[3]La API de Twitter permite a los desarrolladores interactuar programáticamente con Twitter, facilitando el envío y recepción de datos como tweets, perfiles de usuarios y tendencias, así como realizar acciones como publicar tweets y seguir usuarios.

Por el contrario, en una red descentralizada no se identifica un único participante que tenga mayor influencia en la conversación (influencer), por lo que es muy complicado identificar comunidades. Por ejemplo, una red informática descentralizada distribuye las cargas de procesamiento de información en varios dispositivos en lugar de depender de un servidor central.

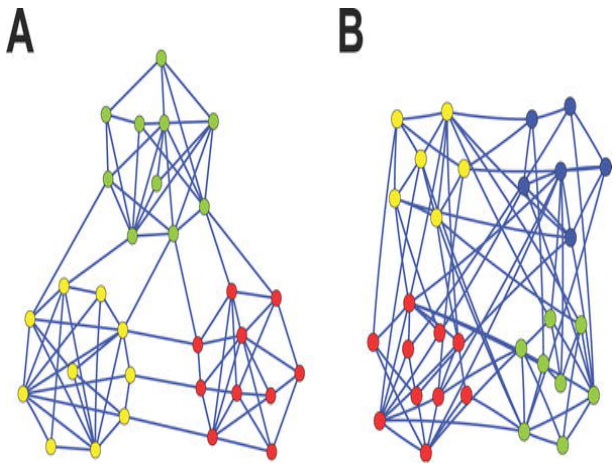


Figura 3.4: Representación centralizada y descentralizada. (2021, November 8). https://is.gd/Q21xRR

Después, se analizó las comunidades detectadas para explorar las interacciones y relaciones dentro de cada una. Identificamos líderes de opinión, usuarios influyentes, etc.

3.2.2. Tipos de comunidades

La relevancia de los algoritmos en la teoría de grafos radica en su capacidad para identificar y medir la importancia de los vértices dentro de un grafo. Estos algoritmos son fundamentales para analizar la estructura de redes complejas, ya que permiten identificar los nodos más relevantes en términos de transmisión de información e influencia. Además, tienen aplicaciones en diversas áreas, como la optimización de motores de búsqueda, el análisis de redes sociales, la identificación de nodos críticos en redes de transporte y, especialmente, la detección de comunidades en redes complejas.

El concepto de modularidad para este tipo de grados es de gran importancia, ya que es un proceso de optimización que permite detectar las estructuras de las diferentes comunidades en la red. Una alta modularidad indica que los nodos dentro de una comunidad están más densamente conectados entre sí que con nodos de otras comunidades, lo que facilita la identificación de grupos bien definidos. Sin embargo, en redes con conexiones aleatorias, como se ilustra en la figura 3.5, la modularidad tiende a ser baja. Esto se debe a que las conexiones suelen distribuirse de forma uniforme, dificultando así la identificación de comunidades cohesivas, a menos que la densidad de aristas sea significativamente alta.

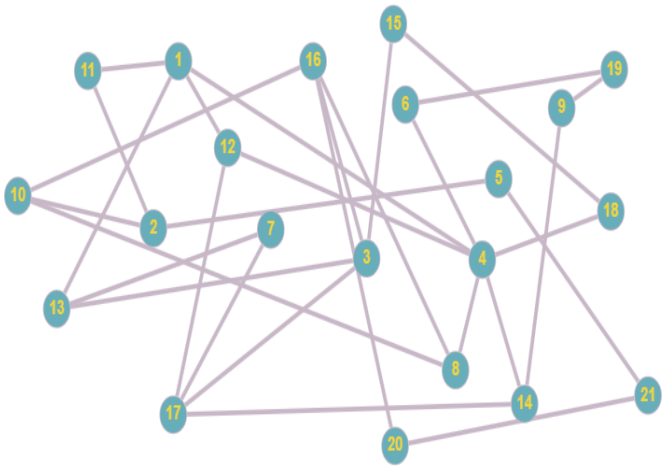


Figura 3.5: Grafo aleatorio. Creación propia. https://graphonline

Para la detección de comunidades se empleo el algoritmo de Louvain, el cual se encuentra en la libreria igraph la cual se implemento mediante un scrip R. Este proceso comienza con la recopilación de datos de la red social en cuestión, en este caso, nos centramos en los retweets en la ciudad. Estos datos incluyen la interconexión de las publicaciones compartidas, amigos en común, listas y más. Una vez visualizada esta red para comprender mejor su estructura y conexiones.

Luego, se aplica los diferentes algoritmos de centralidad[4] que Louvain proporciona para la detección de comunidades. Estos algoritmos ayudan a identificar nodos importantes dentro de la red y cómo se relacionan entre sí. Una vez se han aplicado estos algoritmos, generamos un grafo que representa estas interacciones y comunidades detectadas. Este grafo permite visualizar y analizar de manera más clara cómo se agrupan y organizan las interacciones en nuestra red.

3.2.3. Louvain.

El algoritmo de Louvain, propuesto por [50] (Vincent. 2008), es una herramienta clave para identificar comunidades en conjuntos de datos con conexiones entre elementos, como redes sociales. Su objetivo es encontrar grupos donde las conexiones internas sean fuertes y las externas, débiles, maximizando la "modularidad". Esta métrica evalúa la agrupación de elementos: cuanto mayor sea, más claras serán las comunidades.

Es versátil, aplicándose a conjuntos con conexiones ponderadas o sin peso. Por ejemplo, en Twitter, se puede usar para detectar grupos de usuarios con intereses compartidos. El algoritmo busca maximizar las conexiones internas en las comunidades y minimizar las externas.

La figura 3.6 ilustra la interpretación de la modularidad. Cuanto mayor sea este valor, mejor definidas estarán las comunidades en la red.

[4]La centralidad en la detección de comunidades en un grafo es como identificar quiénes son los nodos más importantes en una red. Estos nodos tienen muchas conexiones con otros nodos en su grupo y son fundamentales para mantener la estructura de la red.

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

donde:

- Q: La modularidad es una medida de qué tan buenos son los grupos encontrados en un conjunto de datos, como una red social. Cuanto mayor sea Q, mejor será la división de la red en grupos.
- A_{ij}: Esto representa cuán fuerte es la conexión entre dos personas (o nodos) i y j en la red. Si dos personas están muy conectadas, A_{ij} será alto; si no están conectadas, será 0.
- k_i y k_j: Representan cuántas conexiones tiene cada persona en total. Si una persona tiene muchos amigos o conexiones, su k será alto.
- m: Es la suma total de todas las conexiones en la red.
- $\delta(c_i, c_j)$: Nos dice si dos personas i y j están en el mismo grupo o comunidad. Si lo están, esto será 1; de lo contrario, será 0.

La fórmula de modularidad básicamente compara la cantidad real de conexiones entre dos personas con la cantidad que esperaríamos si las conexiones se distribuyeran al azar. Luego, multiplica esta diferencia por 1 si las personas están en el mismo grupo y por 0 si no lo están. Finalmente, suma todas estas diferencias para todas las personas en la red y lo divide por el total de conexiones en la red.

En conjunto, la fórmula de modularidad evalúa la diferencia entre la conexión real entre los nodos i y j y la conexión esperada si las conexiones se distribuyeran al azar. Maximizar esta fórmula conduce a encontrar una partición del grafo en comunidades que estén densamente conectadas entre sí y menos conectadas con otras comunidades. El objetivo del algoritmo es encontrar la partición que maximice la modularidad, lo que nos indica que hemos identificado comunidades bien definidas y conectadas internamente.

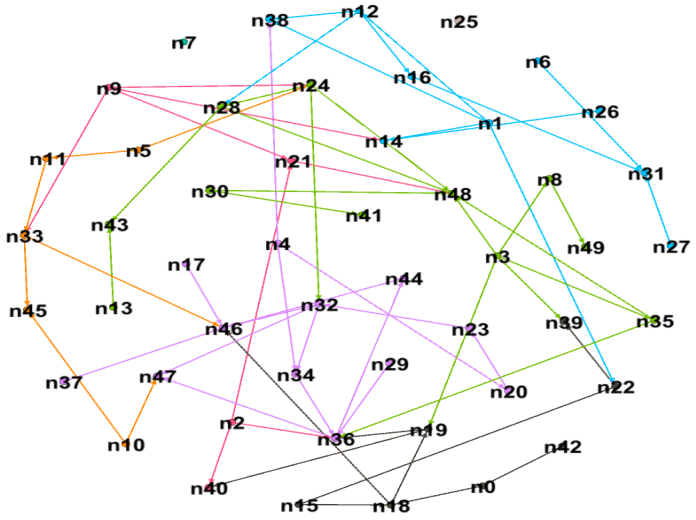


Figura 3.6: Modularidad. Creación propia. Grafo. https://gephi.org/users/download/.

3.2.4. Vector propio (eigenvector) y Centralidad.

A partir de los tweets se crea una red de retweets representada como grafos, para su posterior análisis. Inicialmente, se generaron redes de retweets relacionadas con un hashtag específico y redes de menciones de usuarios en ese hashtag. Sin embargo, también se puede filtrar sub-redes que representan la red de un usuario en particular. La centralidad del vector, como explica (Martínez Arbas 2016) [53], está relacionada con el número de retweets o menciones que recibe cada usuario en la red. Por otro lado, la centralidad del eigenvector muestra en la figura 3.7 a los usuarios más influyentes en la red.

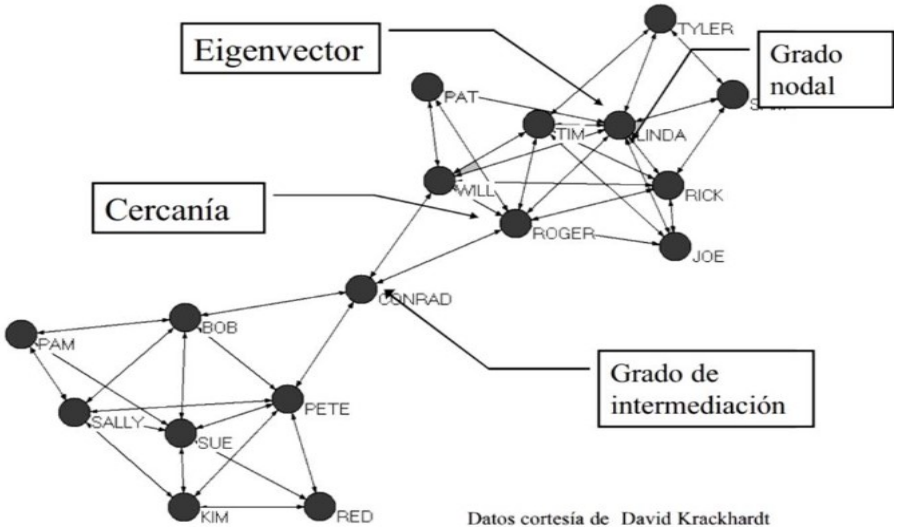


Figura 3.7: Componentes de una red de grafos con centralidad. https://bit.ly/acortado

En el contexto de la detección de comunidades y la importancia de los nodos en un grafo, el eigenvector juega un papel crucial. Este concepto, introducido por (Bonomo 2014) [51], considera la influencia de los nodos cercanos en la importancia de un nodo específico en la red. Esencialmente, el eigenvector condiciona la importancia de un nodo según la relevancia de los nodos conectados a él.

En la práctica, diversos protocolos, como el propuesto por (Martínez 2014) [52], utilizan medidas de centralidad eigenvector para identificar nodos relevantes en un grafo. Estos métodos aprovechan el análisis algebraico de matrices para realizar un análisis dinámico y eficiente de los grafos. Estos enfoques son útiles para encontrar focos importantes dentro de una variedad de estructuras, como imágenes, páginas web o cualquier otro tipo de red.

3.3. Comunidades en la red social Twitter.

Las redes sociales, como Twitter, sirven como plataformas fundamentales para la interacción y el intercambio de información en la era digital. Estas plataformas permiten que las personas establezcan vínculos en diversos contextos, accediendo a información o servicios específicos y participando en interacciones comunicativas dentro de una red social se puede entender mejor en el diagrama de la figura 3.8. Desde una perspectiva técnica, una red social se define como un conjunto de conexiones entre individuos representados como nodos y sus relaciones como aristas, visualizadas mediante grafos, revelando patrones en las relaciones entre actores sociales [56] (Abascal, 2015).

Figura 3.8: Pastel del estudio de la red social Twitter.
(Análisis de Redes Sociales, Mapeo de Conexiones y Redes Mediante Social Media Analytics - FasterCapital, s. f.)

La Teoría de Grafos se emplea para analizar estas interacciones en plataformas digitales como Twitter, donde los nodos representan a los individuos y las aristas simbolizan las relaciones que los conectan. La representación visual como se menciona [55] (Gillen y Merchant, 2013) de patrones de relación entre actores sociales se logra mediante grafos, una estructura que muestra conjuntos de objetos unidos por relaciones y sus interacciones.

> Los *hashtags* (etiquetas) son una serie de caracteres simples o compuestos que van precedidos por la tecla numeral, también llamada "gato", # (...) el *hashtag* en Twitter opera como una especie de clúster o racimo de nodos o lugares de enunciación articulados por él o los términos de referencia, en el que los usuarios del *hashtag* colaboran y se encuentran, discrepan, y disputan sentidos o atacan, descalifican, contra-argumentan o amenazan. [54] (Reguillo, 2018, p.17).

La clasificación de redes sociales se basa en su propósito principal, ofreciendo una visión general de categorías comunes. Estas plataformas son vitales en el mundo digital, conectando personas y proporcionando oportunidades como comunicación, formación de relaciones, intercambio de conocimientos y promoción. Sin embargo, su uso conlleva riesgos como exposición de información personal y propagación de noticias falsas.

3.4. Twitter y R: Herramientas para el análisis de redes sociales.

Como ya se mencionó en la sección 2.1 del capítulo 2 sobre como se solicita el acceso a las llaves, Twitter juega un papel fundamental en la difusión de información y la participación pública en temas de relevancia global, a través del análisis de temas específicos utilizando los *hashtags* de los temas de tendencia, que hoy en día se han convertido en elementos cruciales para la organización y propagación de contenido, subrayando su importancia en la investigación.

3.4.1. Extracción de datos en Twitter.

En cuanto a la extracción de datos Twitter reporta 90 variables que ayudarán a identificar patrones de interacción entre usuarios, así como tendencias y temas emergentes en la plataforma. Utilizamos la API generada (como se habló en la sección 2.1) de Twitter para recopilar esta información de manera ética y garantizar su integridad. Su estructura como red digital facilita la distribución de información y la comunicación bidireccional entre usuarios, convirtiéndola en una fuente invaluable de datos para las investigaciones.

El perfil de usuario en Twitter, incluyendo elementos como la biografía, la foto de perfil y el número de seguidores, impacta en la credibilidad y la influencia de un usuario en la red, ya que el perfil 3.9 puede decir mucho del usuario, si es verificado, no verificado o un Bot[5].

Figura 3.9: Interfaz y partes del Twitter. (2012, December 18). Interfaz. Un Enchufing de Social Media a la sangre.

1. Perfil de usuario: Aunque este punto no se relaciona directamente con el análisis de datos en RStudio, entender la información disponible en el perfil de usuario de Twitter puede ser útil para comprender mejor la dinámica de la red social y cómo influye en la difusión de información.

2. Tendencias (*Trending Topics*): La capacidad de identificar y analizar los temas y hashtags más populares en Twitter puede ser crucial para comprender las conversaciones y tendencias en la plataforma. RStudio puede ser utilizado para analizar estos temas y su relevancia en el contexto del análisis de datos.

3. Siguiendo y seguidores: Estos puntos proporcionan información sobre las interacciones de seguimiento en Twitter, es decir, cuántas cuentas sigue una cuenta y cuántas cuentas la siguen. Analizar estas interacciones puede ser útil para comprender las relaciones entre los usuarios y su influencia en la difusión de contenido.

4. Notificaciones y línea de tiempo (*Timeline*): Estos puntos pueden proporcionar información valiosa sobre la interacción de los usuarios en Twitter, incluyendo menciones, retweets y

[5]Un Bot en una red social es un programa automatizado que imita actividades humanas, como publicar, interactuar con usuarios, promover productos, y monitorear contenido.

likes. Analizar estas interacciones puede ayudar a comprender la dinámica de la red social y la influencia de los usuarios.

5. Tendencia: Al igual que con los Trending Topics, analizar las tendencias en Twitter puede proporcionar información útil sobre los temas y eventos más relevantes en la plataforma. RStudio puede ser utilizado para explorar y visualizar estas tendencias en el contexto del análisis de datos.

3.5. R para el análisis de datos.

La importancia de R y RStudio como una herramienta es esencial para el análisis de datos en diversos contextos, incluida la minería de datos en redes sociales. RStudio permite simplificar el análisis de datos al proporcionar una interfaz (Figura 3.10), además de complementar con herramientas (librerías, igraph, etc.) para la manipulación y visualización de datos.

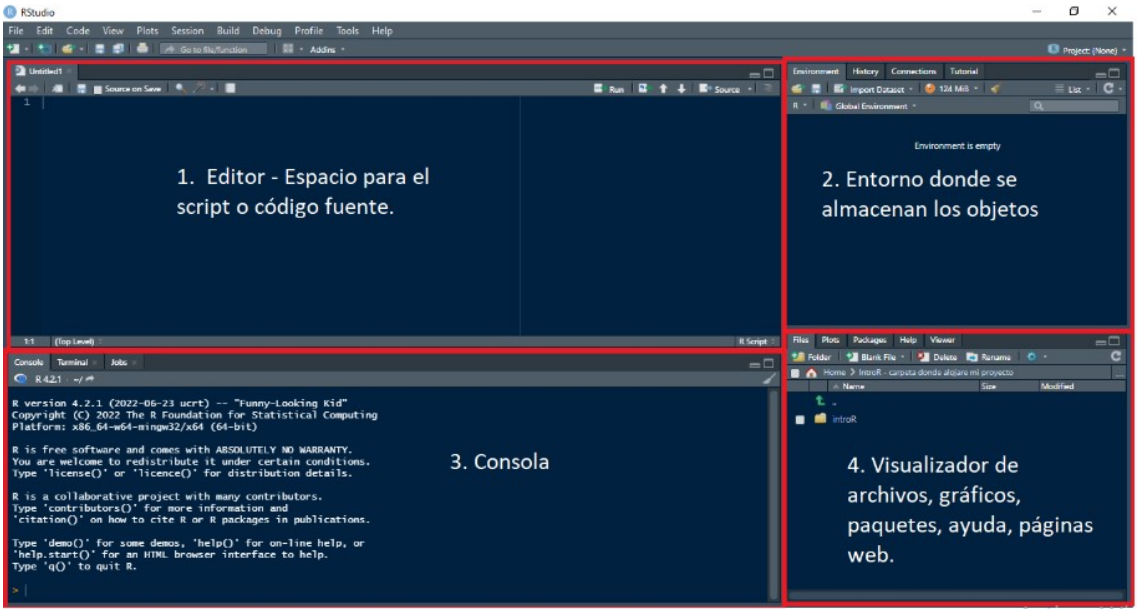


Figura 3.10: (Pantalla principal de RStudio)

En la práctica RStudio ha demostrado ser una herramienta invaluable en el estudio de la minería y análisis de datos extraídos de Twitter. Su utilidad radica en su aplicación en diversas áreas.

Por otra parte, la capacidad de RStudio para crear visualizaciones de gran calidad y significativas a partir de los datos recopilados como gráficos de redes, mapas de calor, entre otros que son de gran utilidad para comprender y analizar la estructura y dinámica de la penetración de la información en la red.

3.5.1. Extracción de la información por el API.

Claro, aquí tienes la relación con el tema de las API de Twitter y la extracción de bases de datos para entender cómo RStudio facilita y mejora el análisis de datos de Twitter 3.11, es crucial comprender sus ventajas fundamentales:

En primer lugar, simplifica la gestión de datos y tareas de programación relacionadas con la investigación. Su interfaz intuitiva agiliza la manipulación y análisis de datos, lo que es especialmente útil cuando se trabaja con grandes volúmenes de información obtenidos a través de las API de Twitter. Esto me permite concentrarme en los aspectos clave del estudio sin perder tiempo en la manipulación manual de datos.

Además, ofrece un entorno de desarrollo más eficiente para trabajar con R. Esto aumenta la productividad y precisión en los análisis al proporcionar características adicionales que facilitan la extracción de datos de Twitter a través de la API y la posterior manipulación de los mismos para su análisis.

Figura 3.11: Logo de Twitter y R. https://is.gd/uXLJeN

Específicamente en el análisis de datos de Twitter, RStudio juega un papel fundamental:

La limpieza de la base de datos es crucial para garantizar la coherencia y precisión de los análisis posteriores. RStudio permite realizar un proceso exhaustivo de la limpieza de datos, especialmente importante al trabajar con información recopilada a través de las API de Twitter, donde pueden surgir datos incompletos o incoherentes.

Además, ofrece herramientas avanzadas para la realización de rankings, lo que me permite identificar a los usuarios más influyentes en la conversación sobre un tema específico utilizando los datos extraídos de las API de Twitter. Esto es esencial para detectar comunidades o grupos de usuarios que tienen un impacto significativo en la difusión de información relacionada con el tema en cuestión.

También es valioso para analizar cómo la información se propaga entre influencers y usuarios en la plataforma. Utilizando técnicas de análisis de redes sociales y los datos obtenidos a través de las API de Twitter, se puede explorar las interacciones y conexiones entre estos grupos, proporcionando una comprensión más profunda de cómo se difunde la información y cómo se forman las comunidades en la plataforma.

En conjunto, RStudio se convierte en una herramienta esencial para llevar a cabo un análisis completo de los datos de Twitter obtenidos a través de las API. Su versatilidad y capacidad para realizar análisis avanzados me permiten explorar la dinámica de la conversación en línea y comprender mejor cómo los *influencers* y los usuarios contribuyen a la propagación de información en la plataforma.

Capítulo 4

Construcción y análisis de la red de retweets.

4.1. Introducción.

En este capítulo se realizará la construcción de la red de retweets, así como el análisis e identificación de posibles comunidades y participantes con mayor influencia mediante un estadístico (*ranking* o métrica) de cinco de las variables identificadas como los principales factores explicativos en la detección de los usuarios con mayor influencia.

4.2. Construcción de las redes de retweets.

Para la construcción de la red de retweets empleamos la variable *retweet_ count*. En la tabla 4.1 se muestra el análisis exploratorio de la variable.

Min.	Q_1	Q_2	Media	Q_3	Max	desv
0.0	6.0	205.0	315.6	498.0	1775.0	387

Tabla 4.1: Resumen estadístico de la variable retweet.

En la figura 4.1 se muestra el gráfico de caja (boxplot) de la variable `retweet_count`, donde se observa que existen 1,147 (6.5 %) de observaciones atípicas. El usuario @CarlosFountain muestra la mayor actividad con 1,775 retweets, seguido por @manuelizo7 con 1,425. El resto de las interacciones, 16,639 (93.5 %), se pueden considerar como comportamientos normales. También se observa que hay 1,874 usuarios que no realizaron ningún retweet.

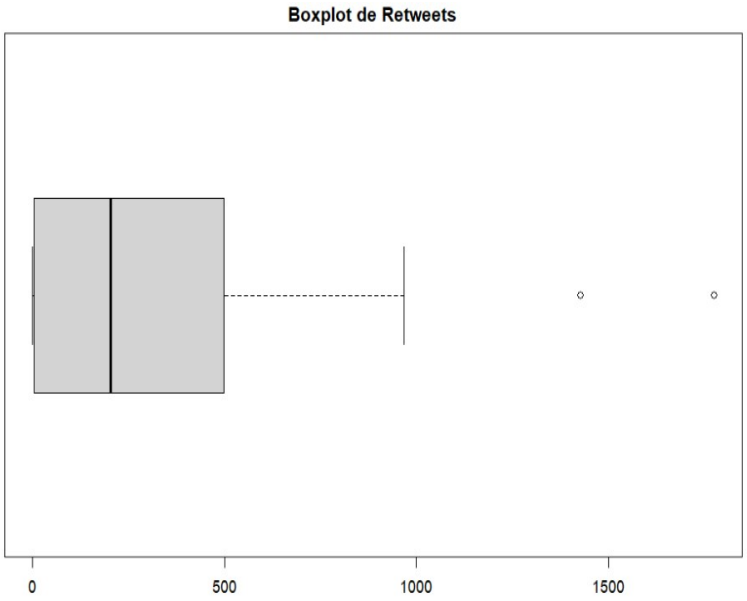


Figura 4.1: Boxplot de retweet.

Por otra parte, 9,012 retweets (50.66 %) caen dentro del rango intercuartílico, 4,363 retweets (24.53 %) caen en el bigote inferior y 3,264 retweets (18.36 %) caen en el bigote superior.

De acuerdo con lo anterior, se puede describir la función de distribución con un comportamiento asimétrico positivo y leptocúrtico, es decir, está sesgada a la derecha y con valores altos en los primeros intervalos del histograma.

Por otra parte, el 50 % de las observaciones centrales está entre el intervalo de 6 y 498 retweets, siendo el promedio 316 retweets, con una gran variabilidad debido a las observaciones extremas (outliers o atípicas).

4.3. Ranking.

Para identificar a los participantes con mayor influencia en la red. El tema de tendencia se construyó una métrica (*ranking*) empleando la función de distribución acumulativa empírica (*ecdf()* (4.2), por sus siglas en inglés), involucrando las siguientes cinco variables cuantitativas que aparecen en cada tweet de la base de datos, las cuales proporcionan el número de interacciones (retweets) entre los participantes.

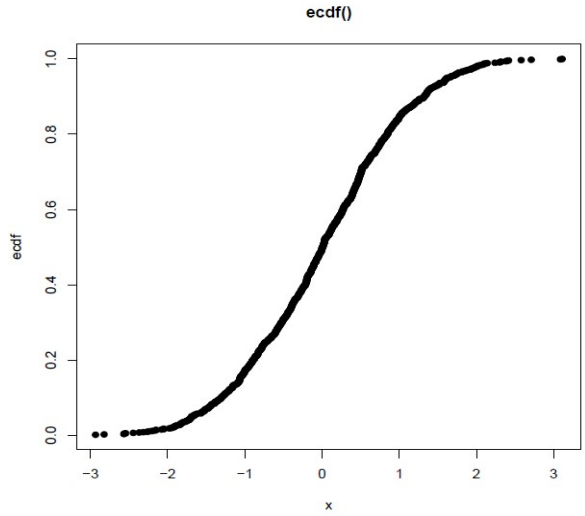


Figura 4.2: *ecdf* de una distribución normal estándar

- followers_percentile = ecdf(followers_count)(followers_count): Calcula el percentil de seguidores para cada usuario basado en su cantidad de seguidores.

- friends_percentile = ecdf(friends_count)(friends_count): Calcula el percentil de amigos (usuarios a los que siguen) para cada usuario basado en su cantidad de amigos.

- listed_percentile = ecdf(listed_count)(listed_count): Calcula el percentil de listados en los que un usuario ha sido incluido.

- favourites_percentile = ecdf(favourites_count)(favourites_count): Calcula el percentil de favoritos (me gusta) que un usuario ha dado.

- statuses_percentile = ecdf(statuses_count)(statuses_count): Calcula el percentil de la cantidad de estados (tweets) que un usuario ha publicado.

- retweet_percentile = ecdf(retweet_count)(retweet_count): Calcula el percentil de la cantidad de tweets que un usuario ha re-subido en su perfil.

Los siguientes pasos se harán en el código para encontrar el ranking que son las cuentas que nos interesan.

a) group_by(screen_name): Agrupa los datos por el nombre de usuario (screen_name). Esto permite tener en cuenta si un usuario cambia su número de seguidores en diferentes momentos del tiempo.

b) summarise: Calcula la puntuación "top_score" para cada usuario, que es la suma de los percentiles calculados anteriormente. Esta puntuación refleja la influencia relativa de un usuario en función de estas métricas. Los usuarios con puntuaciones más altas son considerados más influyentes.

c) ungroup() - Elimina la agrupación, lo que significa que ahora se está trabajando con la tabla de datos en su totalidad.

d) mutate(ranking = rank(-top_score)) ->ranking_de_Usuarios: Agrega una nueva variable "ranking" al conjunto de datos, que se calcula utilizando la función rank() aplicada a la puntuación "top_score". Los usuarios se clasifican en orden descendente (-) según su "top_score", de modo que los usuarios más influyentes obtienen un ranking más bajo.

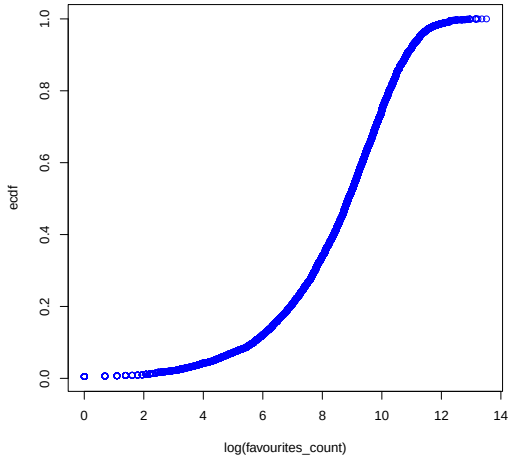
e) ranking_de_Usuarios: Muestra la tabla de datos con el ranking asignado a cada usuario.

f) ranking_de_Usuarios - arrange(ranking): Muestra la tabla ordenada de acuerdo con el ranking, de manera predeterminada en orden decreciente, lo que significa que los usuarios más influyentes aparecerán en la parte superior.

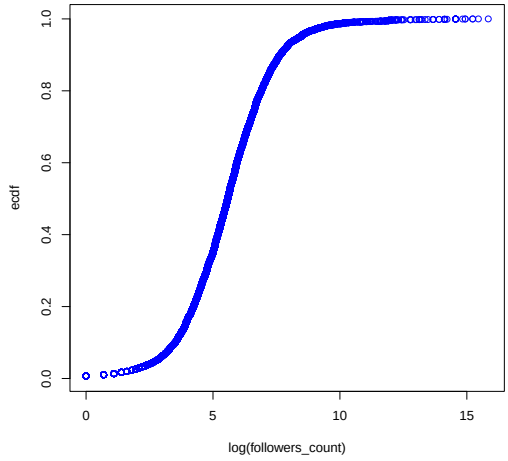
La tabla 4.2 muestra los valores mínimos, promedio y máximo para cada una de las 5 variables.

Estadístico	**Followers_count**	**Friends_count**	**Favourites_count**	**Statuses_count**	**Listen**
Min	0	0	0	1	0
Media	6,044	1,118	19,735	2,8754	36.64
Max	7,601,610	231,206	741,012	1,549,963	24982.00
dev	101296.64	3838.44	37189.11	61492.39	351.93

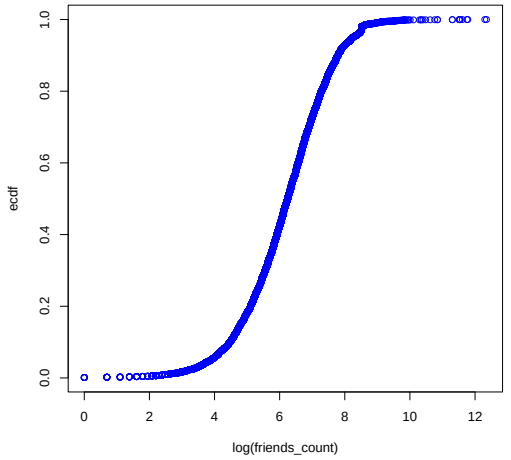
Tabla 4.2: Tabla resumen de los 5 valores (counts).



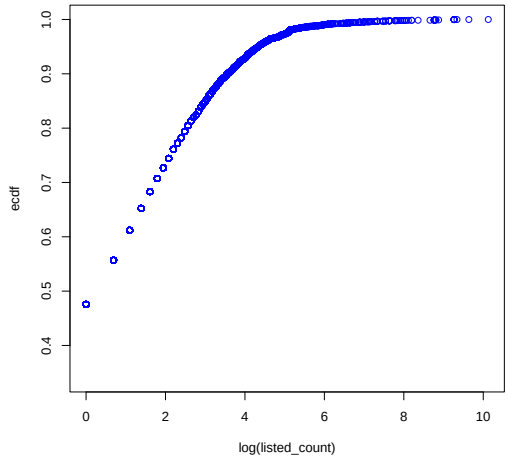
((a)) $favouritesBG\_count$.



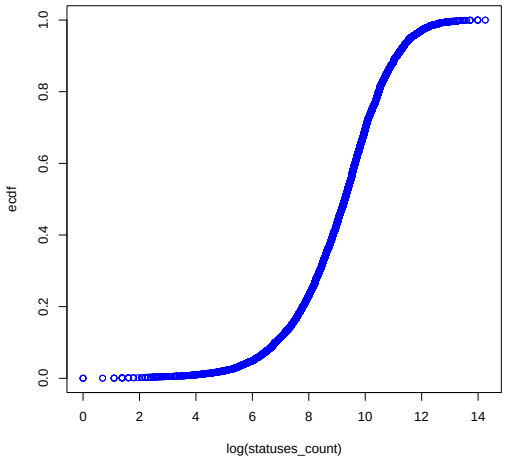
((b)) $followersBG\_count$.



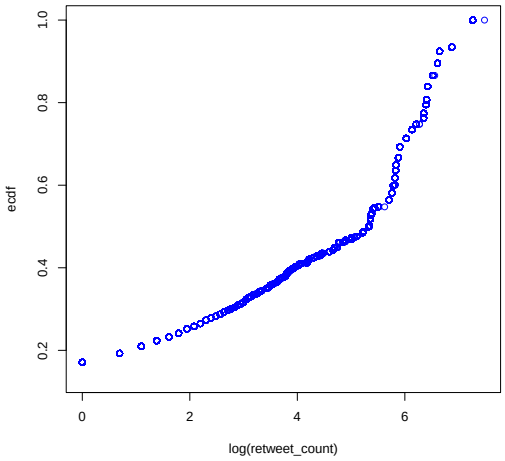
((c)) $friendsBG\_count$.



((d)) $listedBG\_count$.



((e)) $statusesBG\_count$.



((f)) $retweetBG\_count$.

Figura 4.3: Funciones de distribución acumulativa empírica ($ecdf()$) para las seis variables de los tweets de la base general.

La figura 4.3 muestra las funciones de distribución acumulativa empírica *ecdf* de la base general empleadas en la formación del *ranking*. Por otra parte, en la tabla 4.3 se presentan los usuarios con mayor ecdf para cada una de las variables. Por ejemplo, el usuario *elinformateMX* es el que tiene el mayor número de *friends_ count* con 231,206 amigos.

User Name	Ranking	Followers	Friends	Favourites	Statuses	Listen	Score-Producto
lopezdoriga	663	1	0.816665355	0.363656809	0.98881142	0.9999438	4.1169077
		7,601,601	1,583	3,389	285,762	15,328	$1.7862724e^{23}$
ellnformanteMX	542	0.9973575	1	0.25514450	0.9861127	0.9955583	4.234173
		245,822	231,206	1,610	244,981	1,174	$2.63176e^{22}$
CarlosValenz15	299	0.8872709	0.79438885	1	0.9780164	0.7816822	4.441358
		1,852	14,12	741,012	188,543	11	$4.018871e^{18}$
lasillarota	203	0.9977510	0.94703700	0.608231193	1	0.9984820	4.551501
		357,028	3,674	11,704	1,549,963	3,634	$8.647329e^{22}$
El_ Universal_ Mx	63	0.9999438	0.99561453	0.776172279	0.99926909	1	4.771
		5,146,615	13,555	24,912	805,343	24,982	$3.496535e^{25}$
Valores Maximos		1	1	1	1	1	1
		760,1610	231,206	741,012	1,549,963	24,982	$5.042878e^{28}$

Tabla 4.3: Tabla con los valores generales (counts), las funciones de distribución acumulativa empírica (*ecdf*) y el puntaje para el clasificado número 1, así como para el usuario con el mayor número de seguidores.

4.3.1. Red de retweets, base general.

En la tabla 4.4 se muestran los 10 usuarios con mayor *ranking* tomando las 6 variables en la base de datos generales. Cabe destacar que el usuario clasificado en el primer lugar (*influencer*) no es aquel que tiene el mayor valor *ecdf* en las 6 variables utilizadas. De hecho, solo en la variable *statuses_ count* es donde tiene el valor más alto. Sin embargo, la suma de sus 6 valores de *ecdf* es la más alta. También se puede observar el comportamiento de estos mismos usuarios con los valores de sus *ecdf* en la figura 4.4.

User Name	Followers	Friends	Favourites	Statuses	Listen	Score-Producto
ricky_alvarado_	0.9991499	0.9998867	0.9977332	0.999603	0.9933696	4.979915
	123,328	126,788	347,559	754,876	273	$1.119971e^{24}$
adiazpi	0.9987533	0.99966	0.9856058	0.9960898	0.9967698	4.967144
	90,965	100,236	154,011	379,348	414	$2.205403e^{23}$
fconsydig	0.9915562	0.9988099	0.9922362	0.9969965	0.9911028	4.958571
	17,978	19,602	212,630	440,671	218	$7.198421e^{21}$
ferbelaunzaran	0.9966266	0.9983695	0.9654785	0.9922411	0.9961205	4.952035
	165,432	18,331	93,357	320,733	1,275	$1.157728e^{23}$
FrancRodrigu	0.9786911	0.9942089	0.9947712	0.998482	0.9591814	4.929972
	11,540	10,460	226,006	588,174	94	$1.508312e^{21}$
LAURAZAPATAM	0.99449	0.98645	0.9687395	0.9843135	0.9888114	4.927454
	140,759	6,006	99,163	224,675	354	$6.667594e^{21}$
memobarba	0.9963454	0.9998876	0.9724502	0.9561453	0.9951085	4.923377
	163,373	130,105	104,355	121,658	1,059	$2.857750e^{23}$
BerthisBonita	0.9852131	0.9973575	0.9912291	0.9774542	0.9632857	4.920499
	18,020	15,112	207,490	185,878	102	$1.071278e^{21}$
myleneymylene	0.9555268	0.984426	0.9983695	0.9887552	0.9883054	4.919420
	5,046	5,555	367,221	279,664	340	$9.787557e^{20}$
floydu2	0.9875745	0.9988193	0.9884179	0.9763297	0.9560891	4.913664
	24,222	24,532	173,286	180,012	87	$1.612602e^{21}$

Tabla 4.4: Tabla con los valores (counts), las funciones de distribución acumulativa empírica (*ecdf*) y el puntaje para las cinco variables de los diez primeros usuarios en el ranking general.

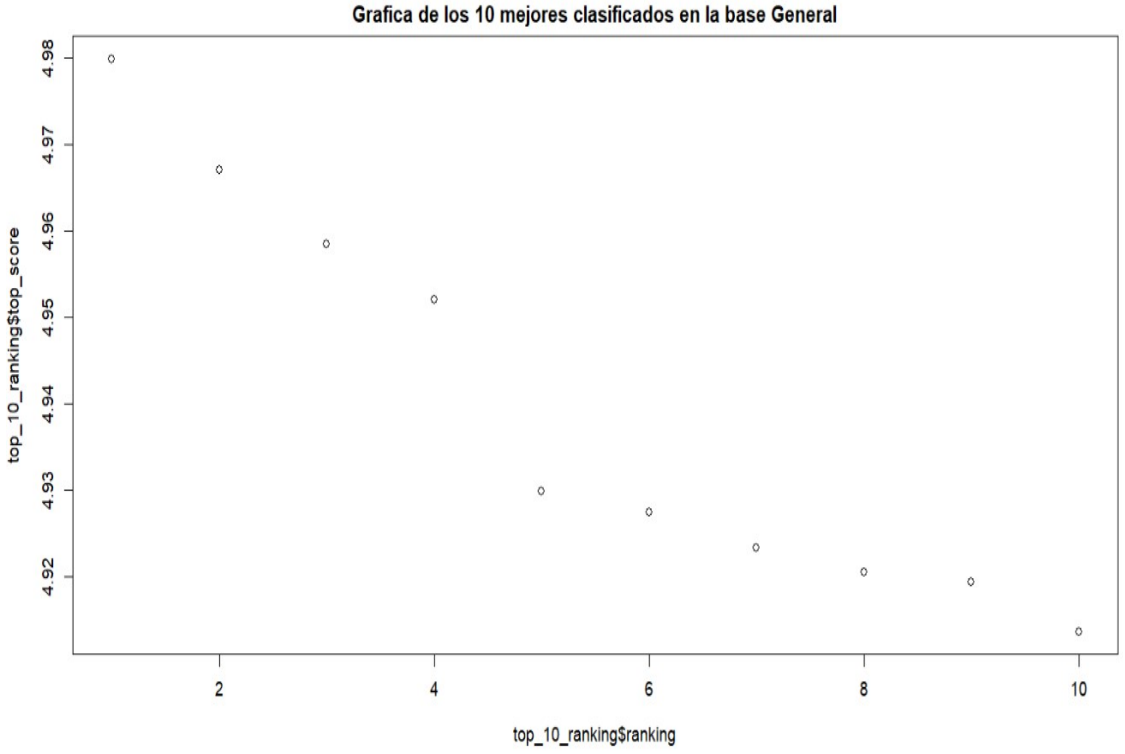


Figura 4.4: Ranking vs top score para los 10 primeros clasificados en la base general.

En la figura 4.5 se muestra la red generada con la variable *retweets_count*, utilizando el algoritmo de Louvain (se debe buscar la referencia de Louvain), y mediante el paquete igraph de

R se genera la red de retweets. A través del código de colores, podemos observar que se pueden distinguir 8 comunidades. En la tabla 4.5 se presentan los usuarios mejor clasificados para cada comunidad.

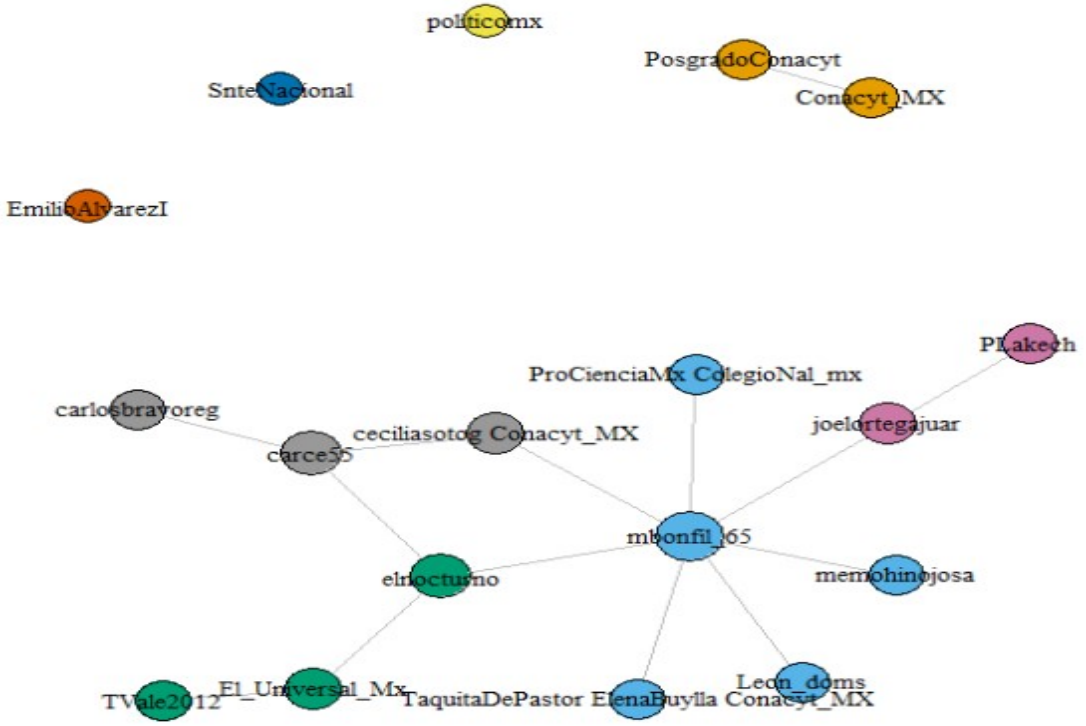


Figura 4.5: Red de retweets base general (Algoritmo Louvain).

Comunidad	Usuarios	Ranking	Score
C_1	Número de usuarios: 2 Usuario más influyente: Conacyt_MX	2,517	3.261167
C_2	Número de usuarios: 5 Usuario más influyente: mbonfil_65	610	4.181905
C_3	Número de usuarios: 4 Usuario más influyente: elnocturno	504	4.265603
C_4	Número de usuarios: 2 Usuario más influyente: politicomx	1,279	3.795791
C_5	Número de usuarios: 2 Usuario más influyente: SnteNacional	3,066	3.052761
C_6	Número de usuarios: 1 Usuario más influyente: EmilioAlvarezI	1,430	3.721266
C_7	Número de usuarios: 2 Usuario más influyente: joelortegajuar	Cuenta con mayor número de menciones sobre el tema de tendencia.	
C_8	Número de usuarios: 3 Usuario más influyente: carce55	1,519	4.564286

Tabla 4.5: Ranking y top_score que ocupan en la base general de los más influyentes por comunidad.

Como se puede observar en la tabla 4.5 los usuarios más influyentes por comunidad no están dentro de los primeros 10 rankeados. Por ejemplo, en la comunidad 3 (C_3) que consta de cuatro usuarios, el usuario más influyente es **elnocturno**, el cual ocupa el ranking 404 y un score de

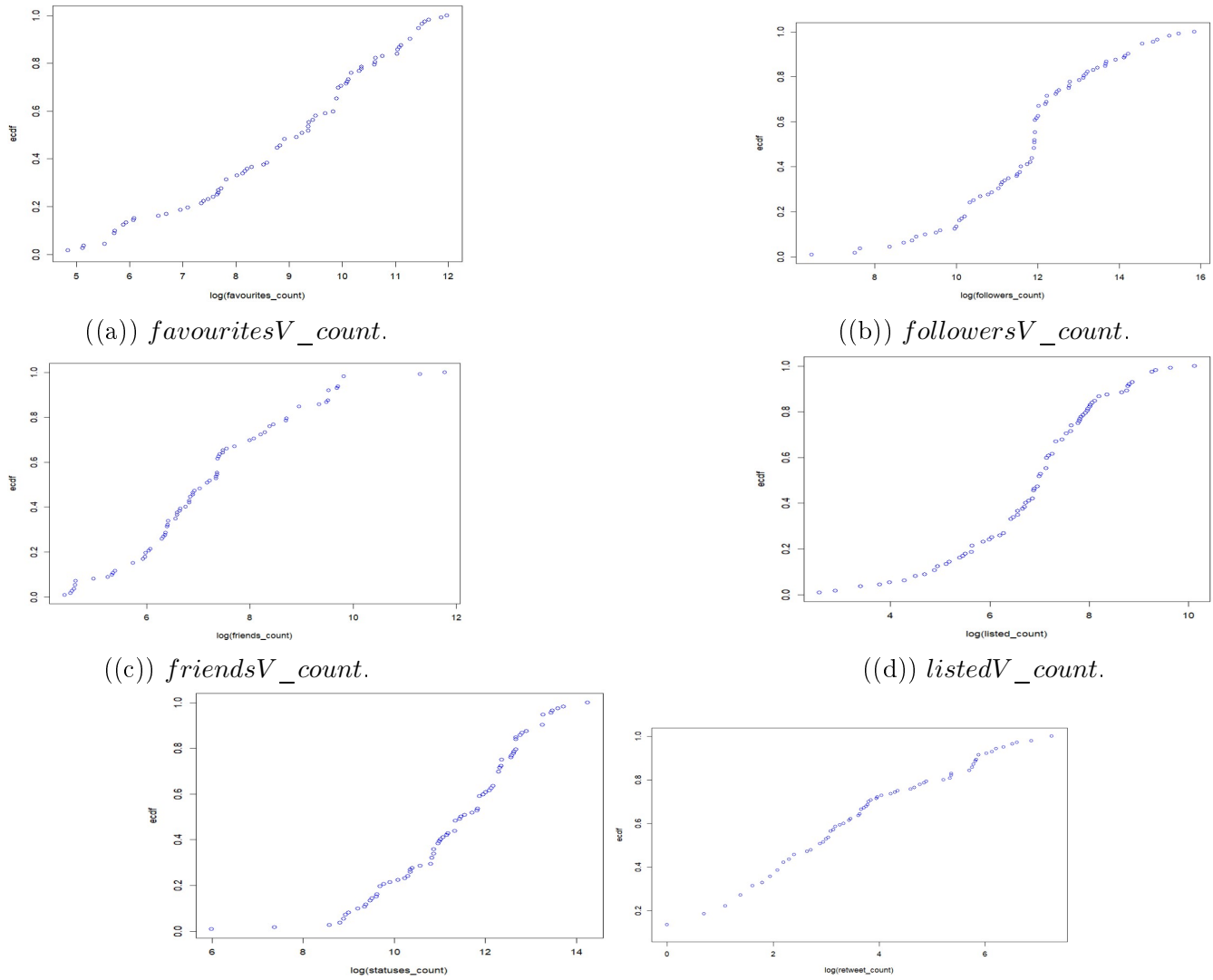
4.265603. Por otra parte, el usuario más influyente de la comunidad 7 (C_7), no fue un participante directo del tema de conversación, sino que fue el que tuvo el mayor número de menciones (**count**) en la variable *mentions_screen_name*.

4.3.2. Red base de datos verificada.

En general, las cuentas (usuarios) en Twitter se pueden clasificar como verificadas y no verificadas. Los usuarios verificados, como se puede ver en el capítulo anterior en la sección de "Extracción de datos en Twitter" en la figura 1.2 etiquetados con una marca de verificación o una paloma azul, son aquellos que están confirmados y autentificados como cuentas de usuarios reales.

La base general se divide en cuentas verificadas y no verificadas, siendo las primeras 140 usuarios de un total de 9,951, lo cual equivale al 1,4 % de los usuarios participantes en el tema de conversación.

En la gráfica 4.6 se muestran las 6 funciones de distribución acumulativa empírica (*ecdf*) (5 para formar el ranking y la de retweet para generar la red) para las cuentas verificadas. Además, en la tabla 4.6 se presentan los usuarios mejor clasificados para cada una de las 5 variables del ranking. Por ejemplo, el usuario *dieguez_* ocupa el primer lugar en la variable *favourites_count* con 158,260 menciones.



((a)) $favouritesV\_count$. ((b)) $followersV\_count$.

((c)) $friendsV\_count$. ((d)) $listedV\_count$.

((e)) $statusesV\_count$. ((f)) $retweetV\_count$.

Figura 4.6: Funciones de distribución acumulativa empírica ($ecdf()$) para las cinco variables de los tweets verificados.

User Name	Ranking	Followers	Friends	Favourites	Statuses	Listen	Score-Producto
lopezdoriga	9	1	0.564285714	0.4	0.692857143	0.992857143	3.65
		7,601,610	1,583	3,389	285,762	15,328	$1.786274e^{23}$
memobarba	10	0.600000000	1	0.978571429	0.492857143	0.435714286	3.507143
		163,373	130,105	104,355	121,658	1,059	$2.85775e^{23}$
*dieguez_*	21	0.278571429	0.750000000	1	0.557142857	0.292857143	2.878571
		67762	4,685	158,260	156,748	707	$5.567861e^{21}$
lasillarota	4	0.721428571	0.714285714	0.564285714	1	0.807142857	3.807143
		357,028	3,674	11,704	1,549,963	3,634	$8.647329e^{22}$
El_Universal_Mx	1	0.992857143	0.835714286	0.785714286	0.950000000	1	4.564286
		514,6615	13,555	24,912	805,343	24,982	$3.496535e^{25}$
Valores Maximos		1	1	1	1	1	1
		7,601,610	130,105	158,260	1,549,963	24,982	$6.060651e^{27}$

Tabla 4.6: Tabla con los valores (counts) verificados, las funciones de distribución acumulativa empírica ($ecdf$) y el puntaje para el clasificado número 1 y el usuario con el mayor número de seguidores.

User_Name	Followers	Friends	Favourites	Statuses	Listen	Producto-Score
El_Universal_Mx	0.9928571	0.8357143	0.7857143	0.95	1	4.57142857142857
	5,146,615	13,555	24,912	805,343	24,982	$3.496535e^{25}$
Pajaropolitico	0.95	0.9214286	0.7428571	0.9285714	0.9785714	4.48214285714286
	20,99,801	13,652	20,436	582,038	10,562	$3.601371e^{24}$
ferbelaunzaran	0.6357143	0.9714286	0.9571429	0.7571429	0.5428571	4.03571428571429
	165,432	18,331	93,357	320,733	1,275	$1.157728e^{23}$
lasillarota	0.7214286	0.7142857	0.5642857	1	0.8071429	3.90178571428571
	357,028	3,674	11,704	1,549,963	3,634	$8.647329e^{22}$
isopixel	0.4714286	0.8285714	0.8428571	0.7714286	0.7428571	3.83035714285714
	150,372	13,219	41,125	351,076	2,629	$7.545069e^{22}$
LeonKrauze	0.7714286	0.8214286	0.8214286	0.3857143	0.7714286	3.78571428571429
	1,388,059	16,493	9,319	137,434	7,154	$2.097587e^{23}$
CanalOnceTV	0.8142857	0.9928571	0.7714286	0.4071429	0.7142857	3.75
	1,095,795	80,998	23,856	69,512	2,382	$3.505926e^{23}$
JohnMAckerman	0.7714286	0.8214286	0.8214286	0.3857143	0.7714286	3.70535714285714
	707,270	11,378	31,871	64,730	3,027	$5.025333e^{22}$
lopezdoriga	1	0.5642857	0.4	0.6928571	0.9928571	3.63392857142857
	7,601,610	1,583	3,389	285,762	15,328	$1.786274e^{23}$
memobarba	0.6	1	0.9785714	0.4928571	0.4357143	3.58928571428571
	163,373	130,105	104,355	121,658	1,059	$2.857750e^{23}$

Tabla 4.7: Tabla con los valores (counts), las funciones de distribución acumulativa empírica (*ecdf*) y el puntaje para las cinco variables de los diez primeros usuarios en el ranking de cuentas verificadas.

4.3.3. Ranking para la base de datos de Tweets verificados.

En la tabla 4.7 se presenta el ranking de las 10 primeras cuentas verificadas, donde se observa que el usuario *El_ Universal_ Mx* ocupa el primer lugar. Sin embargo, en la base general se encuentra en la posición 63 con un *top_ score* de 4.77091.

Como podemos observar, el usuario con mayor influencia sólo muestra un *ecdf* mayor en la variable *listen_ count* con 24,982 menciones. Contrastando con el usuario rankeado en el décimo lugar (*memobarba*) con un *top_ score* de 3.507143, el cual tiene dos variables (Friends y Favourites) el *ecdf* más grande (1 y 0.9785, respectivamente). El comportamiento de estos diez usuarios también se puede ver representado en la figura 4.7.

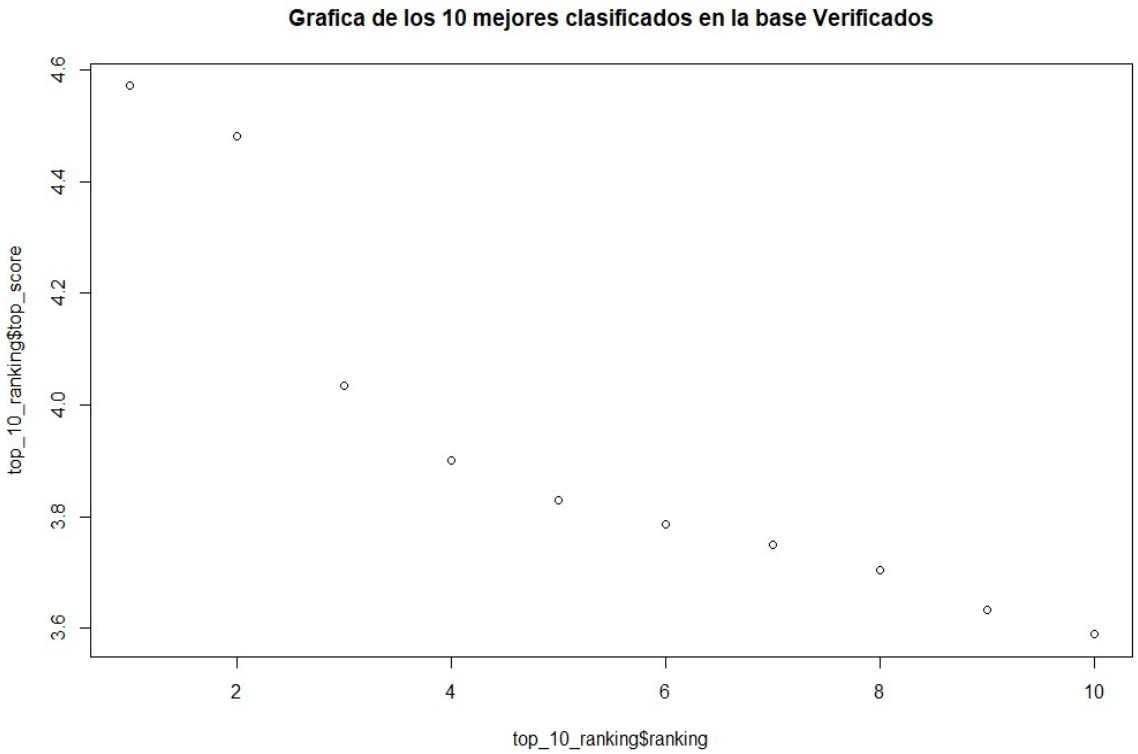


Figura 4.7: Diagrama de dispersión del ranking versus top score para los 10 primeros clasificados en la base de usuarios verificados.

En la figura 4.8 se muestra la red de retweets para la base de usuarios verificados. Como podemos observar mediante el código de colores y el algoritmo de Louvain, se pueden distinguir 5 comunidades. En la tabla 4.8 se presenta el *ranking* y el *top_score* para los usuarios más influyentes en cada una de las 5 comunidades, tanto en la base general como en la base de usuarios verificados.

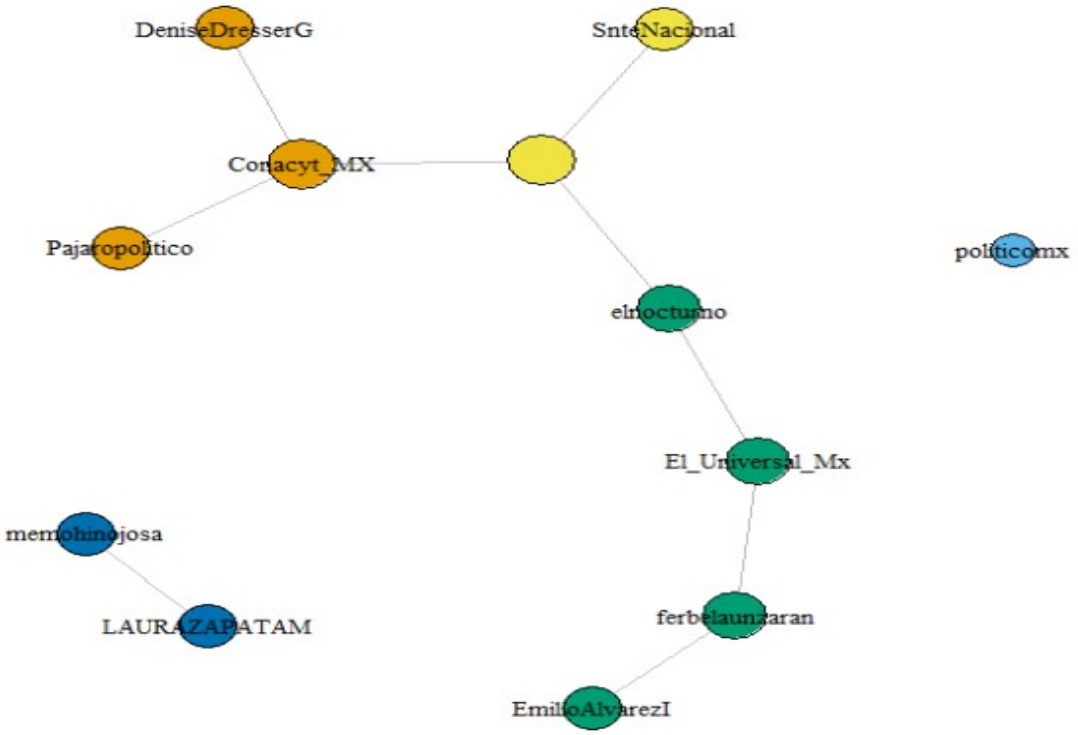


Figura 4.8: Red de retweets base verificados (algoritmo Louvain)

Comunidad	Usuarios más influyentes	Base verificados		Base general	
		Ranking	Score	Ranking	Score
C_1	Número de usuarios:3	53	1.6696429	2,517	3.261167
	Conacyt_Mx				
C_2	Número de usuarios:1	47	1.892857	1,279	3.795791
	politicomx				
C_3	Número de usuarios:4	3	4.035714	4	4.952035
	ferbelaunzaran				
C_4	Número de usuarios:2	65	0.8839286	3066	3.052761
	SnteNacional				
C_5	Número de usuarios:2	18	3.133929	6	4.927454
	LAURAZAPATAM				

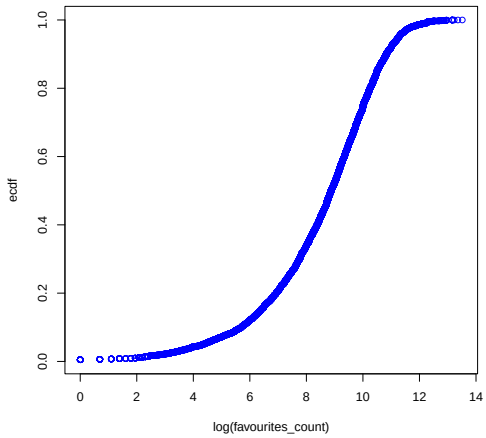
Tabla 4.8: Ranking y top_score que ocupan los usuarios mas influyentes en la base verificados por comunidad y en la base general.

Como podemos observar el ranking y top_score de los más influyentes por cada una de las comunidades es diferente para la base verificados y no verificados, ya que la primera cuenta con 140 usuarios y la base general 16,679 usuarios.

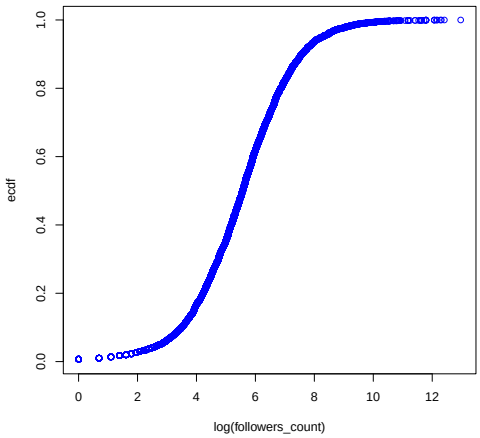
4.3.4. Red base de datos no verificada.

Como ya señalamos, la base de datos no verificados cuenta con 17,646 usuarios, es decir, el 99 % de los usuarios que participaron en el trending topic.

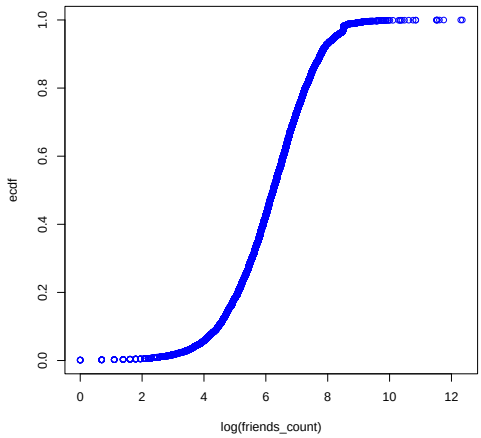
En la figura 4.9 se muestran las gráficas de las funciones de distribución acumulativa empírica (*ecdf*) para las 5 variables que forman el ranking y la variable utilizada para realizar la red de retweets en la base de datos no verificados.



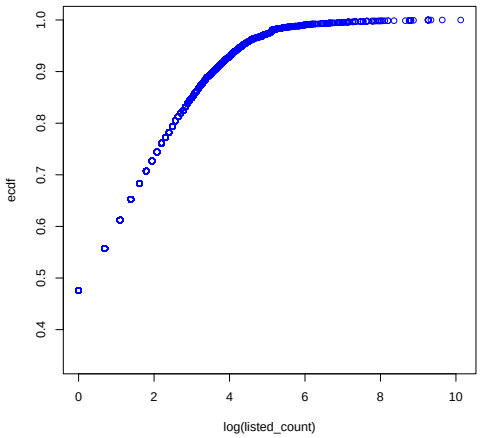
((a)) $favouritesNV\_count$.



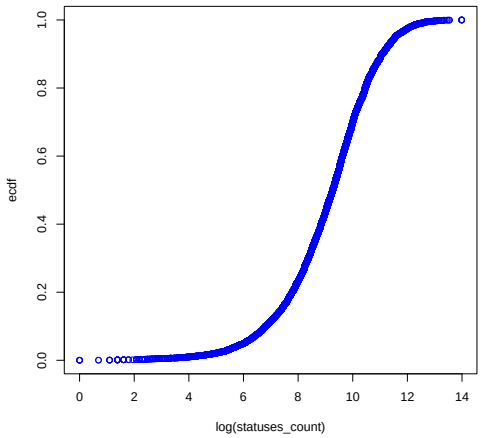
((b)) $followersNV\_count$.



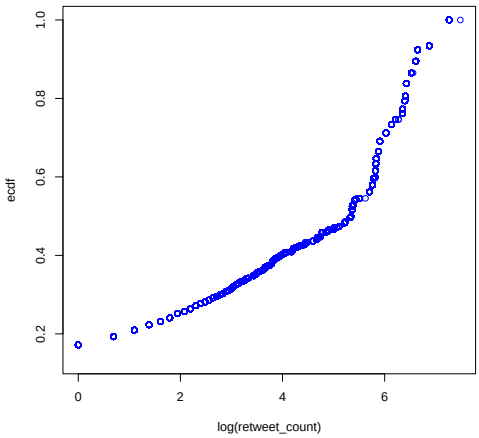
((c)) $friendsNV\_count$.



((d)) $listedBG\_count$.



((e)) $statusesNV\_count$.



((f)) $retweetNV\_count$.

Figura 4.9: Funciones $ecdf()$ para las seis variables de los tweets base no verificados.

En la tabla 4.9 se presentan los usuarios que ocupan el primer lugar para cada una de las 5 variables del ranking en la base de datos no verificados. Por ejemplo, el usuario *NotiGDL* ocupa

la primera posición para la variable *Followers* con 427,598 seguidores.

User Name	Ranking	Followers	Friends	Favourites	Statuses	Listen	Score-Producto
NotiGDL	648	1	0.99937663	0.173127054	0.9807888	0.9986966	4.151989
		427,598	40,893	727	186,459	678	$1.60706e^{21}$
ElInformanteMX	519	0.9999433	1	0.25512864	0.9884960	0.9996033	4.243171
		245,822	231,206	1,610	244,981	1,174	$2.63176e^{22}$
CarlosValenz15	282	0.8941970	0.79638445	1	0.9811289	0.7878839	4.459594
		1,852	1,412	741,012	188,543	11	$4.018871e^{18}$
AdrianaT9735	878	0.9365862	0.07797801	0.999319959	1	0.9927462	4.00663
		2,979	66	420,980	1,186,827	253	$2.482826e^{19}$
RicardoAlemanMx	860	0.9996600	0.62620424	0.405871019	0.9894027	1	4.021138
		174,639	767	4,414	258,562	1,776	$2.715042e^{20}$
Valores Maximos		1	1	1	1	1	1
		427,598	231,206	741,012	1,186,827	1,776	$1.544153e^{26}$

Tabla 4.9: Ranking de la base no verificados de la función de distribución acumulativa empírica (*ecdf*) para los primeros lugares para cada una de las 5 variables.

4.3.5. Red de retweets, base no verificado.

En la tabla 4.10 se presentan los 10 primeros rankeados. Como se puede observar, el usuario *ricky_alvarado* se encuentra en la primera posición con un *top_score* de 4.989678. También se observa que ocupa la primera posición en 3 de las 5 variables empleadas (*Followers*, *Friends* y *Statuses*).

User Name	Followers	Friends	Favourites	Statuses	Listen	Producto-Score
ricky_alvarado_	0.9991499	0.9998867	0.9977332	0.9996033	0.9933696	4.989678
	123,328	126,788	347,559	754,876	273	$1.119970655544106e^{24}$
adiazpi	0.9987533	0.99966	0.9856058	0.9960898	0.9967698	4.977485
	90,965	100,236	15,4011	379,348	414	$2.20540285221735e^{23}$
fconsydig	0.9915562	0.9988099	0.9922362	0.9969965	0.9911028	4.971027
	17,978	19,602	212,630	44,0671	218	$7.19842136513745e^{21}$
FrancRodrigu	0.9858325	0.9956364	0.9947297	0.9990366	0.9662813	4.943200
	11,540	10,460	226,006	588,174	94	$1.50831183466818e^{21}$
BerthisBonita	0.9922929	0.9979599	0.9911595	0.9805622	0.9704182	4.935233
	18,020	15,112	207,490	185,878	102	$1.07127770079934e^{21}$
myleneymylene	0.9628244	0.9862858	0.9983566	0.9911595	0.994673	4.933905
	5,046	5,555	367,221	279,664	340	978755691347641000000
floydu2	0.9943897	0.9990366	0.988326	0.9794854	0.9632778	4.927808
	24,222	24,532	173,286	180,012	87	$1.61260180825573e^{21}$
Cruzamaranta	0.9880993	0.989516	0.9976199	0.9911028	0.955514	4.924368
	13,056	6,619	326,317	278,030	75	588024126197668000000
AntonioChairesC	0.9802221	0.9911028	0.994843	0.9942197	0.9578375	4.920082
	9,095	7,529	230,044	326,756	78	401484778680591000000
diegodelunamx	0.9550606	0.9822623	0.9807888	0.9992066	0.9981866	4.917969
	4,395	5,004	128,471	611,458	496	856898915133095000000

Tabla 4.10: Tabla con los valores (counts), las funciones de distribución acumulativa empírica (*ecdf*) y el puntaje para las cinco variables de los diez primeros usuarios en el ranking de la base de datos no verificados.

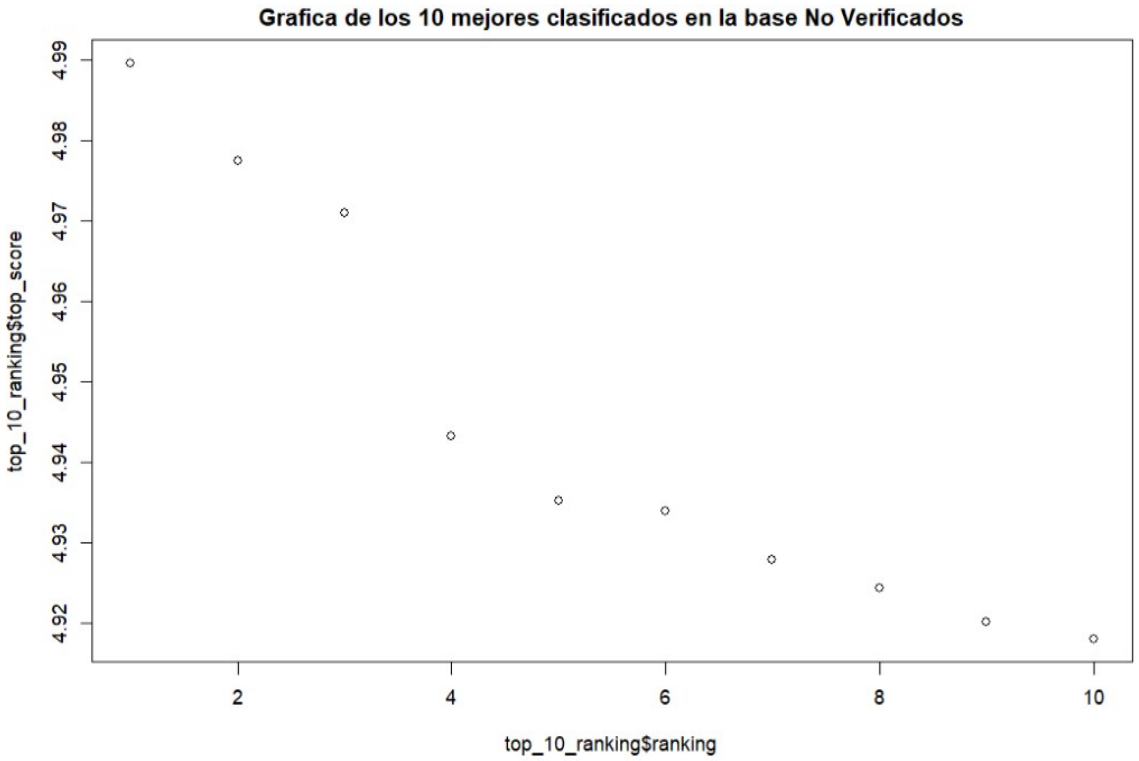


Figura 4.10: Diagrama de dispersión del ranking versus top score para los 10 primeros clasificados en la Base de usuarios no verificados.

En la figura 4.11 se presenta la red generada con la variable *retweets_ count*, realizada mediante el algoritmo de Louvain. El algoritmo de Louvain ayuda para la detección de comunidades, el comportamiento de estos diez usuarios también se puede ver representado en la figura 4.10 y es usado principalmente en grandes redes o grafos, pues es uno de los algoritmos más rápidos en encontrar comunidades en estas redes. Aparte de detectar las comunidades, nos muestra la jerarquía de las comunidades a diferentes escalas [60] (Cruces S., 2022).

Su principio de funcionamiento es maximizando la puntuación de modularidad para cada comunidad en cada iteración para así evaluar cómo de densamente está conectada esta comunidad comparándolos con los de una red aleatoria y conseguir una partición óptima del grafo en comunidades. Este algoritmo nos ayuda a identificar patrones y estructuras que no son posible a simple vista [61] (E. Hodler, 2021.) y utilizando el paquete igraph de R. A través del código de colores, podemos observar que se pueden distinguir 7 comunidades. En la tabla 4.11 se muestran los primeros rankeados para cada comunidad, junto con su *ranking* y *top_ score* que ocupan tanto en la base general como en la base de datos no verificados.

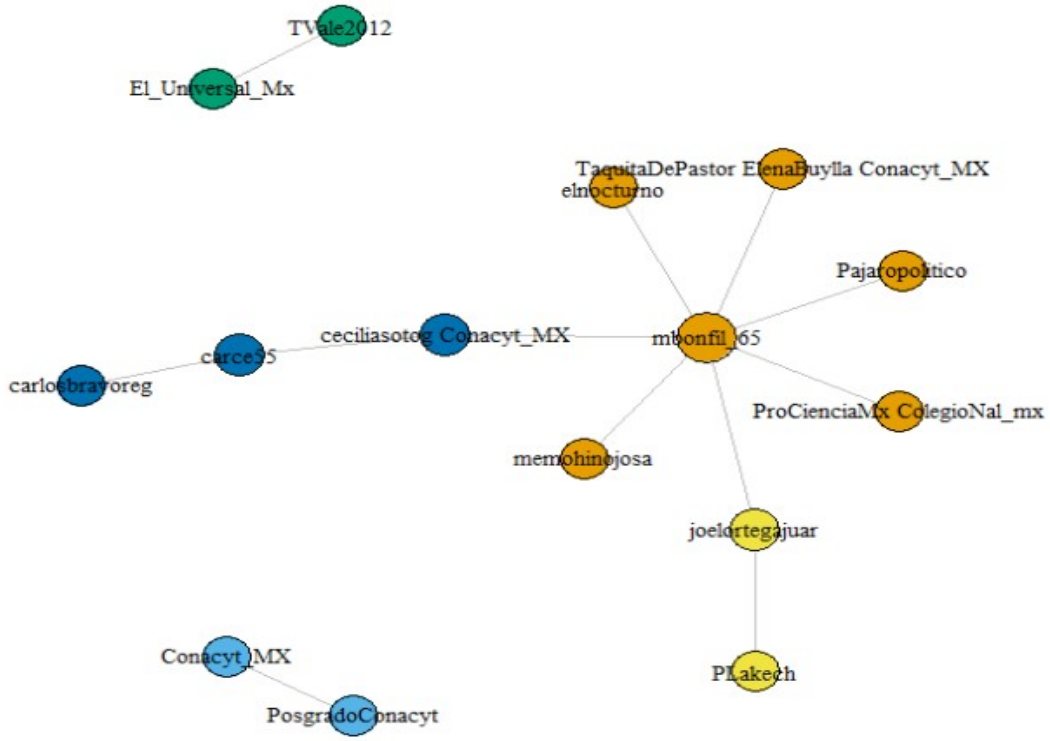


Figura 4.11: Red de retweets base no verificados (algoritmo Louvain)

Comunidad	Usuarios más influyentes	Base no verificados		Base general	
		Ranking	Score	Ranking	Score
C_1	Número de usuarios:6	579	4.199252	610	4.181905
	mbonfil_65				
C_2	Número de usuarios:2	7,155	1.654132	7,231	1.643324
	PosgradoConacyt				
C_3	Número de usuarios:2	486	4.276031	516	4.258948
	TVale2012				
C_4	Número de usuarios:2	Cuenta con mayor número de menciones sobre el tema en tendecia.			
	joelortegajuar				
C_5	Número de usuarios:3	1,471	3.69566	1,519	3.6780982
	carce55				

Tabla 4.11: Ranking y top_score que ocupan los usuarios más influyentes por comunidad para la base de datos no verificados y en la base general.

Capítulo 5

Conclusiones.

El primer punto a resaltar tiene que ver con la base de datos empleada. Se tuvo el inconveniente de que cambiaron las políticas de uso e inclusive el nombre de Twitter a X, además de la inscripción y gratuidad para ser considerado como desarrollador y uso de la API. Sin embargo, los algoritmos desarrollados y el análisis se pueden considerar atemporales, es decir que se pueden aplicar a cualquier base de datos extraída de Twitter tomando las mismas variables, previa limpieza y preparación de los datos. Es así, que la base de datos utilizada comprendió un periodo que va del 21 de septiembre de las 17:39hrs. del 2019 hasta las 18:01hrs. del 23 de septiembre del mismo año, con trending topic #Conacyt, el cual se derivó de las diferencias entre la directora de CONACYT (Dra. María Elena Álvarez-Buylla) y el Dr. Antonio Lazcano, por su remoción de los comités de evaluación en los que participaba.

La dimensión de la base de datos (base general) durante este periodo fue de 17,786 Twetts de 9,951 cuentas distintas y el 98.4 % de ellos fueron escritos en español. Como se mencionó, la base general se filtró en dos bases de datos, usuarios verificados la cual contenía 140 tweets (menos del 1 %) y no verificados la cual contenía 17,646 tweets (99.21 %).

En la tabla 5.1 se muestra un resumen de los usuarios más influyentes en tema de conversación para las bases de datos mencionadas y sus respectivas comunidades.

Base	User_name	Ranking		Score	
		General	Comunidad	General	Comunidad
General	ricky_alvarado_	1		4.979915	
Comunidades					
C_1	Conacyt_MX (V)	2,517		3.261167	
C_2	mbonfil_65 (NV)	610		4.181905	
C_3	elnocturno (V)	504		4.265603	
C_4	politicomx (V)	1,279		3.795791	
C_5	SnteNacional (V)	3,066		3.052761	
C_6	EmilioAlvaresI (V)	1,430		3.721266	
C_7	joelortegajuar	Usuario en lista de distribución con mayor menciones.			
C_8	carce55 (NV)	1,519		3.6780982	
Verificados	El_Universar_Mx	1		4.564286	
Comunidades					
C_1	Conacyt_MX	2,517	53	3.261167	1.6696429
C_4	politicomx	1,279	47	3.795791	1.892857
C_3	ferbelauzaran	4	3	4.952035	4.045714
C_4	SnteNacional	3,066	65	3.052761	0.8939286
C_5	LAURAZAPATAM	6	18	4.927454	3.133929
No Verificado	ricky_alvarado_	1		4.989678	
Comunidad					
C_1	mbonfil_65	610	579	4.181905	4.199252
C_2	PosgradoConacyt	7,231	7,155	1.643324	1.654132
C_3	TVale2012	516	486	4.258948	4.276031
C_4	joelortegajuar	Usuario en lista de distribución con mayor menciones.			
C_5	carce55	1,519	1,471	3.678098	3.69566

Tabla 5.1: Usuarios más influyentes para la base general, verificados, no verificados y sus respectivas comunidades

Como se mencionó, no necesariamente el usuario con mayor valor nominal o *ecdf* en alguna de las cinco variables que forman el ranking es el más influyente, sino que el ranking se conforma de una combinación de cada uno de las cinco *ecdf()* para cada usuario. Por otra parte, no existe una definición precisa de lo que es un influencer, es decir, no solo importa cuantos seguidores tiene, sino también "las relaciones" con los demás usuarios del tema de conversación. Por ejemplo, si un usuario cuenta con cien mil seguidores y esos seguidores no siguen a nadie, los potenciales lectores de un mensaje enviado o tuiteado de esa cuenta sólo llegaría a cien mil posibles lectores. En contraste, si un usuario tiene mil seguidores, pero estos tienen en promedio cinco mil seguidores cada uno, un mensaje tuiteado por este usuario tendría un potencial de cinco millones de posibles lectores. Esto último se conoce como la paradoja de la amistad (**Friendship Paradix Redux**) planteada y explicada por Nathan O. Hodas et al. [63] que establece *tus amigos tienen más amigos que tú, en promedio*. Esta paradoja surge porque las personas extremadamente populares (deportista, artistas, políticos, etc.), están sobre representados cuando se promedia sobre los amigos.

Así, por ejemplo, en las tablas 4.3, 4.6 y 4.9 los usuarios con valores nominales mayores en cada una de las variables no son los más influyentes de acuerdo a la construcción del ranking empleado para identificar los usuarios más influyentes tablas 4.4, 4.5, 4.7, 4.8, 4.10 4.11.

Las futuras líneas de investigación que, a mi juicio, podrían ser:

- Realizar análisis temporales para capturar la evolución de la influencia de los usuarios a través del tiempo o si solamente son tendencias esporádicas. Este enfoque permitiría identificar patrones (estacionarios, de tendencia, estacionales o cíclicos) de eventos relevantes y quizás cambios en la dinámica en la construcción de la red social.
- Realizar un análisis multifactorial o clusters para identificar comunidades y sus respectivos influencers en cada una de ellas.
- Dado la naturaleza multidisciplinaria del tema se podrían agregar variables diseñadas por expertos en comunicación, sociología y políticos para una interpretación más robusta de los resultados, es decir, la intersección de diversas disciplinas podría enriquecer la comprensión las complejidades del fenómeno (análisis de redes complejas).

Finalmente, el aprendizaje y habilidades adquiridos durante el desarrollo de la investigación fueron variados, abarcando áreas clave para la formación de una investigación versátil. La redacción de un artículo en un tema tan multidisciplinario no sólo fortaleció mi capacidad de síntesis y comunicación científica, sino que también me permitió entender la importancia de integrar conocimientos de diferentes campos para abordar problemas complejos.

Aprendí a manejar la información de manera eficiente y descubrí el poder de las bases de datos en la investigación de temas en tendencia. Esto incluyó la capacidad de rastrear y analizar datos en tiempo real, quiénes, dónde y a qué hora se publicaron tweets, lo que me permitió obtener una visión más precisa y dinámica del comportamiento social en las plataformas digitales.

El dominio de los distintos tipos de análisis estadísticos fue otro aspecto clave de mi aprendizaje. Aprender a aplicar estos análisis y representarlos en comunidades, como lo hace el algoritmo *Louvain* [60], me enseñó a interpretar patrones y relaciones ocultas en los datos. Esta habilidad es esencial para transformar datos en información significativa que puede influir en la toma de decisiones.

Además, adquirí conocimientos técnicos específicos, como el uso de software estadístico y herramientas de visualización de datos. La experiencia práctica en la aplicación de algoritmos de clustering y análisis de redes, me preparó para enfrentar futuros desafíos en cualquier campo que requiera el análisis de datos complejos.

Apéndice A

Apéndices

A.1. Tabla con la descripción con las variables que genera cada tweet

Variables de estudio.	
Descripción.	**Detalles.**
No. 1 User_id. Identidad de Usuario	El User_id en R es un valor numérico único que identifica a cada usuario en Twitter. Se utiliza para operaciones y análisis precisos, como el seguimiento de seguidores y el análisis de interacciones, y es distinto del nombre de usuario (username).
No. 2 Status_Id. Id de estudio	El status_ID en R es un identificador numérico único de cada tuit en Twitter. Sirve para identificar y referenciar tuits específicos, facilitando el seguimiento de interacciones y el análisis de datos.
No. 3 Created_at. Creado en	Los campos Created_at deben devolver la fecha/hora en que fue publicado el tuit.
No. 4 Screen_name. Nombre de Usuario	El nombre de usuario debe tener entre 4 y 15 caracteres, solo puede contener letras, números y guiones bajos, sin espacios. El nombre para mostrar puede tener hasta 50 caracteres.
No. 5 Text. Texto	Número de caracteres que tiene el tuit.
No. 6 Source. Fuente	Tipo de dispositivo desde el que se tuiteó.
No. 7 Display_text_width. Ancho del texto visualizado	El número de caracteres del tuit.
No. 8 Reply_to_status_id. Respuesta al id de estado	Si el tuit es una respuesta, este campo contendrá el ID del tuit original, permitiendo representar una cadena de respuestas.
No. 9 Reply_to_user_id. Reproducir al ID de usuario	Si el tuit es una respuesta, se crea un identificador numérico del usuario que publicó el tuit al que se está respondiendo.
No. 10 reply_to_screen_name. Responder a un usuario	Es una respuesta a un tuit de otro usuario que comienza con el @nombredeusuario de la persona a la que se responde. Para responder, se hace clic en el botón "Responder" de un tuit. Las respuestas a tus tuits comienzan con tu @nombredeusuario y aparecen en tu pestaña de Notificaciones.

Continuará en la siguiente página.

Tabla A.1 – Continuación de la página anterior.

Descripción.	Detalles.
No. 11 is_quote. Es cita	Esta variable ayuda a analizar y clasificar los tuits según si son citas. Permite entender la estructura de las conversaciones y la interacción entre usuarios. ***TRUE***: El tuit es una cita de otro tuit. ***FALSE***: El tuit no es una cita y es independiente.
No. 12 is_retweet. Es retweet	Esta variable analiza y clasifica los tuits como retuits o no, proporcionando información sobre la viralidad y difusión del contenido en Twitter. El análisis de retuits ayuda a identificar la popularidad y alcance de un tuit o usuario. ***TRUE***: El tuit es un retuit de otro tuit, compartido para que los seguidores lo vean. ***FALSE***: El tuit no es un retuit y ha sido creado originalmente por el usuario.
No. 13 favorite_count. Recuento de favoritos	La variable favorite_count, obtenida a través de la API de Twitter, indica cuántos usuarios han marcado un tuit como favorito usando el botón "Me gusta". Un número alto en favorite_count sugiere una mayor apreciación o relevancia del contenido publicado.
No. 14 retweet_count. Recuento de retuits	La variable retweet_count indica cuántos usuarios han compartido un tuit en su línea de tiempo, reflejando la popularidad y el alcance del tuit. Un número alto en retweet_count significa que más usuarios han compartido ese tuit. El nombre de la columna puede variar según la estructura del conjunto de datos.
No. 15 reply_count. Recuento de repeticiones	El dato reply_count en R indica el número de respuestas que ha recibido un tuit en Twitter, obtenido a través de la API de Twitter. Este valor refleja la interacción y participación del público, destacando la cantidad de respuestas directas y comentarios generados en torno al tuit. Analizar reply_count es crucial para evaluar la participación y el nivel de conversación generado por el contenido en Twitter.
No. 16 quote_count. Recuento de citas	El dato quote_count en Twitter representa cuántas veces un tuit ha sido citado con comentario. Este valor, obtenido mediante la API de Twitter, indica cuántos usuarios han retuiteado un tuit añadiendo un comentario personal. El quote_count refleja la cantidad de participación generada por un tuit a través de citas, mostrando el nivel de participación que ha provocado el contenido en la plataforma.
No. 17 hashtags. Etiquetas	Este nos da una recién creada función que cuenta el número de elementos de una lista de Hashtags en tendencia.

Continuará en la siguiente página.

Tabla A.1 – Continuación de la página anterior.

Descripción.	Detalles.
No. 18 urls_url. URLS_URL	El dato urls_url en Twitter representa las URLs incluidas en un tuit. Esta variable, obtenida mediante la API de Twitter, almacena los enlaces a sitios web externos, artículos, imágenes, vídeos u otros contenidos relacionados que están vinculados en el tuit. Analizar urls_url permite identificar las fuentes de información o recursos externos mencionados en los tuits. Esto proporciona una comprensión más completa del contenido compartido y las referencias utilizadas en Twitter.
No. 19 urls_t.co. URLS_t.co	Cuando se obtienen datos de tuits a través de la API de Twitter, la variable urls_t.co almacena las URLs acortadas por el servicio de Twitter conocido como "t.co". Estas URLs acortadas ayudan a reducir la longitud de las direcciones URL originales, permitiendo a los usuarios compartir enlaces dentro del límite de caracteres establecido por Twitter.
No. 20 urls_expanded_url. URLs Expandidas	El dato urls_expanded_url en Twitter representa las URLs completas y expandidas incluidas en un tuit. Esta variable, obtenida mediante la API de Twitter, almacena las direcciones URL originales y completas a las que apuntan los enlaces dentro del tuit.
No. 21 media_url. URL Multimedia.	El dato media_url en Twitter representa la URL de los archivos multimedia adjuntos, como imágenes, videos o GIFs, en un tuit. Esta variable, obtenida mediante la API de Twitter, almacena la dirección URL que enlaza al contenido multimedia compartido junto con el tuit.
No. 22 media_t.co. t.co Multimedia.	Este término posiblemente esté relacionado con una columna personalizada o definida por el usuario en un conjunto de datos específico.
No. 23 media_expanded_url. URL ampliada de multimedia.	Es importante notar que Twitter usa su servicio de acortamiento de URL (t.co) para comprimir las direcciones en los tuits, lo que puede ocultar la URL completa de los medios adjuntos. La columna media_expanded_url proporciona la URL expandida del archivo multimedia, es decir, la dirección completa después de redireccionarse desde la URL acortada.
No. 24 media_type. Tipo de medio	El análisis de multimedia_type permite clasificar y entender los diversos tipos de medios compartidos en los tuits. Esto es útil para identificar qué tipo de contenido multimedia está presente en los datos y realizar análisis específicos basados en los tipos de medios adjuntos.
No. 25 ext_media_url. URL extendida de multimedia.	El dato ext_media_url parece ser otra columna inventada por el usuario en un contexto específico, posiblemente relacionada con una columna personalizada o definida por el usuario en un conjunto de datos particular.
No. 26 ext_media_t.co. t.co extendida de multimedia.	Dato inventado por el usuario para un contexto particular.

Continuará en la siguiente página.

Tabla A.1 – Continuación de la página anterior.

Descripción.	Detalles.
No. 27 ext_media_expanded_url. URL ampliada de medios ext.	El dato ext_media_expanded_url representa la URL expandida de un medio adjunto en un tuit de Twitter, pero se refiere a un enlace externo fuera del dominio de Twitter.
No. 28 ext_media_type. Tipo de medio ext.	El dato ext_media_type representa el tipo de medio adjunto externo en un tuit de Twitter. Este valor indica la categoría del medio adjunto, como "photo" (imagen), "video" (video), "audio" (audio) u otros tipos de contenido.
No. 29 mentions_user_id. Menciona el id de usuario	El dato mentions_user_id en Twitter representa el identificador único de un usuario mencionado en un tuit. Es crucial para rastrear interacciones y evaluar la relevancia de menciones en conversaciones específicas. Este análisis permite entender las dinámicas de conversación, la participación de usuarios y la influencia en la plataforma.
No. 30 mentions_screen_name. Menciona el nombre en la pantalla	El dato mentions_screen_name en Twitter muestra el nombre de pantalla único de un usuario mencionado en un tuit, indicado por "@" en el texto del tuit. Este valor es crucial para identificar qué usuario ha sido mencionado en el contenido del tuit y facilita el análisis de interacciones, relaciones entre usuarios, participación e influencia en la plataforma.
No. 31 Long. Largo	Twitter ha lanzado nuevas funciones para sus usuarios. "Notas de Twitter" permite crear contenidos largos de hasta 2.500 palabras con títulos, imágenes y medios incrustados. Esta función está disponible solo en la versión web y para usuarios seleccionados en Canadá, Ghana, Reino Unido y Estados Unidos. Además, para suscriptores de Blue, Twitter ha introducido tuits de hasta 10.000 caracteres y soporte para texto en negrita y cursiva, similar a plataformas de boletines rivales.
No. 32 quoted_status_id. Identificador de estado cotizado	Dato inventado por el usuario para un contexto particular.
No. 33 quoted_text. Texto citado	El dato quoted_text en Twitter representa el texto completo del tuit que ha sido citado por otro usuario. Esta variable almacena el contenido textual del tuit original al que se hace referencia en el tuit actual. El valor de quoted_text facilita el acceso al contenido completo del tuit original citado, sin necesidad de buscarlo directamente en la plataforma.

Continuará en la siguiente página.

Tabla A.1 – Continuación de la página anterior.

Descripción.	Detalles.
No. 34 quoted_created_at. Cita creada en.	El dato quoted_created_at en Twitter representa la fecha y hora en que se mencionó por primera vez a un usuario en un tuit. Esta variable almacena la marca de tiempo que indica cuándo ocurrió la mención del usuario en el tuit. El valor de quoted_created_at proporciona información crucial sobre la secuencia temporal de las menciones en el tuit, ayudando a entender el contexto y las interacciones. Este análisis permite examinar la cronología de las menciones a un usuario específico y comprender mejor las dinámicas de interacción y flujo de información en Twitter.
No. 35 quoted_source. Fuente citada	Dato inventado por el usuario para un contexto particular.
No. 36 quoted_favorite_count. Conteo de favoritos citados	El dato quoted_favorite_count en Twitter indica cuántas veces un tuit citado ha sido marcado como favorito por otros usuarios. Esta variable refleja el nivel de aprobación o interés que ha generado el tuit citado entre la comunidad de usuarios.
No. 37 quoted_retweet_count. Número de retweets citados	El dato quoted_retweet_count en Twitter representa cuántas veces un tuit citado ha sido retuiteado por otros usuarios. Esta variable indica la cantidad de veces que el tuit citado ha sido compartido, reflejando su nivel de difusión y resonancia entre la comunidad de usuarios.
No. 38 quoted_user_id. ID de usuario citado	El dato quoted_user_id en Twitter representa el identificador único del usuario que publicó el tuit original citado en otro tuit. Esta variable almacena el número único que identifica al usuario autor del tuit citado, permitiendo seguir sus interacciones y menciones específicas relacionadas con ese tuit.
No. 39 quoted_screen_name. Nombre de usuario citado	El dato quoted_screen_name en Twitter representa el nombre de pantalla del usuario que publicó el tuit original citado en otro tuit. Este nombre de pantalla es el identificador único del usuario precedido por "@". El valor de quoted_screen_name permite identificar al autor original del tuit citado y rastrear sus contribuciones en la conversación. El análisis de esta variable ayuda a explorar la relación entre el tuit citado y su autor, facilitando la comprensión de la influencia y el alcance del contenido citado en base al autor original del tuit.
No. 40 quoted_name. Nombre citado	Dato inventado por el usuario para un contexto particular.

Continuará en la siguiente página.

Tabla A.1 – Continuación de la página anterior.

Descripción.	Detalles.
No. 41 quoted_followers_count. Recuento de seguidores citados.	El dato quoted_followers_count en Twitter representa la cantidad de seguidores que tiene el autor del tuit original citado en otro tuit. Esta variable indica el tamaño de la audiencia del autor del tuit citado. El valor de quoted_followers_count es útil para evaluar el alcance potencial del contenido citado, ya que un mayor número de seguidores puede significar una audiencia más amplia y un mayor impacto en la difusión del mensaje. El análisis de quoted_followers_count permite entender el alcance y la relevancia del contenido citado en función del tamaño de la audiencia del autor original del tuit.
No. 42 quoted_friends_count. Cuenta de amigos citados.	El dato quoted_friends_count en Twitter representa la cantidad de usuarios que sigue el autor del tuit citado. Esta métrica indica las conexiones sociales del autor en la plataforma. El valor de quoted_friends_count proporciona información sobre la actividad social y las conexiones del autor original del tuit citado en Twitter. El análisis de esta métrica puede ayudar a entender la red social y las interacciones del autor, así como evaluar su grado de conexión con otros usuarios. Es importante señalar que el número de amigos (o usuarios seguidos) no determina por sí solo la calidad o relevancia del contenido citado.
No. 43 quoted_statuses_count. Cuenta de estados citados.	El dato quoted_statuses_count en Twitter representa el número de tuits publicados por el autor del tuit citado. Esta variable indica la actividad y participación del autor en la plataforma, ofreciendo una medida de su experiencia y grado de interacción en la red social.
No. 44 quoted_location. Ubicación citada.	El dato quoted_location en Twitter representa la ubicación indicada por el autor del tuit citado en su perfil. Esta variable almacena información sobre la ubicación declarada por el autor, que puede ser útil para entender el contexto geográfico del tuit o analizar patrones geográficos en las discusiones en Twitter. Es importante notar que la disponibilidad de datos en quoted_location puede variar, ya que la ubicación en los perfiles de Twitter es opcional y depende de la configuración individual de cada usuario.
No. 45 quoted_description. Descripción citada.	El dato quoted_description en Twitter representa la descripción o biografía proporcionada por el autor del tuit citado en su perfil. Esta variable almacena información sobre los intereses, ocupación o características relevantes del autor, ofreciendo detalles adicionales para entender su identidad y áreas de experiencia. El valor de quoted_description es útil para evaluar la credibilidad o relevancia del contenido citado, proporcionando insights sobre la identidad y contexto del autor original del tuit.

Continuará en la siguiente página.

Tabla A.1 – Continuación de la página anterior.

Descripción.	Detalles.
No. 46 quoted_verified. cita de cuenta verificada.	El dato quoted_verified en Twitter indica si la cuenta del autor del tuit citado está verificada. Esta variable almacena un valor booleano, siendo **TRUE** si la cuenta está verificada y **FALSE** si no lo está. Una cuenta verificada en Twitter ha sido confirmada como auténtica por la plataforma, especialmente para figuras públicas, celebridades y organizaciones importantes. La marca de verificación ayuda a identificar cuentas confiables y auténticas en la plataforma. El dato quoted_verified es útil para evaluar la credibilidad y autenticidad del autor del tuit citado, especialmente en contextos donde se comparte información crítica o relevante en Twitter.
No. 47 retweet_status_id. Id. de estado de retweet.	El dato retweet_status_id en Twitter representa el identificador único del tuit que ha sido retuiteado por otro usuario. Esta variable permite rastrear y vincular el retuit con el tuit original correspondiente, facilitando el análisis de la propagación y la viralidad del contenido en la plataforma. Seguir el retweet_status_id proporciona acceso al contenido original retuiteado y ayuda a entender su impacto en la comunidad de usuarios de Twitter.
No. 48 retweet_text. Retwittear texto.	Dato inventando por el usuario para un contexto particular. Lo más probable es que sea el texto del retweet seleccionado.
No. 49 retweet_created_at. Retweet creado.	El dato retweet_created_at en Twitter representa la fecha y hora en que se realizó un retuit. Esta marca de tiempo es útil para rastrear la secuencia temporal de los retuits y entender cómo se propaga el contenido en la plataforma. El análisis de retweet_created_at permite evaluar la influencia y resonancia de un tuit según el número y la distribución temporal de los retuits, proporcionando información sobre la difusión y el impacto del contenido en Twitter.
No. 50 retweet_source. Fuente de retweet.	El dato retweet_source en Twitter indica la fuente o la aplicación desde la cual se realizó un retuit. Esto proporciona información sobre cómo se distribuye y se comparte el contenido en la plataforma, incluyendo detalles sobre la aplicación oficial de Twitter, aplicaciones de terceros o clientes específicos utilizados para retuitear. El análisis de retweet_source puede revelar las preferencias de los usuarios y las diversas formas en que interactúan con el contenido, así como las fuentes más comunes de retuits y la distribución del contenido a través de diferentes aplicaciones o clientes de Twitter.
No. 51 retweet_favorite_count. Retweet favorito.	Dato inventado por el usuario para un contexto particular, en este caso sería un retweet favorito de otra cuenta.
No. 52 retweet_retweet_count. Recuento de retweets.	El dato retweet_retweet_count en Twitter indica cuántas veces un tuit ha sido retuiteado por otros usuarios en la plataforma. Este valor refleja la difusión y la resonancia del tuit entre los usuarios, siendo una métrica clave para evaluar su viralidad, alcance e influencia en Twitter.

Continuará en la siguiente página.

Tabla A.1 – Continuación de la página anterior.

Descripción.	Detalles.
No. 53 retweet_user_id. ID de usuario de retweet.	El dato retweet_user_id en Twitter representa el identificador único del usuario que ha realizado el retuit de un tuit específico. Este valor permite rastrear y vincular el retuit con el perfil de usuario correspondiente, facilitando el análisis de quiénes son los usuarios que han compartido y difundido un contenido particular en la plataforma.
No. 54 retweet_screen_name. Retuitear nombre de usuario.	El dato retweet_screen_name en Twitter representa el nombre de pantalla del usuario que ha realizado el retuit de un tuit específico. Este valor permite identificar al usuario que ha compartido el contenido original con sus seguidores, facilitando el análisis de las interacciones y el comportamiento de los usuarios en relación con ese tuit.
No. 55 retweet_name. Retwittear nombre.	Dato inventado por el usuario para un contexto particular.
No. 56 retweet_followers_count Recuento de seguidores de retweets.	El dato retweet_followers_count en Twitter representa la cantidad de seguidores que tiene el usuario que ha realizado el retuit de un tuit específico. Este valor es útil para evaluar la influencia y el alcance del usuario en la plataforma, ya que refleja la audiencia potencial que puede alcanzar a través de sus retuits.
No. 57 retweet_friends_count. Recuento de amigos de retweets.	El dato retweet_friends_count en Twitter representa la cantidad de usuarios que sigue el usuario que ha realizado el retuit. Esta métrica ofrece insights sobre las conexiones sociales del usuario en la plataforma, aunque no determina por sí sola la calidad o relevancia del retuit realizado.
No. 58 retweet_statuses_count. Recuento de estados de retweets.	El dato retweet_statuses_count en Twitter indica cuántos tuits ha publicado el usuario que ha realizado el retuit. Esto ofrece información sobre la actividad y la participación del usuario en la plataforma, aunque no determina por sí solo la calidad o relevancia del retuit realizado.
No. 59 retweet_location. Ubicación de retweet.	Dato inventado por el usuario para un contexto particular, para conocer la ubicación desde donde se retweeteo.
No. 60 retweet_description Descripción de retweet.	Dato inventado por el usuario para un contexto particular que puede ser una breve descripción del texto del retweet.
No. 61 retweet_verified. Retweet verificado.	El dato retweet_verified indica si la cuenta del usuario que realizó el retuit está verificada en Twitter. Esta verificación confirma la autenticidad de ciertas cuentas de interés público como celebridades o figuras públicas. Es útil para evaluar la credibilidad del usuario que realiza el retuit en el contexto del contenido compartido en Twitter.
No. 62 place_url. URL del lugar.	El dato place_url en Twitter almacena la URL asociada a un lugar geográfico mencionado en un tuit. Esta URL proporciona información adicional sobre el lugar, como su ubicación exacta, descripción y fotos. Es útil para comprender el contexto geográfico de los tuits y explorar más sobre los lugares mencionados en Twitter.

Continuará en la siguiente página.

Tabla A.1 – Continuación de la página anterior.

Descripción.	Detalles.
No. 63 place_name. Nombre del lugar.	El dato place_name en Twitter almacena el nombre del lugar geográfico mencionado en un tuit. Este valor proporciona información sobre la ubicación asociada al tuit, como una ciudad, país, región u otra descripción geográfica. Es útil para entender el contexto geográfico de los tuits y realizar análisis basados en la ubicación, identificando patrones de contenido y tendencias relacionadas con lugares específicos. Es importante tener en cuenta que no todos los tuits incluyen información de lugar, por lo que la disponibilidad de datos en la columna place_name puede variar.
No. 64 place_full_name. Nombre completo del lugar.	El dato place_full_name en Twitter almacena el nombre completo del lugar geográfico mencionado en un tuit. Este valor proporciona detalles detallados sobre la ubicación mencionada, como el nombre completo de la ciudad, estado o provincia, país y otros detalles geográficos relevantes. Analizar place_full_name es útil para etiquetar con precisión la ubicación asociada a un tuit específico, permitiendo análisis más detallados y contextuales basados en la ubicación geográfica. Es importante tener en cuenta que no todos los tuits mencionan un lugar y la disponibilidad de datos en la columna place_full_name puede variar.
No. 65 place_type. Tipo de lugar.	El dato place_type en Twitter indica el tipo de lugar geográfico mencionado en un tuit. Este valor categoriza el lugar mencionado en categorías como "city" (ciudad), "country" (país), "poi" (punto de interés), "neighborhood" (vecindario) u otros tipos de lugares geográficos. Analizar place_type es útil para etiquetar y categorizar la geolocalización asociada a un tuit, permitiendo análisis más precisos y contextualizados basados en la ubicación geográfica. Es importante destacar que no todos los tuits mencionan un lugar y la disponibilidad de datos en la columna place_type puede variar según la información proporcionada por los usuarios y la disponibilidad en la API de Twitter.
No. 66 Country. País.	El dato country en Twitter representa el país asociado a un lugar geográfico mencionado en un tuit. La variable country almacena el código ISO de dos letras del país según la norma ISO 3166-1 alpha-2, como "US" para Estados Unidos, "GB" para Reino Unido, "FR" para Francia, entre otros. El análisis de country permite identificar y etiquetar el país mencionado en un tuit, facilitando análisis basados en la ubicación geográfica y proporcionando insights sobre la distribución geográfica de los tuits. Es importante notar que no todos los tuits incluyen referencia a un país específico, y la disponibilidad de datos en la columna country puede variar según la información proporcionada por los usuarios y la API de Twitter.

Continuará en la siguiente página.

Tabla A.1 – Continuación de la página anterior.

Descripción.	Detalles.
No. 67 Country_code Código de país.	El dato country_code en Twitter representa el código alfanumérico (generalmente en formato ISO 3166-1 alpha-2) asociado a un lugar geográfico mencionado en un tuit, como "US" para Estados Unidos o "GB" para Reino Unido. Facilita el análisis geográfico de los tuits y proporciona información sobre la distribución geográfica del contenido. No todos los tuits incluyen este dato, y su disponibilidad puede variar según la información proporcionada por los usuarios y la API de Twitter.
No. 68 geo_coords. Coordenadas geográficas.	El dato geo_coords en Twitter proporciona las coordenadas geográficas (latitud y longitud) asociadas a un tuit, ofreciendo una ubicación precisa del lugar mencionado. Es esencial para realizar análisis geoespaciales, como visualización y mapeo en mapas, identificación de patrones espaciales, análisis de proximidad y cálculo de distancias. La disponibilidad de estos datos puede variar entre los tuits debido a configuraciones de privacidad y restricciones de la API de Twitter, lo que también puede afectar su precisión.
No. 69 coords_coords. Coordinadas.	El dato coords_coords en Twitter representa las coordenadas geográficas (latitud y longitud) asociadas a un tuit, proporcionando una ubicación precisa del lugar mencionado. Es crucial para el análisis geoespacial, permitiendo visualizar y mapear tuits, identificar patrones espaciales, realizar análisis de proximidad y calcular distancias. La disponibilidad y precisión de estos datos pueden variar según las configuraciones de privacidad de los usuarios y las limitaciones de la API de Twitter.
No. 70 bbox_coords. Coordenadas de bbox.	El dato bbox_coords en Twitter representa las coordenadas geográficas que definen un cuadro delimitador (bounding box) asociado a un lugar mencionado en un tuit. Este cuadro está definido por cuatro pares de coordenadas de latitud y longitud que encierran el área geográfica del lugar. Es esencial para visualizar y mapear tuits, identificar áreas geográficas específicas y realizar análisis espaciales. La disponibilidad y precisión de estos datos pueden variar según la configuración de privacidad y las restricciones de la API de Twitter.
No. 71 status_url. URL del estado.	El dato status_url se refiere a la URL directa a un tuit específico en Twitter. Esta variable almacena el enlace que lleva directamente al tuit correspondiente en la plataforma. La URL del tuit permite acceder directamente a ese contenido, compartirlo, referenciarlo o ver las interacciones asociadas. Es útil para analizar la difusión de un tuit, compartir contenido relevante o acceder a información específica en Twitter. Cada tuit tiene su propia URL única, y la columna status_url en los datos recopilados almacena estos enlaces para cada tuit.

Continuará en la siguiente página.

Tabla A.1 – Continuación de la página anterior.

Descripción.	Detalles.
No. 72 Name. Nombre.	El dato name en Twitter se refiere al nombre de usuario o cuenta asociada a un tuit específico. Esta variable almacena el nombre único o "handle" de la cuenta en Twitter que publicó el tuit. Por ejemplo, para la cuenta @OpenAI, el nombre almacenado en la columna name sería "OpenAI". El dato name es esencial para identificar la autoría del tuit y diferenciar entre diferentes usuarios en los datos de Twitter. Permite realizar análisis específicos sobre la actividad, el contenido y las interacciones de usuarios individuales en la plataforma. Es importante destacar que name se refiere al nombre de usuario en Twitter y no al nombre completo del usuario fuera de la plataforma.
No. 73 Location. Localización.	El dato location en Twitter se refiere a la ubicación geográfica asociada al perfil de un usuario. Esta variable almacena la descripción proporcionada por el usuario, que puede incluir detalles como una ciudad, estado, país u otra referencia geográfica. El dato location es útil para realizar análisis relacionados con la distribución geográfica de los usuarios en Twitter, identificar patrones regionales o geográficos en las interacciones y el contenido compartido. Es importante tener en cuenta que la precisión y la consistencia de los datos en la columna location dependen de la información proporcionada por los usuarios. Algunos usuarios pueden proporcionar ubicaciones precisas, mientras que otros pueden ingresar información más general o incluso datos falsos. Por lo tanto, el dato location debe ser interpretado con precaución y utilizado como un indicador general de la ubicación geográfica del usuario en Twitter.
No. 74 Descroption. Descripción.	El dato description en Twitter se refiere a la descripción o biografía del perfil de un usuario. Esta variable almacena la información descriptiva proporcionada por el usuario, que puede incluir detalles sobre su persona, intereses, profesión, aplicaciones u otra información relevante que el usuario decida compartir. La descripción en la columna description ofrece una visión breve del usuario y puede ser útil para entender mejor su identidad, intereses o área de especialización. Además, puede utilizarse para análisis de texto, identificación de tendencias o búsquedas basadas en palabras clave dentro de las descripciones de los usuarios. Es importante tener en cuenta que los usuarios tienen libertad para proporcionar cualquier descripción en su perfil, lo que puede variar desde información precisa hasta descripciones generales o creativas. Por lo tanto, la interpretación de la columna description debe considerarse en contexto y verificar su relevancia para el análisis específico que se esté realizando.
	Continuará en la siguiente página.

Tabla A.1 – Continuación de la página anterior.

Descripción.	Detalles.
No. 75 URL. URL.	El dato url en Twitter se refiere a la URL o enlace asociado al perfil de un usuario. La variable url almacena la dirección web proporcionada por el usuario en su perfil de Twitter, la cual puede dirigir a un sitio web personal, un blog, una página de empresa u otro recurso relevante relacionado con el usuario. Esta URL es útil para acceder directamente al recurso mencionado en el perfil del usuario, permitiendo obtener más información sobre él, su trabajo, su empresa o sus intereses. Es importante destacar que no todos los usuarios tienen una URL especificada en su perfil, por lo que la disponibilidad de datos en la columna url puede variar. Además, algunos usuarios podrían haber proporcionado una URL inválida, inexistente o dejar el campo en blanco.
No. 76 Protected. Protección.	El dato protected en Twitter indica si el perfil de un usuario está configurado como privado o protegido. La variable protected almacena un valor booleano donde: ***TRUE***: Indica que el perfil del usuario está configurado como privado o protegido. Esto significa que solo las personas aprobadas por el usuario pueden ver sus tuits, seguirlo y acceder a su contenido. Los tuits de un usuario con perfil protegido no son visibles para el público en general. ***FALSE***: Indica que el perfil del usuario no está protegido y es público. En este caso, cualquier persona puede ver los tuits, seguir al usuario y acceder a su contenido. El dato protected es útil para determinar la configuración de privacidad de un usuario en Twitter y puede influir en cómo se accede y se analiza su contenido.
No. 77 followers_count. Seguidores.	El dato followers_count en Twitter indica el número de seguidores de un usuario. Es esencial para medir su popularidad y alcance en la plataforma. Este número puede cambiar debido a ganancias y pérdidas de seguidores, y algunos perfiles privados no muestran esta información públicamente.
No. 78 Friends_count. Número de amigos.	El dato friends_count en Twitter indica cuántas cuentas sigue un usuario. Es crucial para entender las conexiones sociales y el interés del usuario por el contenido de otros usuarios. Este número puede variar con el tiempo debido a seguir o dejar de seguir cuentas, y algunos perfiles pueden tener esta información oculta si son privados.

Continuará en la siguiente página.

Tabla A.1 – Continuación de la página anterior.

Descripción.	Detalles.
No. 79 Listed_count. Recuento enumerado.	El dato listed_count en Twitter indica cuántas veces un usuario ha sido incluido en listas creadas por otros usuarios. Este número refleja su reconocimiento y relevancia percibida en la plataforma. La cantidad de listas puede cambiar con el tiempo debido a adiciones o eliminaciones por parte de otros usuarios, y la visibilidad de esta métrica puede variar según la configuración de privacidad de las listas creadas.
No. 80 Statuses_count. Conteo de estados.	El dato statuses_count en Twitter indica el número total de tuits publicados por un usuario. Este valor es crucial para evaluar la actividad y el compromiso del usuario en la plataforma, así como para identificar usuarios influyentes con un alto volumen de publicaciones. Es importante considerar que el número de tuits puede variar debido a nuevas publicaciones y eliminaciones de tuits anteriores por parte de los usuarios.
No. 81 Favourites_count. Conteo de Favoritos.	El dato favourites_count en Twitter indica el número total de tuits que un usuario ha marcado como favoritos. Este valor es útil para entender los intereses y preferencias del usuario, así como su nivel de participación y actividad en la plataforma. Es importante considerar que el número de tuits marcados como favoritos puede cambiar con el tiempo debido a las nuevas marcaciones o eliminaciones por parte de los usuarios, y que algunos perfiles pueden tener esta información oculta públicamente debido a configuraciones de privacidad.
No. 82 Account_created_at. Cuenta creada en.	El dato account_created_at en Twitter indica la fecha y hora en que se creó la cuenta de usuario, registrado en formato UTC (Coordinated Universal Time). Es útil para determinar la antigüedad de la cuenta, proporcionando contexto sobre su tiempo de existencia en la plataforma y potencial experiencia del usuario en Twitter. Este dato facilita el análisis temporal de la actividad del usuario, pero no revela detalles específicos sobre el contenido o la actividad de los tuits publicados.
No. 83 Verified. Verificado.	El dato verified en Twitter indica si una cuenta de usuario ha sido verificada por Twitter. Si el valor es **TRUE**, significa que la cuenta está verificada, confirmada como auténtica por Twitter mediante un distintivo azul. Esto es común para figuras públicas, marcas reconocidas o entidades de interés público. Si el valor es **FALSE**, la cuenta no está verificada. La verificación ayuda a identificar cuentas legítimas y confiables en la plataforma, aunque no todas las cuentas pueden ser verificadas y el proceso es selectivo por parte de Twitter.
Continuará en la siguiente página.	

Tabla A.1 – Continuación de la página anterior.

Descripción.	Detalles.
No. 84 Profile_url. URL de perfil.	El dato profile_url en Twitter es la dirección URL que lleva directamente al perfil de un usuario en la plataforma. Esta URL permite acceder a la biografía del usuario, sus tuits, seguidores, a quienes sigue y otra información relevante. Analizar el profile_url es útil para obtener detalles sobre el usuario, investigar su actividad, evaluar su influencia o acceder a su contenido completo. Cada usuario tiene una URL de perfil única, almacenada en la columna profile_url en los datos recopilados.
No. 85 profile_expanded_url. URL expandida de perfil.	Dato inventado por el usuario para un contexto particular.
No. 86 account_lang. Lengua de cuenta.	El dato account_lang en Twitter se refiere al idioma principal configurado en el perfil de un usuario. Almacena el código de dos letras del idioma según el estándar ISO 639-1. Este dato es útil para entender la preferencia lingüística del usuario en Twitter y puede ser utilizado para personalizar la experiencia de contenido, segmentar usuarios por idioma o analizar la diversidad lingüística de una comunidad. Es importante destacar que account_lang no indica necesariamente el idioma de los tuits individuales que el usuario pública.
No. 87 Profile_banner_url. URL de perfil ampliado.	El dato profile_banner_url en Twitter se refiere a la URL de la imagen de portada asociada al perfil de un usuario. Almacena la dirección URL de esta imagen, que se muestra en la parte superior del perfil del usuario en la plataforma. Esta imagen permite personalizar visualmente el perfil y puede proporcionar información adicional sobre el usuario, su marca o sus intereses. Es útil para acceder directamente a la imagen de portada para visualización o análisis, aunque no todos los usuarios tienen una imagen de portada configurada, lo que puede afectar la disponibilidad de datos en esta columna.
No. 88 profile_background_url. URL de fondo de perfil.	El dato profile_background_url en Twitter se refiere a la URL de la imagen de fondo asociada al perfil de un usuario. Almacena la dirección URL de esta imagen, que se muestra como fondo del perfil del usuario en la plataforma. Esta imagen permite personalizar visualmente el perfil y puede proporcionar información adicional sobre el usuario, su marca o sus intereses. Es útil para acceder directamente a la imagen de fondo para visualización o análisis, pero no todos los usuarios tienen una imagen de fondo configurada, lo que puede afectar la disponibilidad de datos en esta columna.

Continuará en la siguiente página.

Tabla A.1 – Continuación de la página anterior.

Descripción.	Detalles.
No. 89 Profile_image_url. URL de la imagen de perfil.	El dato profile_image_url en Twitter se refiere a la URL de la imagen de perfil asociada a la cuenta de usuario. Almacena la dirección URL de esta imagen, que representa generalmente la fotografía o avatar del usuario en su perfil. Este dato es útil para acceder directamente a la imagen de perfil, visualizar el aspecto visual del perfil del usuario y realizar análisis de imágenes si es necesario. Es importante considerar que la disponibilidad de datos en esta columna puede variar y algunas cuentas pueden no tener una imagen de perfil configurada.
No. 90 @. Arriba.	En el contexto de R, el símbolo @ se utiliza para acceder directamente a los componentes de un objeto o para obtener atributos específicos de un objeto de ciertas clases. Por ejemplo, si tienes un objeto llamado "obj" de una clase específica y deseas acceder a un atributo llamado "attr", puedes hacerlo utilizando "obj@attr". Es importante señalar que el uso de @ no es considerado una práctica recomendada en R, ya que puede comprometer los principios de encapsulamiento y acceso seguro a los objetos. En su lugar, se recomienda utilizar las funciones y métodos proporcionados por las clases para acceder y manipular los atributos de manera adecuada y segura. Si tienes un caso específico en mente donde necesites utilizar @ o si deseas más información sobre su uso en un contexto particular, estaré encantado de proporcionarte más detalles al respecto.

Tabla A.1: Descripción de las 91 variables generadas por cada tweet.

A.2. Terminología de la minería y análisis de datos.

Agrupación	Almacenar los datos de forma consecutiva, casi de la misma manera de la que se quiere acceder a ellos; así el acceso necesitará menos operaciones.
Agregación	Es un tipo de proceso de minería de datos e información en el que los datos son buscados, recopilados y presentados en un formato resumido basado en informes para lograr objetivos o procesos.
Algoritmo de aprendizaje automático	Un conjunto de instrucciones diseñado para mejorar el rendimiento de un modelo de aprendizaje automático a medida que se exponen a más datos.
Análisis de datos (Data Analysis)	Proceso de examinar, limpiar y transformar datos con el objetivo de descubrir información útil y tomar decisiones informadas.
Aprendizaje no supervisado	Modelo de aprendizaje automático que se utiliza cuando no hay etiquetas o resultados previos disponibles para la información de entrada.
Aprendizaje supervisado	Modelo de aprendizaje automático que se utiliza cuando se dispone de etiquetas o resultados previos para la información de entrada.
Árboles de decisión (Decision Trees)	Modelo de aprendizaje automático que utiliza una estructura de árbol para representar posibles resultados y decisiones basadas en condiciones y características de entrada.
Clustering	Proceso de agrupar objetos similares en conjuntos (clústeres) basándose en sus características.
Conjunto de datos	Conjunto de información utilizado para entrenar, validar o probar un modelo de aprendizaje automático.
Minería de datos (Data Mining)	Proceso de descubrir patrones y relaciones interesantes en grandes conjuntos de datos.
Modelado de datos (Data Modeling)	Proceso de diseñar la estructura de datos y sus relaciones en una base de datos.
Redes neuronales (Neural Networks)	Modelo de aprendizaje automático inspirado en la estructura del cerebro humano.
Redondeo	Es la operación o proceso a través del cual se modifica un número o dígito hasta que alcance un valor determinado de acuerdo a una serie de normas.
Regresión	Análisis estadístico que se utiliza para determinar la relación entre una variable dependiente y una o más variables independientes.
Selección de características	Proceso de identificar las características o atributos más relevantes para un modelo de aprendizaje automático.
Visualización de datos (Data Visualization)	Proceso de representar datos e información gráficamente para facilitar la comprensión y el análisis.

Tabla A.2: Terminología de la minería y Análisis de datos.

A.3. Terminología de Twitter.

Follower (seguidor)	Es el usuario que sigue la cuenta. Puede ver todos los tweets que se publican.
Following (siguiendo a)	Es la acción de seguir a otro usuario dentro de esta red.
Hashtag	El rey de Twitter. Se representa con un icono de almohadilla # y permite añadir tras él los términos que se quieran. Se utiliza para facilitar búsquedas. Por ejemplo, usando #WordPress en el buscador se encuentra un listado de los usuarios que han utilizado ese término en sus tweets.
IM (Instant Message) (Mensaje instantáneo)	Es un mensaje directo y privado que se envía a un usuario de Twitter.
Lista	Es un listado que puedes configurar con las cuentas favoritas. Se pueden crear la cantidad de listas que se quieran y otorgarles un nombre. Por ejemplo, se puede tener una lista relacionada con los gustos de cada persona, compras, empresas, deportes, etc.
Me gusta	Esta está representado por un icono de corazón. Se relaciona con un click si ha gustado un tweet.
Mención	Es una forma de mencionar a otro usuario en un tweet, utilizando el símbolo @ seguido de su nombre de usuario.
Retweet (RT)	Es la re-publicación de un tweet lanzado por otro usuario.
Time Line (Línea de tiempo)	Es la parte de la cuenta en la que se puede ver por orden cronológico los mensajes de los usuarios que sigues.
Trending topic	Son los temas más comentados del momento, es decir, las palabras con más menciones de la red social en un determinado periodo de tiempo.
Tweet (tuit)	Es cada uno de los mensajes que se publica. Hay que recordar que cada uno de ellos contiene hasta 280 caracteres (se amplió de 140 a 280 caracteres en 2018) sin contar el material multimedia que incluyas.
Twittero	Es cada usuario registrado. Se representa con @NombreDelUsuario.

Tabla A.3: Terminología empleada en Twitter.

A.4. Terminología de RStudio.

Biblioteca (Library)	Es el lugar donde se almacenan los paquetes instalados en RStudio.
Consola	Un espacio en RStudio donde se pueden ingresar comandos y ver los resultados de la ejecución del código.
Data frame	Es una estructura de datos en R que organiza los datos en filas y columnas, similar a una tabla en una base de datos.
Función (Function)	Conjunto de instrucciones que realiza una tarea específica en R.
Gráfico (plot, curve, etc.)	Es una representación visual de datos utilizando RStudio, generado por las funciones que grafican en R.
Knitr	Es un paquete de R que permite la creación de documentos dinámicos que combinan código, resultados y texto explicativo
Markdown	Es un lenguaje de marcado que se puede utilizar para crear documentos con formato en RStudio.
Matriz (Matrix)	Estructura de datos en R que organiza los datos en una tabla bidimensional de filas y columnas.
Paquete (Package)	Es una colección de funciones y datos en R que se pueden utilizar para realizar tareas específicas.
R	Es un lenguaje de programación estadístico y de análisis de datos. Script: un archivo que contiene código R que se puede ejecutar en RStudio.
RStudio	Es un entorno de desarrollo integrado (IDE) para el lenguaje de programación R.
Rtweet	La librería rtweet se especializa en la obtención de datos de Twitter y ofrece funciones para autentificarse en la API de Twitter, buscar y descargar tweets, y extraer información sobre usuarios y tendencias. También incluye herramientas para realizar análisis de redes sociales y visualizaciones interactivas.
Tidyverse	La librería tidyverse es una colección de paquetes de R que se centra en la manipulación y visualización de datos. Ofrece una serie de funciones que hacen que el proceso de limpieza y análisis de datos sea más fácil y eficiente. Entre sus paquetes se encuentran ggplot2 para visualización de datos, dplyr para manipulación de datos, y tidyr para dar formato a los datos.
Vector	Es una estructura de datos en R que almacena una colección ordenada de valores del mismo tipo.

Tabla A.4: Terminología empleada en RStudio.

Bibliografía

[1] Minería de datos para descubrir estilos de aprendizaje. (s. f.). CATEGORIZACIÓN. Recuperado 21 de diciembre de 2022, de https://rieoei.org/historico/deloslectores/1674Duran.pdf

[2] Castrillón, J., García, J., Anaya, M., Rodríguez, D., de la Rosa, D., y Caballero, C. (2008). Bases de datos, motores de búsqueda e índices temáticos: herramientas fundamentales para el ejercicio médico. Salud Uninorte, 24(1), 96–119. sitio web: http://www.scielo.org.co/scielo.php?script=sci_arttextpid=S0120-55522008000100011

[3] López, C. (2007). Minería de datos: técnicas y herramientas. Madrid: Editorial Paraninfo

[4] Vieira, L., Ortiz, L., Ramirez, S. (2009). Introducción a la Minería de Datos Rio de Janeiro: E-papers Servicios Editoriales.

[5] Greyrat, R. (2022, 5 julio). Transformación de datos en minería de datos – Barcelona Geeks. https://barcelonageeks.com/transformacion-de-datos-en-mineria-de-datos/

[6] Timarán-Pereira, S. R., Hernández-Arteaga, I., Caicedo-Zambrano, S. J., Hidalgo-Troya, A. y AlvaradoPérez, J. C., El proceso de descubrimiento de conocimiento en base de datos. (2016), El proceso de descubrimiento de conocimiento en base de datos. (2016). ediciones.ccc.edu.com. Recuperado 26 de diciembre de 2022, de https://ediciones.ucc.edu.co/index.php/ucc/catalog/download/36/40/230-1?inline=1

[7] Óscar P. M. [PDF] Minería de Datos. Abstract. Existencia de herramientas automáticas que no hacen necesario el ser un experto en estadística Potencia de computo - Free Download PDF. (s. f.-d). https://silo.tips/download/mineria-de-datos-abstract-existencia-de-herramientas-automaticas-que-no-hacen-ne

[8] RPubs - Reglas de asociación. (2020c, febrero 27). https://rpubs.com/Cristina_Gil/Reglas_Asociacion

[9] Cant, P. #. (n.d.). V# PROVEEDOR ESTADO CIUDAD VP v#. Wordpress.com. Retrieved January 3, 2023, from https://unefazuliasistemas.files.wordpress.com/2011/04/introducion-a-los-sistemas-de-bases-de-datos-cj-date.pdf

[10] Yoan, A., González, R., Francisco, J., Trinidad, M., Ariel, J., Ochoa, C., Shulcloper, J. R., Erro, L. E., Tonantzintla, S. M., Yoan Rodríguez González, A., Martínez Trinidad, J. F., Jesús, A. (n.d.). Minería de Reglas de Asociación sobre Datos Mezclados. Inaoep.Mx. Retrieved January 3, 2023, from https://ccc.inaoep.mx/portalfiles/file/CCC-09-001.pdf

[11] Sánchez Barrera, H. E. (2018). Clasificación del abecedario dactilológico mexicano utilizando minería de datos.

[12] M. Garre, J. J. Cuadrado, M. A. Sicilia, D. Rodríguez y R. Rejas, 110 Çomparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software,- evista Española de Innovación, Calidad e Ingeniería del Software, vol. 3, nº 1, pp. 6-22, 2007.

[13] Sairy, Ch., Edu.Ec. Retrieved January 3, 2023, from https://dspace.unl.edu.ec/Fernanda.pdf

[14] (Domínguez, 2001) Domínguez, E. (2001). "Regresión Logística, un ejemplo de su uso en endocrinología", Instituto Nacional de Endocrinología, Revista Cubana Endocrinol, Cuba.

[15] (Molinero, 2006) Vallejos, S. (2006). "Minería de Datos", Facultad de Ciencias Exactas, Naturales, y Agrimensura, Universidad Nacional del Nordeste.

[16] Bello, E. (2021). ¿Qué es el minado de Datos o Data Mininig? Técnicas y pasos a seguir. Thinking for Innovation. https://www.iebschool.com/blog/data-mining-mineria-datos-big-data/

[17] Rouse, M. (2019, 14 octubre). Sistema de gestión de bases de datos o DBMS. Computer-Weekly.es. https://www.computerweekly.com/es/definicion/Sistema-de-gestion-de-bases-de-datos-o-DBMS

[18] Ana V,. (2009). Retrieved January 5, 2023, from http://file:///C:/Users/pc/Downloads/-Ana_ValenciaAgudelo_2009.pdf

[19] Integración de Datos. (n.d.). Tecnologias-informacion.com. Retrieved January 6, 2023, from https://www.tecnologias-informacion.com/integracion.html

[20] What Is ETL? (2018, June 21). Sas.com. https://www.sas.com/es_mx/insights/data-management/what-is-etl.html

[21] Extracción, transformación y carga de datos (ETL). (n.d.). Microsoft.com. Retrieved January 6, 2023, from https://learn.microsoft.com/es-es/azure/architecture/data-guide/relational-data/etl

[22] Suárez, Y. R., Amador, A. D. (2009). Herramientas de minería de datos. Revista Cubana de Ciencias Informáticas, 3(3-4), 73-80.

[23] Morgado G., T., Ponce-de-León-Lima, D. A., Rosete-Suárez, A. (2017). Descubrimiento de conocimiento en bases de datos históricas de una empresa comercializadora. Ingeniería Industrial, 38(3), 289–297. http://scielo.sld.cu/scielo.php?script=sciarttextpid=S1815-59362017000300007

[24] Hernández O., J. (2004). Introducción a la Minería de Datos: Pearson

[25] Martínez, M. C. B. B. (n.d.). MINERÍA DE DATOS. Buap.Mx. Retrieved January 9, 2023, from http://bbeltran.cs.buap.mx/NotasMD.pdf

[26] josevallep. (2005, July 25). Bodega de datos (Data warehouse). Monografias.com. https://www.monografias.com/trabajos24/bodega-de-datos/bodega-de-datos

[27] Gráficos de distribución. (s. f.). © Copyright IBM Corp. 2014. https://www.ibm.com/docs/es/cognos-analytics/10.2.2?topic=SSEP7J_10.2.2/com

[28] García, M. M., Quintales, L. A. M., Peñalvo, F. J. G., Martín, M. J. P. (2002). Obtención y validación de modelos de estimación de software mediante técnicas de minería de datos. Revista colombiana de computación, 3(1), 53-71.

[29] Figuerola, C. G., Berrocal, J. L. A., Rodríguez, A. F. Z., Rodríguez, E., Reina, G. (2004). Algunas técnicas de clasificación automática de documentos. Cuadernos de documentación multimedia, ISSN-e, 1575-9733.

[30] Toro Ocampo, E. M., Molina Cabrera, A., Garcés Ruiz, A. (2006). Pronóstico de bolsa de valores empleando técnicas inteligentes. Tecnura, 9 (18), 57-66.

[31] TÉCNICAS DE PREDICCIÓN. (n.d.). Www.ub.edu. Retrieved January 29, 2023, from http://www.ub.edu/aplica_infor/spss/cap8-5.htm

[32] (N.d.). Retrieved January 29, 2023, from http://file:///C:/Users/pc/Downloads/Dialnet-ImplementacionDeLasTecnicasDePrediccionEnLaGenerac-7272003.pdf

[33] Carrasquilla, A., Chacón, A., Núñez, K., Gómez, O., Valverde, J., Guerrero, M. (2016). Regresión lineal simple y múltiple: aplicación en la predicción de variables naturales relacionadas con el crecimiento microalgal. Tecnología en Marcha, 29 (suppl. 5), 33-45. doi: https://dx.doi.org/10.18845/tm.v29i8.2983

[34] Gonzalez, L. (2020, 18 agosto). Aprendizaje no Supervisado. Aprende IA. https://aprendeia.com/aprendizaje-no-supervisado-machine-learning/

[35] Lin, G. y L. Chen. (2006). Identification of homogeneousregions for regional frequency analysis using the selforganizing map. Journal of Hydrology 324, pp 1–9.

[36] Barrios, A. F. y Carvajal, Y. (2006). Regionalización de índices de aridez y agresividad climática en Colombia utilizando análisis multi variado. Conformación estadística de una base de datos nacional homogénea.Tesis de grado. Universidad del Valle. Facultad de Ingeniería. Santiago de Cali

[37] Usar análisis de centralidad. (n.d.). Arcgis.com. Retrieved March 17, 2023, from https://pro.arcgis.com/es/pro-app/latest/help/analysis/link-charts/centrality.htm

[38] Continuos, A. (n.d.). Reglas de Asociación. Inaoep.Mx. Retrieved April 26, 2023, from http://ccc.inaoep.mx/ emorales/Cursos/NvoAprend/Acetatos/reglasAsociacion.pdf

[39] Álvarez Nuñez, M. F., y Parra Muñoz, J. A. (2013). Teoría de grafos.

[40] Wasserman, S. (1994). Social Network Analysis: Methods and applications. Vol 8.Cambridge University Press

89

[41] Iribarren, J. L., y Moro, E. (2011). Affinity Paths and information diffusion in social networks.Social Networks, 33(2), 134–142.

[42] Christakis, N. A., y Fowler, J. H. (2008). The collective dynamics of smoking in a large social network.New England journal of medicine, 358(21), 2249–2258.

[43] Siete puentes, un camino: Königsberg. (2004, febrero). Recuperado 15 de junio de 2023, de https://revistasuma.fespm.es/sites/revistasuma.fespm.es/IMG/pdf/45/069-078.pdf

[44] hasperue 2014 extraccion, Extracción de conocimiento en grandes bases de datos utilizando estrategias adaptativas, Hasperué, Waldo, 2014, Editorial de la Universidad Nacional de La Plata (EDULP)

[45] Comunicación /, C. A. (n.d.). Investigación y gestión de las redes digitales. Cuadernosartesanos.org. Retrieved September 12, 2023, from http://www.cuadernosartesanos.org/cac50.pdf#page=7

[46] Herrera, M., Wright, R., Abraham, E., Izquierdo, J., y Perez-Garcia, R. (2016). Condiciones hidráulicas sobre medidas de centralidad en grafos para la evaluación de la resiliencia de redes de distribución de agua. Acta Universitaria, 26, 82-90.

[47] Moronta, J., Rocco, C. M. (2016). Comparación de algoritmos de detección de comunidades en sistemas eléctricos de potencia. Revista de la Facultad de Ingeniería Universidad Central de Venezuela, 31(4), 51-60.

[48] Alvarez Nuñez, M. F. (2013). Teoría de grafos.

[49] Arenado Serrano, S. (2023). Analíticas de redes y grafos.

[50] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre, "Fast unfolding of community hierarchies in large networks," CoRR, 2008, [Online]. Available

[51] F. Bonomo, "introducción a la teoría de grafos." http://www-2.dc.uba.ar/personal/fbonomo/grafos/curso_grafos_handout080909.pdf. Visto en 09/05/2014.

[52] G. Martınez, "mejora de una metodología para la identificación de website keyobjects mediante la aplicación de tecnologías eye-tracking, análisis de dilatación pupilar y algoritmos de web mining," Universidad de Chile, 2014.

[53] Martínez Arbas, Á. (2016). Captura y transformación de grafos para análisis de redes sociales.

[54] Reguillo, R. (2018). Paisajes insurrectos. Jóvenes, redes y revueltas en el otoño civilizatorio. Cultura y Representaciones Sociales, 12(24), 433.

[55] Gillen, J. y Merchant, G. (2013). Twitter as a dialogic and linguistic practice. Language Sciences, 35, 47-58.

[56] Abascal, R. (2015). Comunicación política en 140 caracteres: el caso Ayotzinapa, en Razón y Palabra. 19 (92), 1-30.

[57] Sotelo, F., Fernandez, J., Jaramillo-Morillo, D., Urbano, F. A., Ordoñez, C. C. (2023). Desarrollo de una Aplicación para el Posicionamiento de Marcas en la Red Social Facebook a Través de la Minería de Datos. Revista Ibérica de Sistemas e Tecnologias de Informação, (E63), 288-299.

[58] Shmueli, Jason; Yeo, Jewelry; Tomé Damniel (2015). "Leveraging the power of a Twitter netweork for library promotion". https://doi.org/10.1016/j.acalib.2014.12.004

[59] Smith, Marc; Rainie, Lee; Shneiderman, Ben; Himelboim, Itai (2014). "Mapping Twitter topic networks: From polarized crowds to community clusters". Pew Research Center. Internet y technology, 20 Febr. http://www.pewinternet.org/2014/02/20/mapping-twittertopic-networks-from-polarized-crowds-to-community-clusters

[60] Cruces S., Graph and Network Analytics, Universidad de Sevilla, 2022.

[61] M. N. A. E. Hodler, A Comprehensive Guide to Graph Algorithms un Neo4j, neo4j, 2021.

[62] Carmona, O. G., Gárces, L. P. (2004). Inferencia estadística utilizando redes neuronales artificiales. Scientia et technica, 10(26), 31-36.

[63] Hodas, N., Kooti, F., Lerman, K. (2013). Friendship paradox redux: Your friends are more interesting than you. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 7, No. 1, pp. 225-233).