Ciencia de datos aplicada a un foro de discusión tecnológico

Julieth Tatiana García Zuluaga

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor
Javier Fernando Botia Valderrama, Doctor (PhD)

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Analítica y Ciencia de Datos
Medellín, Antioquia, Colombia
2024

Cita	(García Zuluaga, 2024)
Referencia **Estilo APA 7 (2020)**	García Zuluaga, J. T. (2024). *Inteligencia Artificial aplicada a un foro de discusión tecnológica* Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia.

Especialización en Analítica y Ciencia de Datos, Cohorte VI.

Centro de Investigación Ambientales y de Ingeniería (CIA).

Tabla de contenido

Lista de tablas

Lista de figuras

Siglas, acrónimos y abreviaturas

APA	American Psychological Association
Cms.	Centímetros
ERIC	Education Resources Information Center
Esp.	Especialista
MP	Magistrado Ponente
MSc	Magister Scientiae
Párr.	Párrafo
PhD	Philosophiae Doctor
PBQ-SF	Personality Belief Questionnaire Short Form
PostDoc	PostDoctor
UdeA	Universidad de Antioquia

Resumen

La información presente en los foros de discusión de la Vicepresidencia de Tecnología puede ser una fuente invaluable para la empresa, ofreciendo un entendimiento profundo de las necesidades y preferencias de los usuarios, así como identificando las fortalezas y debilidades de su conocimiento.

Este proyecto se enfoca en la aplicación de la inteligencia artificial para segmentar y clasificar a los usuarios que participan en los foros, con el objetivo de optimizar la interacción y personalización de contenido, mejorar la toma de decisiones basada en datos y aumentar la eficiencia en la gestión de las comunidades de usuarios.

Además, el análisis de sentimientos de las publicaciones del foro permitirá obtener información valiosa sobre la percepción de los usuarios hacia la tecnología y los servicios ofrecidos por la empresa. Esta información puede ser utilizada para identificar áreas de mejora, optimizar la comunicación y la gestión de las comunidades, y fortalecer la relación con los usuarios.

La segmentación y clasificación de los usuarios, junto con el análisis de sentimientos, permitirá a la empresa tomar decisiones más informadas, como la implementación de capacitaciones personalizadas, con el fin de mejorar las competencias necesarias y optimizar la gestión del conocimiento en la organización. Como resultado, se anticipa una mejora en la eficiencia y efectividad de los equipos, lo que se traducirá en un aumento de la productividad y la calidad de los proyectos desarrollados por la Vicepresidencia de Tecnología.

Palabras clave: Foros de discusión, Segmentación de usuarios, Análisis de sentimientos, Procesamiento de Lenguaje Natural.

El código fuente y los datos utilizados para este trabajo se encuentran disponibles en el siguiente repositorio de GitHub: https://github.com/TatianaGarcia1128/project_udea_2024

Abstract

The information present in discussion forums within the Technology Vice Presidency can be an invaluable resource for the company, providing a deep understanding of user needs and preferences, as well as identifying the strengths and weaknesses of their knowledge.

This project focuses on applying artificial intelligence to segment and classify users participating in these forums, with the aim of optimizing interaction and content personalization, improving data-driven decision-making, and increasing efficiency in managing user communities.

Furthermore, sentiment analysis of forum posts will provide valuable insights into user perceptions of technology and services offered by the company. This information can be used to identify areas for improvement, optimize communication and community management, and strengthen user relationships.

User segmentation and classification, combined with sentiment analysis, will allow the company to make more informed decisions, such as implementing personalized training programs to enhance necessary skills and optimize knowledge management within the organization. As a result, improvements are anticipated in team efficiency and effectiveness, translating into increased productivity and project quality within the Technology Vice Presidency.

Keywords: Discussion forums, User segmentation, Sentiment analysis, Natural Language Processing.

The source code and data used for this work are available in the following GitHub repository: https://github.com/TatianaGarcia1128/project_udea_2024

1. Descripción del problema

Las herramientas colaborativas son aplicaciones dinámicas que se caracterizan por tener comunidades donde el mayor énfasis se da a la contribución y participación de los usuarios; dentro de las herramientas existentes, los foros de discusión se destacan por ser generalmente utilizados para resolver inconvenientes que pudieran surgir en cualquier momento. Para poder acceder a la información almacenada en los foros de discusión, a menudo es necesario navegar por varios hilos hasta dar con una solución factible o con la solución adecuada, razón por la cual, se consideró la utilización de las características de calidad para evaluar las soluciones encontradas en diferentes foros de discusión [1].

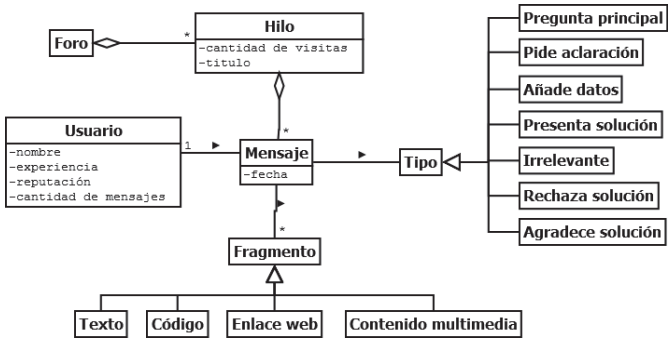

Figura 1:Modelo conceptual de foros de discusión técnicos

En la empresa se implementó un foro de discusión hace más de un año con el propósito de abordar inquietudes y problemas cotidianos de los ingenieros de software. Se desea aprovechar esta fuente de datos para respaldar la toma de decisiones estratégicas por parte de los líderes, particularmente en relación con la gestión del conocimiento y del talento en la organización, así como para maximizar el uso efectivo de la herramienta.

Estos datos ofrecen la posibilidad de obtener una comprensión profunda de las dificultades y desafíos específicos que enfrentan los equipos en el desarrollo de soluciones tecnológicas en diversas áreas de la Vicepresidencia de Tecnología. Al realizar un análisis detallado de estos datos, es factible identificar patrones, tendencias y áreas problemáticas específicas, proporcionando diagnósticos valiosos sobre el uso de la herramienta. Además, mediante el análisis de sentimientos de las publicaciones del foro, se puede obtener información sobre la percepción de los ingenieros de software hacia la herramienta, la calidad de las soluciones, y la efectividad de la colaboración.

Para abordar esta problemática, se llevará a cabo una segmentación de usuarios mediante algoritmos de agrupamiento de datos. Esta segmentación permitirá clasificar a los usuarios en grupos homogéneos basados en sus comportamientos, necesidades y niveles de participación. Además, se aplicarán algoritmos de clasificación para categorizar las discusiones y extraer temas recurrentes, lo que ayudará a los líderes a entender mejor las áreas de interés y preocupación de los ingenieros de software.

Con estas estrategias, se espera no solo optimizar el uso del foro de discusión, sino también impulsar iniciativas de formación y desarrollo profesional más efectivas, fomentar una colaboración más dirigida y mejorar la resolución de problemas dentro de la organización. En última instancia, el análisis avanzado de estos datos contribuirá a fortalecer la capacidad de innovación y la eficiencia operativa de la Vicepresidencia de Tecnología.

1.1. Problema de negocio

Aplicar la inteligencia artificial en el análisis de los datos provenientes de un foro de discusión mediante modelos de agrupamiento de datos, con el fin de potenciar la toma de decisiones estratégicas centradas en la gestión del conocimiento, la administración del talento y la optimización de la herramienta. Además, se utilizará el análisis de sentimientos para identificar las áreas de mejora y adaptar las estrategias de comunicación a las necesidades de la comunidad. Se busca aumentar la participación y retención de los usuarios del foro a través de la segmentación con el modelo de agrupamiento de datos seleccionado, obteniendo estos resultados en las siguientes métricas:

- **Tasa de participación (Engagement Rate):** Aumentar la tasa de participación del foro en un 20%; es decir, que el 20% más de los usuarios sean clasificados como “Participativos” por el modelo.
- **Tasa de interacción:** Aumentar la tasa de interacción promedio de los usuarios “Participativos” en el foro a un mínimo de 3 acciones por usuario.
- **Tasa de retención de usuarios activos:** Reducir la tasa de abandono de los usuarios “Participativos” en un 5%; es decir, que un 5% más de los usuarios “Participativos” iniciales sigan participando en el foro dentro de un umbral de retención de 0.5; donde el

umbral de retención representa una cantidad de tiempo específica en la que un usuario debe regresar al foro después de su última visita para ser considerado retenido.

- **Proporción de publicaciones con sentimientos neutros:** Aumentar en un 20% la proporción de publicaciones clasificadas como neutrales por el modelo de análisis de sentimiento.
- **Proporción de publicaciones con sentimientos positivos:** Mantener o aumentar ligeramente (5%) la proporción de publicaciones clasificadas como positivas por el modelo de análisis de sentimiento.

1.2. Aproximación desde la analítica de datos

A continuación para cada caso de aplicación desde la analítica se contemplan los siguientes aspectos:

Figura 2: Metodología segmentación de usuarios

Etapa de recopilación de información:

- Recolección de los datos del foro a partir de un backup de la base de datos del año 2022.
- Estudio de las tablas y campos que conforman la base de datos del foro con el objetivo de seleccionar la información más relevante para el estudio.

- Ejecución de estructura de datos de encolamientos en Postgresql y Python para la generación del conjunto de datos insumo de este estudio.

Etapa de preprocesamiento y limpieza de datos:

- Eliminación de variables innecesarias en el conjunto de datos debido a que se encuentran completamente nulas o no son relevantes para el estudio.
- Validación de los valores únicos en los campos del conjunto de datos para identificar aquellas variables que contienen solo un único valor, ya que estas no aportarían variabilidad al modelo y podrían ser eliminadas sin afectar el rendimiento del análisis.
- Creación de conjunto de datos agrupados para evitar redundancia.
- Limpieza de caracteres especiales en la variable de biografía del conjunto de datos, seguida de la creación de una bolsa de palabras, lematización, y extracción de las principales características. Finalmente, estas características se concatenarán al conjunto de datos original para su posterior análisis.
- Ajustar el tipo de datos de las columnas que lo requieran.
- Manejo de valores nulos mediante técnicas de imputación simple (valor constante) y mediante técnicas de imputación avanzada (KNN).
- Validación de registros duplicados.

Análisis exploratorio de datos:

- Para realizar un análisis más detallado de cada variable presente en el conjunto de datos, se realiza una separación de las columnas de acuerdo con su tipo.
- Gráficos temporales para las variables de tiempo con el objetivo de encontrar tendencias o patrones similares entre algunas de ellas.
- Se realizará un análisis de su frecuencia utilizando gráficos de barras y de torta para las variables categóricas. Estas visualizaciones proporcionarán una representación clara y concisa de la distribución de las categorías dentro de cada variable, lo que facilitará la identificación de patrones y tendencias relevantes en los datos.
- Visualización de la distribución de densidad de las variables numéricas y análisis de valores atípicos mediante gráficos de boxplot; este análisis proporcionará información sobre la

forma y la dispersión de los datos, permitiendo determinar la preparación adecuada de los mismos.

- Análisis de la correlación de las variables numéricas mediante la matriz de correlación.

Preparación del conjunto de datos:

- Para convertir las variables temporales a tipo entero mediante operaciones, se aplicarán transformaciones adecuadas que permitan expresar las fechas o intervalos de tiempo de manera numérica. Estas transformaciones facilitarán el análisis y modelado de los datos al representar el tiempo de una manera más manejable y compatible con técnicas de modelado numérico.
- Balanceo de las variables categóricas mediante la técnica de sobremuestreo por agregación que consiste en agrupar todas las categorías minoritarias en sola categoría para igualar la cantidad de muestras con la categoría predominante.
- Prueba de normalidad de Shapiro-Wilk sobre las variables numéricas.
- Detección de valores atípicos mediante el método de Z-Score modificado que se usa para datos asimétricos.
- Imputar los valores atípicos detectados con la media de los datos.
- Escalar el conjunto de datos; es decir, transformar los valores de las características a un rango específico de 0 y 1.
- Extraer una muestra de 100 registros con propósito de pruebas para elegir el mejor modelo.

Técnicas de reducción de la dimensionalidad

Evaluar diferentes técnicas de reducción de dimensionalidad con el objetivo de seleccionar aquella que mejor se adapte a la naturaleza de los datos. Las técnicas evaluadas son:

- Principal component analysis (PCA).
- Principal component analysis with kernel functions (Kernel PCA): es una extensión de PCA que permite aplicar PCA en espacios de características no lineales mediante el uso de kernels.
- Scattered Principal Component Analysis (Sparse PCA): es una variante de PCA que impone esparcimientos en los componentes principales lo que permite que solo un pequeño

subconjunto de las características contribuya significativamente a cada componente principal.

- t-distributed Stochastic Neighbor Embedding (t-SNE): es una técnica de reducción de dimensionalidad no lineal utilizada principalmente para visualizar conjuntos de datos de alta dimensionalidad en espacios de dimensiones bajas (generalmente 2D o 3D).

Modelos de Agrupamiento de Datos:

- Evaluar los siguientes algoritmos de agrupamiento de datos para segmentar usuarios con características similares basadas en sus interacciones en el foro:
    - Density-based spatial clustering of applications with noise (DBSCAN): Agrupamiento espacial basado en densidad de aplicaciones con ruido propuesto por Ester, Kriegel, Sander y Xu en 1996. Es el método de aprendizaje no supervisado más utilizado entre los métodos de agrupamiento basados en densidad. [3]

    - Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN): Significa Agrupamiento Espacial basada en la densidad jerárquica de aplicaciones con ruido. Utiliza una técnica de Agrupamiento Jerárquica para construir un árbol de grupos y, a continuación, selecciona los grupos más estables y persistentes en función de su densidad. HDBSCAN puede manejar ruido, valores atípicos y grupos de diferentes formas y tamaños. [4]

    - Spectral Clustering: hacen uso del espectro (valores propios) de la matriz de similitud de los datos para realizar reducción de dimensionalidad antes de la agrupamiento en un menor número de dimensiones. La matriz de similitud se proporciona como una entrada y consta de una evaluación cuantitativa de la similitud relativa de cada par de puntos en el conjunto de datos. [5]

    - Hierarchical Clustering: Los llamados métodos jerárquicos tienen por objetivo agrupar grupos para formar uno nuevo o bien separar alguno ya existente para dar origen a otros dos, de tal forma que, si sucesivamente se va efectuando este proceso de aglomeración o división, se minimice alguna distancia o bien se maximice alguna medida de similitud. [6]

- K-means: Es un algoritmo de agrupamiento no supervisada (clustering) que agrupa objetos en *k* grupos basándose en sus características. El agrupamiento se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo o clúster. Se suele usar la distancia cuadrática. [7]

- Seleccionar el mejor modelo de cada uno, de acuerdo con la combinación de hiperpárametros que proporciona los mejores resultados para los coeficientes de Silueta, Calinski-Harabasz y Davies-Bouldin, maximizando así la calidad y la coherencia de los modelos seleccionados.
- Seleccionar el mejor modelo de los elegidos anteriormente evaluando su comportamiento con la muestra de datos tomadas previamente.
- Aplicar el mejor modelo seleccionado previamente al conjunto de datos original, extraer el vector de características y llevar a cabo un análisis de patrones y grupos. Posteriormente, se procederá a realizar conclusiones sobre la estructura y la interpretación de los grupos identificados.

Figura 3: Metodología análisis de sentimientos

Etapa de recopilación de información: Se realiza el mismo proceso descrito para segmentación de usuarios.

Etapa de preprocesamiento y limpieza de datos:

- Extracción de la variable ‘excerpt’ la cual contiene las publicaciones que se analizarán.
- Creación de una bolsa de palabras (stopwords) en idioma español para limpieza de las publicaciones.
- Tokenización de las publicaciones en español usando la librería NLTK para análisis de sentimientos.
- Validación de registros duplicados.

Análisis exploratorio de datos:

- Visualización de las palabras más frecuentes en el foro mediante gráfico de barras horizontales, mapa de palabras y línea de frecuencia.

Preparación del conjunto de datos:

- Realizar un análisis de sentimiento en los textos de la columna 'excerpt', usando la librería VADER (Valence Aware Dictionary and sEntiment Reasoner) y luego clasifica cada texto como positivo, negativo o neutral. Vader viene de “Valence Aware Dictionary and sEntiment Reasoner” y es la librería que usa Python para el análisis de sentimientos.
- Cambiar los sentimientos por valores numéricos.
- Crear una representación numérica de los datos de texto (publicaciones).
- Dividir el conjunto de datos en entrenamiento y pruebas.

Modelos de clasificación:

- Evaluar los siguientes algoritmos de clasificación para el análisis de sentimientos en el foro de discusión:
    - Regresión logística: La regresión logística es una técnica estadística multivariante que nos permite estimar la relación existente entre una variable dependiente no

métrica, en particular dicotómica y un conjunto de variables independientes métricas o no métricas. [8]

- Árbol de decisión: Los **árboles de decisión son** algoritmos estadísticos o técnicas de *machine learning* que nos permiten la construcción de modelos predictivos de analítica de datos para el Big Data basados en su clasificación según ciertas características o propiedades, o en la regresión mediante la relación entre distintas variables para predecir el valor de otra. [9]

- Support Vector Machine (SVM): El objetivo del algoritmo SVM es encontrar un hiperplano que separe de la mejor forma posible dos clases diferentes de puntos de datos. [10]

- Evaluar el rendimiento de los modelos utilizando métricas de desempeño, la matriz de confusión y las curvas ROC para seleccionar el modelo óptimo.

1.3.Origen de los datos

El foro de discusión se ha desarrollado utilizando una herramienta basada en un proyecto de código abierto y emplea una base de datos Postgresql para almacenar de manera persistente la información. Este foro se encuentra alojado en la nube de AWS, específicamente en el componente RDS Auto Scaling Group, junto con un Bucket S3.

Es importante destacar que el acceso a la base de datos no está disponible para el público en general. Para obtener copias de seguridad de la información almacenada o para llevar a cabo implementaciones específicas en la nube donde reside la base de datos, es necesario solicitar acceso a los administradores de bases de datos (DBAs).

1.4. Métricas de desempeño

Las métricas de machine learning que se usarán para la segmentación de usuarios son [11]:

- Coeficiente de Silueta: combina cohesión y separación para evaluar la calidad de los conglomerados. Mide que tan bien encaja cada puntos de datos dentro de su grupo asignado en comparación con otros grupos. El coeficiente varía de -1 a 1, donde un valor cercano a 1 indica que las muestras están bien agrupadas, mientras que un valor cercano a -1 sugiere que las muestras pueden haber sido asignadas al grupo incorrecto. [12]
- Índice de Calinski-Harabasz (CH): Evalúa la calidad de los conglomerados basándose tanto en la dispersión dentro del conglomerado como en la dispersión entre conglomerados. Un valor más alto del índice Calinski-Harabasz indica grupos más compactos y bien separados. [12]
- Índice de Davies-Bouldin (DB): Mide la similitud promedio entre cada grupo y su grupos más similar, al mismo tiempo que considera la disimilitud promedio entre cada grupo y su grupo menor similar. Se pretende minimizar este índice. [12]

Las métricas de negocio que se usarán para la segmentación de usuarios son: Nivel de participación y actividad de los usuarios en cada segmento.

Las métricas de machine learning que se usarán para el análisis de sentimientos son [13]:

- Exactitud: Esta métrica calcula la proporción de predicciones correctas realizadas por el modelo en relación con el número total de predicciones. Se expresa como un valor entre 0 y 1, donde 1 representa una precisión perfecta. [13]
- Precisión y Recall: La precisión mide la proporción de verdaderos positivos (predicciones correctas) en relación con el total de predicciones positivas realizadas por el modelo; el recall, por otro lado, mide la proporción de verdaderos positivos en relación con el total de ejemplos positivos presentes en los datos de prueba. [13]
- F1-score: Es una métrica que combina la precisión y el recall en un solo valor, proporcionando una medida general del rendimiento del modelo [13]

2. Objetivos

2.1.Objetivo general

Desarrollar una solución para mejorar la participación en el foro de discusión de la Vicepresidencia de Tecnología mediante la segmentación de usuarios y el análisis de sentimientos, utilizando algoritmos de agrupamiento de datos para identificar los usuarios en función de su actividad, interacción y otros indicadores relevantes.

2.2.Objetivos específicos

- Implementar un proceso de exploración de datos por medio de un análisis de datos cuantitativo y/o cualitativo, transformaciones de los datos basados en escalamiento y/o reducción de dimensionalidad, así como el tratamiento de datos atípicos pertinentes.
- Proponer una estrategia para la búsqueda del mejor modelo de agrupamiento de datos que permita encontrar una segmentación que se ajuste al contexto de los usuarios que participan en el foro de discusión de la Vicepresidencia de Tecnología.
- Evaluar el desempeño de los modelos de análisis de sentimientos, incluyendo árboles de decisión, regresión logística y SVM, para la clasificación de sentimientos utilizando métricas como la precisión, exactitud, sensibilidad, especificidad y F1-Score. Los resultados de este análisis se utilizarán para comparar los modelos y seleccionar el modelo con mejor desempeño, con el objetivo de comprender mejor las emociones de los usuarios.

3. Datos

3.1. Datos originales

Los datos se recopilaron en un inicio haciendo una restauración local de un backup proporcionado por el DBA de la empresa que se encarga de administrar la base de datos del foro de discusión a mayo de 2023. Del total de 114 tablas disponibles en la base de datos restaurada se seleccionaron 13, consideradas como las que contienen información relevante para la ejecución de los dos objetivos propuestos. Posterior a esta selección de tablas y campos se crea un query el cual es ejecutado en un script de python para exportar el resultado de las consultas en un archivo .csv.

Conjunto de datos para el objetivo de segmentación de usuarios

El conjunto de datos que será usado presenta las siguientes características:

	id	approved	creation_user	date_of_birth	first_seen_at	flag_level	group_locked_trust_level	last_posted_at	last_seen_at	manual_locked_trust_level	...	location	views_profile	mobile	posts_read	time_read_visit
0	2	True	2022-07-08	NaN	2022-07-08	0	NaN	2022-07-15	2022-07-28	3.0	...	NaN	42	False	0	0
1	2	True	2022-07-08	NaN	2022-07-08	0	NaN	2022-07-15	2022-07-28	3.0	...	NaN	42	False	0	0
2	2	True	2022-07-08	NaN	2022-07-08	0	NaN	2022-07-15	2022-07-28	3.0	...	NaN	42	False	0	0
3	2	True	2022-07-08	NaN	2022-07-08	0	NaN	2022-07-15	2022-07-28	3.0	...	NaN	42	False	0	0
4	2	True	2022-07-08	NaN	2022-07-08	0	NaN	2022-07-15	2022-07-28	3.0	...	NaN	42	False	0	0

5 rows × 40 columns

Figura 4: Muestra conjunto de datos

A continuación, se describe detalladamente cada una de las columnas del conjunto de datos:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 68386 entries, 0 to 68385
Data columns (total 40 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   id                          68386 non-null  int64
 1   approved                    68386 non-null  bool
 2   creation_user               68386 non-null  object
 3   date_of_birth               0 non-null      float64
 4   first_seen_at               68386 non-null  object
 5   flag_level                  68386 non-null  int64
 6   group_locked_trust_level    0 non-null      float64
 7   last_posted_at              61683 non-null  object
 8   last_seen_at                68386 non-null  object
 9   manual_locked_trust_level   15017 non-null  float64
 10  moderator                   68386 non-null  bool
 11  previous_visit_at           68152 non-null  object
 12  silenced_till               0 non-null      float64
 13  trust_level                 68386 non-null  int64
 14  views                       68386 non-null  int64
 15  days_visited                68386 non-null  int64
 16  distinct_badge_count        68386 non-null  int64
 17  first_post_created_at       62287 non-null  object
```

```
 18  flags_agreed               68386 non-null  int64
 19  flags_disagreed            68386 non-null  int64
 20  flags_ignored              68386 non-null  int64
 21  likes_given                68386 non-null  int64
 22  likes_received             68386 non-null  int64
 23  post_count                 68386 non-null  int64
 24  posts_read_count           68386 non-null  int64
 25  time_read                  68386 non-null  int64
 26  topic_count                68386 non-null  int64
 27  topics_entered             68386 non-null  int64
 28  badge_granted_title        68386 non-null  bool
 29  bio_raw                    8872 non-null   object
 30  location                   5654 non-null   object
 31  views_profile              68386 non-null  int64
 32  mobile                     68386 non-null  bool
 33  posts_read                 68386 non-null  int64
 34  time_read_visit            68386 non-null  int64
 35  group_name                 0 non-null      float64
 36  user_count_group           0 non-null      float64
 37  visibility_level           0 non-null      float64
 38  badge_name                 65211 non-null  object
 39  badge_type_name            65211 non-null  object
dtypes: bool(4), float64(7), int64(19), object(10)
                                memory usage: 19.0+ MB
```

```
--------------------------------------
Forma del dataset:
--------------------------------------
(68386, 40)
--------------------------------------
```

El conjunto de datos contiene 68386 muestras y 40 columnas, de las cuales 19 tienen valores tipo entero (int64), 7 de tipo decimal (float64), 4 de tipo booleano (bool) y las 10 restantes son tipo object. De acuerdo la base de datos ocupa 19.0+ MB, y es fácil de procesar.

Al realizar una exploración descriptiva de los datos se obtiene lo siguiente:

	id	date_of_birth	flag_level	group_locked_trust_level	manual_locked_trust_level	silenced_till	trust_level	views	days_visited	distinct_badge_count	...	posts_read_count	time_read
count	68386.000000	0.0	68386.0	0.0	15017.0	0.0	68386.000000	68386.0	68386.000000	68386.000000	...	68386.000000	68386.000000
mean	461.151230	NaN	0.0	NaN	3.0	NaN	1.892668	0.0	78.110958	8.157810	...	522.212236	26276.610330
std	478.208115	NaN	0.0	NaN	0.0	NaN	1.193106	0.0	54.997465	5.570794	...	632.162534	30247.611403
min	2.000000	NaN	0.0	NaN	3.0	NaN	0.000000	0.0	1.000000	0.000000	...	0.000000	0.000000
25%	97.000000	NaN	0.0	NaN	3.0	NaN	1.000000	0.0	30.000000	4.000000	...	86.000000	4609.000000
50%	281.000000	NaN	0.0	NaN	3.0	NaN	2.000000	0.0	67.000000	7.000000	...	218.000000	12175.000000
75%	647.000000	NaN	0.0	NaN	3.0	NaN	3.000000	0.0	121.000000	12.000000	...	759.000000	36742.000000
max	2204.000000	NaN	0.0	NaN	3.0	NaN	4.000000	0.0	199.000000	18.000000	...	3022.000000	102488.000000

8 rows × 26 columns

Figura 4: Información descriptiva del conjunto de datos

Conjunto de datos para el objetivo de análisis de sentimientos

El conjunto de datos que será usado para el análisis de sentimientos posterior a la extracción de la variable objetivo ‘excerpt’ que contiene las publicaciones en el foro tiene la siguiente estructura:

	id	excerpt
0	2	Actualmente estoy utilizando la version 1.6.6 ...
1	2	Actualmente estoy utilizando la version 1.6.6 ...
2	2	Actualmente estoy utilizando la version 1.6.6 ...
3	2	Actualmente estoy utilizando la version 1.6.6 ...
4	2	Actualmente estoy utilizando la version 1.6.6 ...
...	...	...
2091663	2190	se están generando vulnerabilidades de tipo Vu...
2091664	2190	se están generando vulnerabilidades de tipo Vu...
2091665	2190	se están generando vulnerabilidades de tipo Vu...
2091666	2190	se están generando vulnerabilidades de tipo Vu...
2091667	2197	Como puedo ocultar variables en releases anter...

2091668 rows × 2 columns

Figura 5: Estructura de datos para el análisis de sentimientos

3.2. Datasets

Sobre el conjunto de datos se realizan las siguientes acciones de preprocesamiento de datos:

Eliminación de variables irrelevantes: Se eliminarán las variables que contengan únicamente valores nulos, identificadas previamente al consultar la información detallada del conjunto de datos.

Extracción columna de publicaciones: Se selecciona la columna ‘excerpt’ del conjunto de datos la cual contiene las publicaciones dentro del foro de discusión. Adicionalmente, se eliminan los duplicados de estas publicaciones.

Gestión de duplicados: Se validarán valores duplicados para unificar palabras con diferente ortografía pero significado similar. Además, se identificarán y eliminarán variables con un único valor en todas sus muestras.

Creación del conjunto de datos agrupados: El conjunto de datos contiene registros duplicados de usuarios debido a que cada usuario puede recibir múltiples medallas a lo largo de su participación en el foro. Para evitar que estos duplicados distorsionen el análisis, se realiza una agrupación de los registros por ID de usuario. Esta acción permite:

- **Conservar el conteo real de usuarios:** Se mantiene la cantidad precisa de usuarios en el foro sin duplicaciones.

- **Mostrar la última medalla obtenida:** Se conserva la información de la última medalla recibida por cada usuario, reflejando su último logro en el foro.

Esta técnica de agrupación por ID de usuario asegura un análisis preciso y completo de los usuarios del foro, evitando la distorsión de los resultados por la presencia de registros duplicados.

La nueva forma del conjunto de datos, posterior a la agrupación es:

```
----------------------------------------
Forma del dataset:
----------------------------------------
(1857, 28)
----------------------------------------
```

Extracción de características: Se extraerán las principales características de la variable 'bio_raw' mediante técnicas de eliminación de palabras vacías (stopwords) y lematización, con el objetivo de limpiar los datos.

Se realiza la extracción de términos que contiene los valores de la matriz TF-IDF para cada término en cada documento y se obtienen las siguientes dos palabras más relevantes en la biografía de los usuarios que participan en el foro:

```
<class 'pandas.core.frame.DataFrame'>
Index: 1857 entries, 2 to 2204
Data columns (total 2 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   desarrollador  1857 non-null   float64
 1   software       1857 non-null   float64
dtypes: float64(2)
memory usage: 43.5 KB
```

Figura 6: Palabras relevantes en bio_raw

Finalmente, estas dos nuevas columnas son concatenadas al conjunto de datos inicial.

Adicional, se realiza la limpieza de caracteres especiales de la variable ‘excerpt’; así como la eliminación de stopwords y la tokenización de las palabras:

	id	excerpt	token
0	2	actualmente estoy utilizando la version 166 y ...	[actualmente, utilizando, version, 166, genera...
1	2	estamos en una iniciativa que necesitamos real...	[iniciativa, necesitamos, realizar, envio, arc...
2	2	se tiene un proyecto que es un portal que func...	[proyecto, portal, funciona, autogestion, tran...
3	2	se presenta el siguiente error al realizar al ...	[presenta, siguiente, error, realizar, conexio...
4	2	como manejar el tema de variables en los archi...	[manejar, tema, variables, archivos, config, c...
...	...	...	...
2669	2139	hola comunidad quisiera preguntar si alguno co...	[hola, comunidad, quisiera, preguntar, alguno,...
2670	2146	hola quisiera saber cuales son los valores rec...	[hola, quisiera, saber, cuales, valores, recom...
2671	2167	buenas tardes a todos pregunta si al aumentar ...	[buenas, tardes, pregunta, aumentar, memoria, ...
2672	2190	se estan generando vulnerabilidades de tipo vu...	[estan, generando, vulnerabilidades, tipo, vul...
2673	2197	como puedo ocultar variables en releases anter...	[puedo, ocultar, variables, releases, anterior...

2662 rows × 3 columns

Figura 7: Tokenización y limpieza 'excerpt'

Ajuste de tipos de datos: Se ajustarán los tipos de datos para facilitar la exploración y análisis de la información. Por lo que se ajusta el tipo de dato de las variables de tiempo y de las variables categóricas presentes en el conjunto de datos para mejorar su análisis más adelante.

Imputaciones de valores nulos: Se realizan diferentes imputaciones para las columnas que presentan valores nulos; teniendo en cuenta el tipo de dato de cada una de ellas.

Para las variables de tiempo que presentan valores nulos, se consulta cuánto porcentaje de estos valores hay presente en la muestras, con el objetivo de aplicar el mejor método de imputación, y se obtiene que los valores nulos representan más del 10% para cada una de las variables.

Como el porcentaje de valores nulos es mayor al 10% del total de datos en las variables de tiempo, se realiza la imputación variable de los datos mediante el método de vecinos cercanos; el cual encuentra las fechas más similares a la fecha faltante y utiliza su valor para llenar el campo faltante.

Para la variable ‘manual_locked_trust_level’ que presenta valores nulos, se revisa que valores únicos contiene (3 y NaN), siendo 3 como el nivel de confianza alto; por lo que se decide imputar los valores nulos con el valor de 1, para indicar un nivel de confianza más bajo. Adicional, se ajusta el tipo de dato de la variable a ‘object’.

Para las variables categóricas 'location', 'badge_name' y 'badge_type_name' que presentan valores nulos, se lleva el valor ‘sin información’ para imputar estos valores.

Validación final de duplicados: Se realizará una última validación de duplicados tras la aplicación de las transformaciones previas.

Al realizar la consulta de valores duplicados, se obtienen los siguientes registros:

```
----------------------------------------
Registros duplicados
----------------------------------------
```

id	developer	software	creation_user	first_seen_at	last_posted_at	last_seen_at	manual_locked_trust_level	moderator	previous_visit_at	trust_level	...	topic_count	topics_entered	views_profile
1595	0.0	0.0	2022-11-02	2022-11-02	2022-12-02 09:36:00.000000129	2022-11-02	1.0	False	2022-11-23 09:36:00.000000129	0	...	0	1	0
1869	0.0	0.0	2023-01-11	2023-01-11	2022-12-02 09:36:00.000000129	2023-01-11	1.0	False	2022-11-23 09:36:00.000000129	0	...	0	1	0
2186	0.0	0.0	2023-03-22	2023-03-22	2022-12-02 09:36:00.000000129	2023-03-22	1.0	False	2022-11-23 09:36:00.000000129	0	...	0	1	0

3 rows × 29 columns

Figura 8: Registros duplicados

Se conservan estos registros dado que presentan diferente ‘id’; por lo que corresponden a diferentes usuarios que coinciden en sus características.

3.3.Analítica descriptiva

En esta sección, se explora en profundidad el conjunto de datos de usuarios del foro, con el objetivo de obtener información valiosa y revelar patrones interesantes que puedan ayudar en la comprensión del comportamiento de los usuarios. Se utilizan diversas técnicas de visualización y estadísticas descriptivas para analizar las características de los usuarios y sus interacciones en el foro.

El análisis inicia con la partición de las variables del conjunto de datos en 3 listas diferentes de acuerdo con el tipo de dato:

- Lista de variables categóricas
- Lista de variables de tiempo
- Lista de variables numéricas

Una vez se definen las anteriores listas de variables se realiza un análisis exploratorio sobre cada una de ellas.

Análisis exploratorio de variables de tiempo

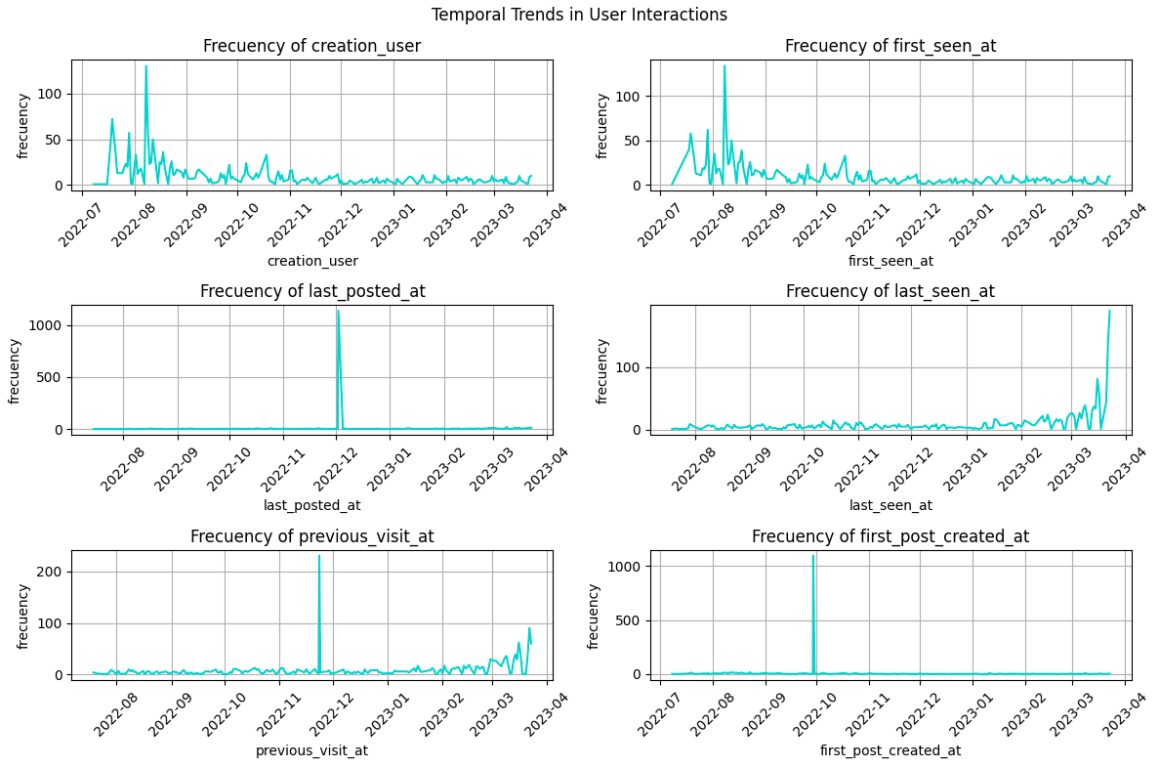

Figura 9: Gráficos temporales de las interacciones de los usuarios

Los gráficos anteriores para las variables de tipo 'DateTime' evidencian que:

- **Alta actividad de registro de usuarios:** Se observa un pico de registros entre julio y septiembre de 2022, coincidiendo con el posible lanzamiento del foro.
- **Correlación entre registro y primera visita:** La gran similitud entre las variables 'creation_user' y 'first_seen_at' sugiere una alta correlación, por lo que una de ellas podría eliminarse en análisis futuros.
- **Escasa actividad de publicaciones:** Se registran pocas publicaciones en general, excepto por un pico muy alto en enero de 2023 que requiere investigación.
- **Aumento de la última aparición de usuarios:** El gráfico de 'last_seen_at' indica una buena acogida del foro por parte de la comunidad.
- **Compromiso creciente de los usuarios:** El gráfico 'previoud_visit_at' muestra un incremento en las visitas previas, reflejando un mayor compromiso y lealtad de los usuarios. Se observa un pico en febrero que necesita ser analizado.

- **Fecha de la primera publicación:** El gráfico 'first_post_created_at' indica que la primera publicación se creó en octubre de 2022.

Análisis exploratorio de variables categóricas:

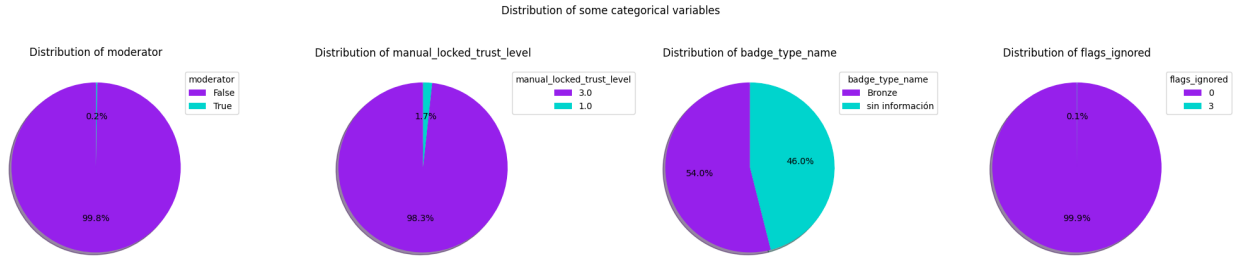

Figura 10: Gráficos de torta de algunas variables categóricas

Para las tres variables categóricas representadas en los gráficos, se observan desbalances significativos:

- **Desbalance en usuarios moderadores:** La proporción abrumadora de usuarios no moderadores (99.8%) frente a moderadores (0.2%) podría dificultar la identificación de patrones relacionados con la moderación. Se recomienda considerar la eliminación de esta variable.
- **Desbalance en el nivel de confianza:** La presencia de un pequeño grupo de usuarios (1.7%) con un nivel de confianza de 0 podría introducir sesgos en el análisis. Se recomienda considerar la eliminación de esta variable.
- **Baja variabilidad en 'flags_ignored':** La variable 'flags_ignored' presenta una distribución muy desbalanceada, con poca variabilidad y poca información significativa. Se recomienda eliminarla del conjunto de datos.
- **Información incompleta sobre las medallas:** La falta de información sobre el tipo de medallas para la mayoría de los usuarios (46%) limita la comprensión completa de la distribución de medallas en el foro.

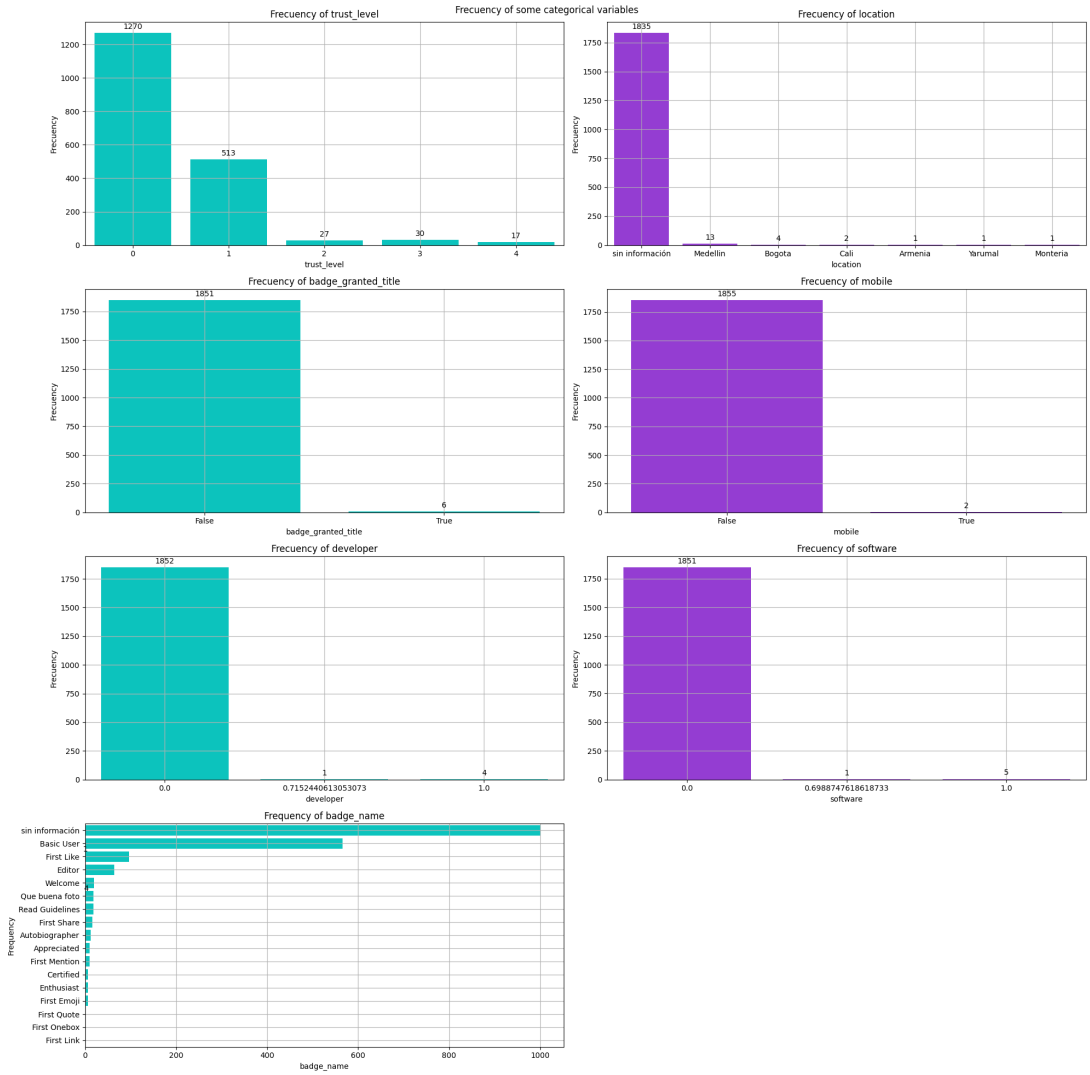

Figura 11: Gráficos de barras de algunas variables categóricas

Los gráficos revelan un marcado desequilibrio en todas las variables categóricas representadas. Se recomienda eliminar las variables 'location', 'badge_granted_title' 'mobile', 'developer' y 'software' debido a que presentan una diferencia significativa, lo que hace imposible equilibrar sus datos. En cuanto a las variables 'badge_name' y 'trust_level', se podría lograr un balanceo agrupando algunas categorías.

Análisis exploratorio de variables numéricas

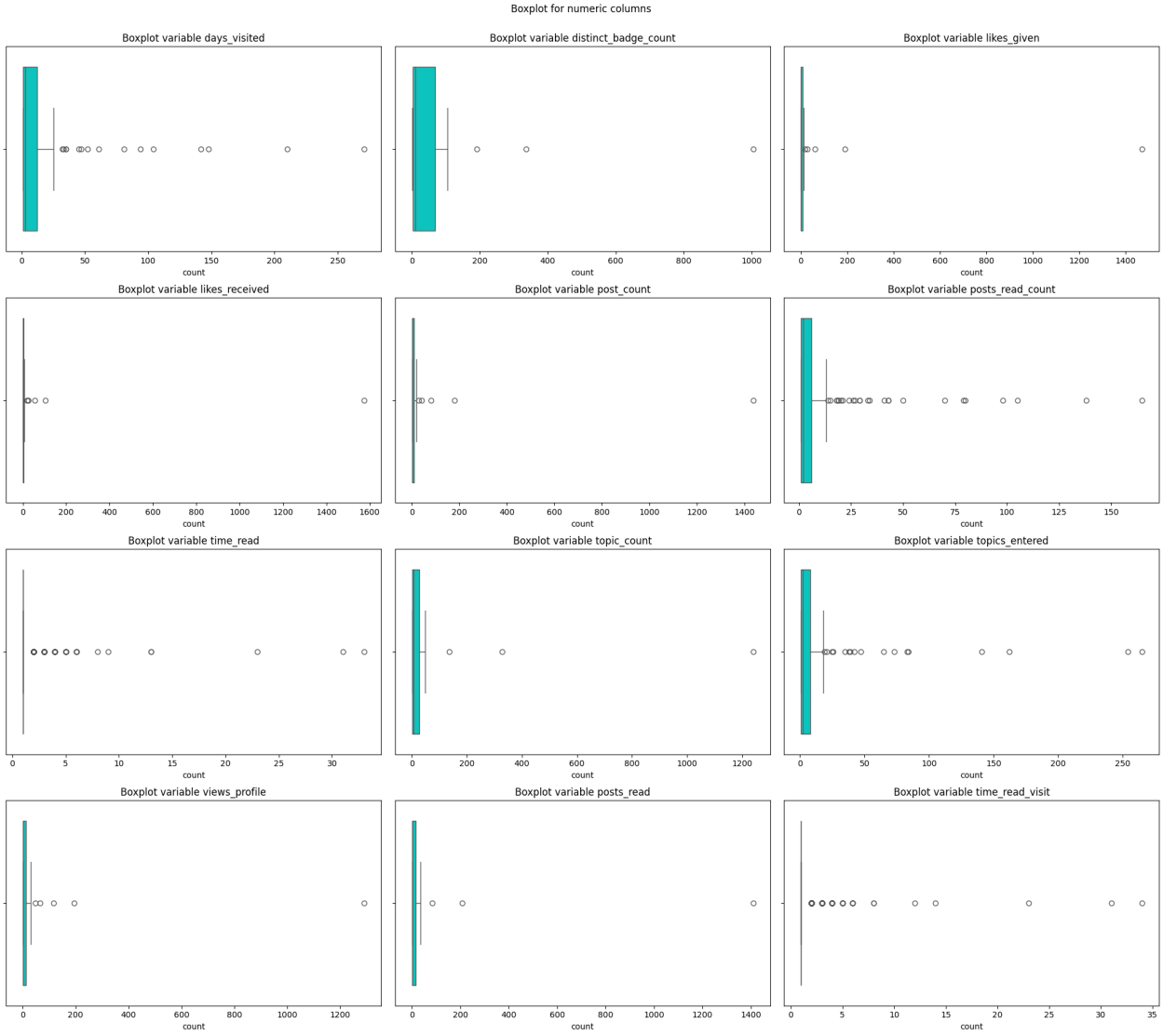

Figura 12: Cajas de bigote de variables numéricas

En los boxplot graficados, se observa la presencia de valores atípicos u outliers al lado derecho de las cajas, los cuales generan una asimetría o sesgo en la cola derecha de las distribuciones. Estos valores atípicos pueden ser resultado de errores de medición, valores extremos o anomalías en los datos.

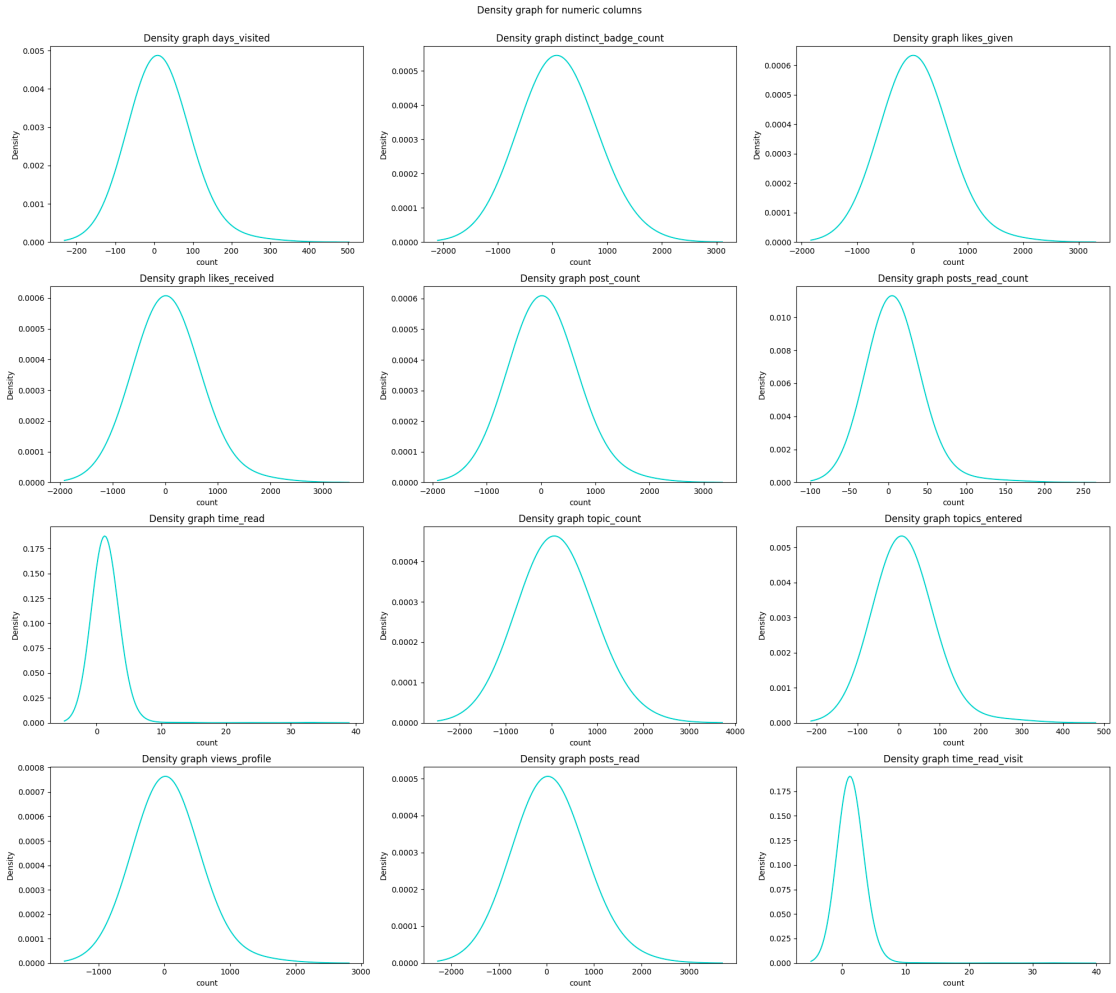

Figura 13: Gráficos de distribución de variables numéricas

Los gráficos de distribución de densidad respaldan la observación hecha en los boxplots, mostrando un sesgo muy marcado hacia la derecha en todas las variables.

Además, los gráficos de densidad también muestran que la mayoría de los datos están concentrados alrededor del valor cero. Esto sugiere que hay una alta frecuencia de valores bajos en las variables, lo que podría ser un indicador de una gran cantidad de ceros o valores faltantes en los datos.

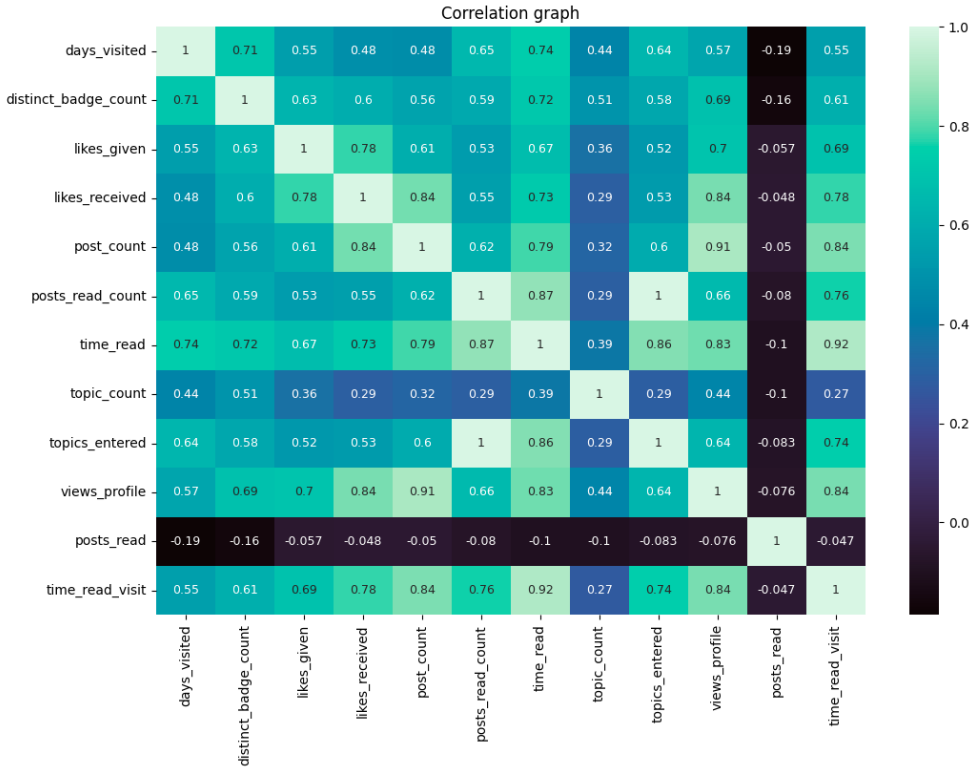

Figura 14: Matriz de correlación variables numéricas

En general, la matriz de correlación muestra que existe una fuerte correlación positiva entre la mayoría de las características del conjunto de datos. Esto indica que las personas que son más activas en el sitio web tienden a tener un comportamiento similar en otras características. Por ejemplo, las personas que visitan el sitio con más frecuencia también tienden a leer más publicaciones, ver más perfiles y visitar el sitio más tiempo.

Mapa de palabras de las publicaciones

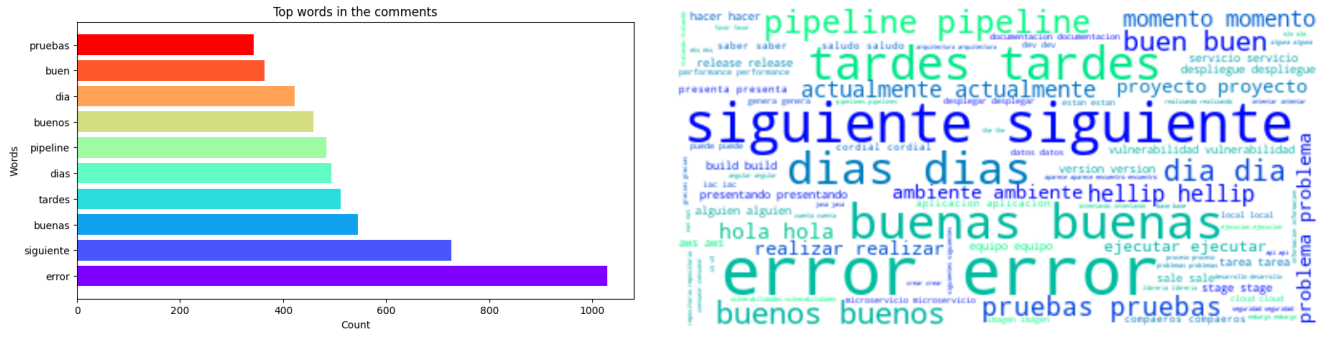

Figura 15: Análisis de palabras en las publicaciones

El gráfico de barras permite identificar de los temas principales de la conversación y las palabras que más se repitieron. Se puede observar que hay una frecuencia considerable de palabras relacionadas con la resolución de problemas técnicos, como "error", "siguiente" y "pruebas". También hay palabras que sugieren un tono positivo y colaborativo, como "buenas" y "buenos".

La nube de palabras sugiere que la conversación o texto se trata de la resolución de problemas técnicos, especialmente relacionados con la implementación de microservicios, la ejecución de pruebas y la resolución de errores. También parece haber una discusión sobre la gestión de proyectos, el desarrollo de pipelines y la comunicación entre equipos.

La presencia de palabras como "buenas" y "buenos" sugiere un tono positivo y colaborativo en la conversación.

4. Proceso de analítica

4.1. Pipeline principal

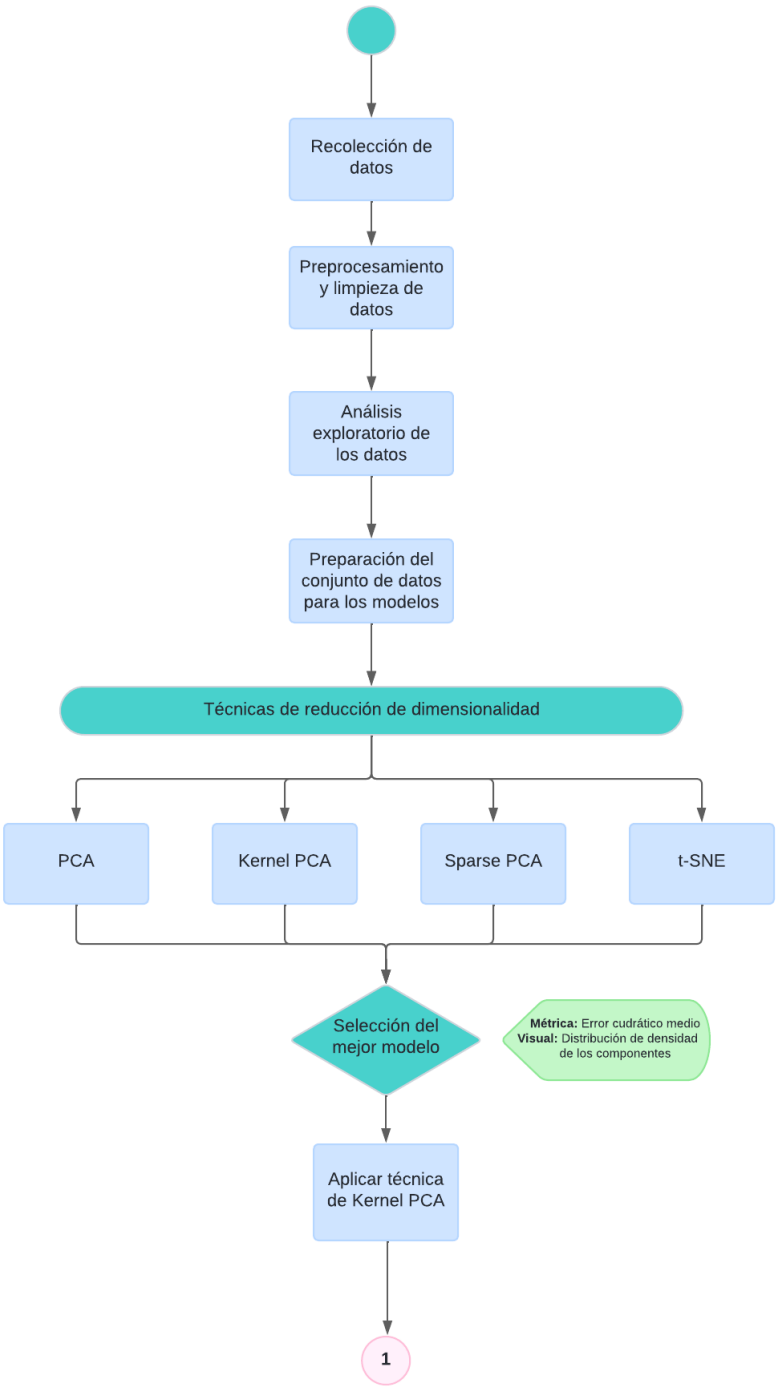

Figura 16: Pipeline principal segmentación de usuarios - Parte 1

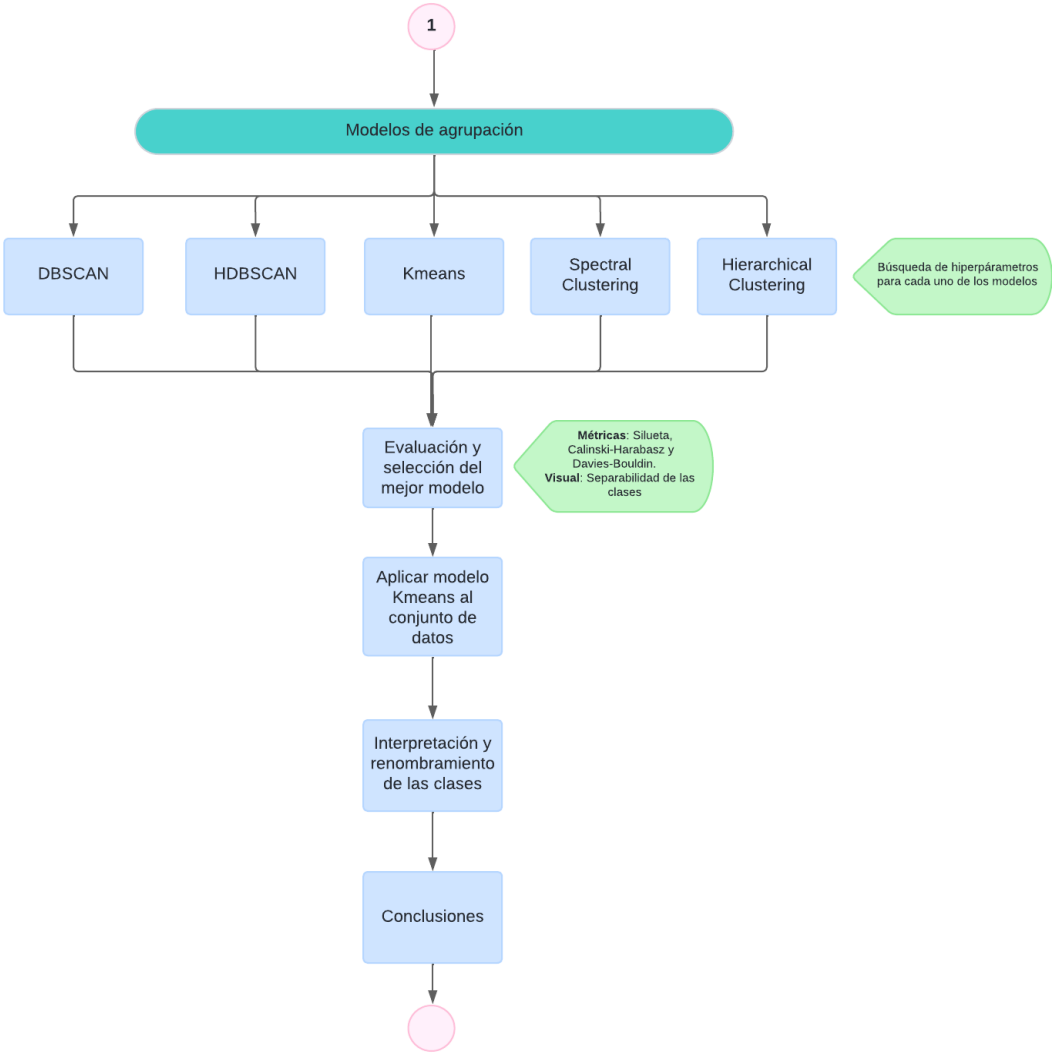

Figura 17: Pipeline principal segmentación de usuarios - Parte 2

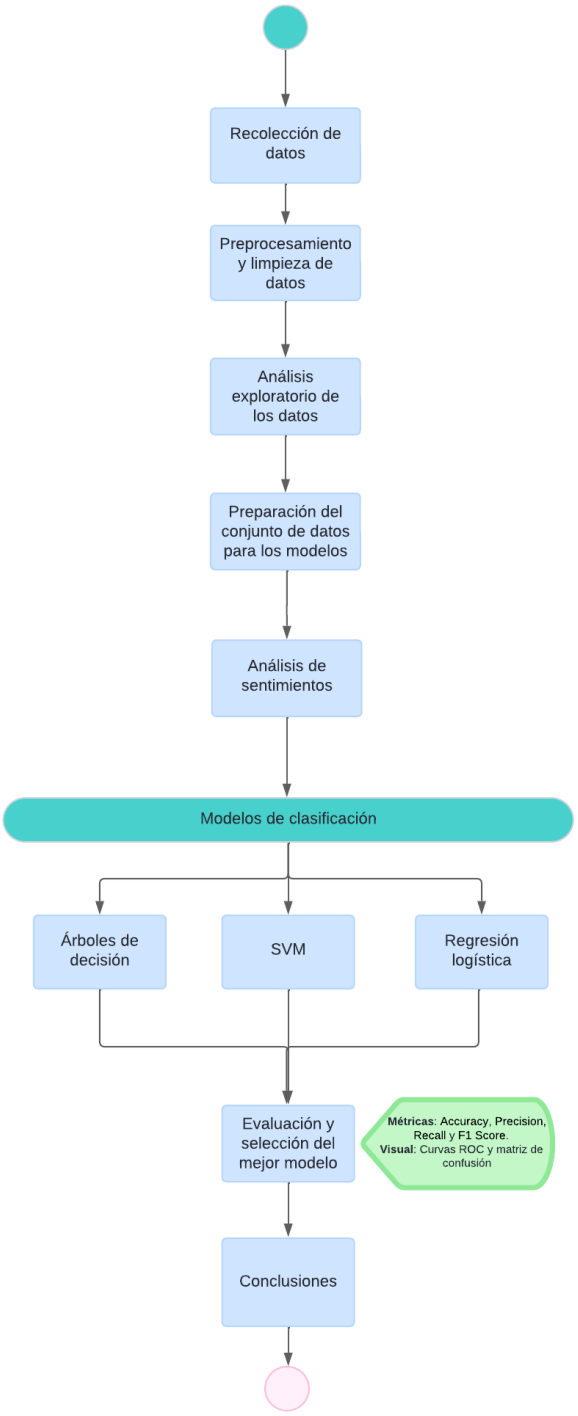

Figura 18: Pipeline principal análisis de sentimientos

4.2.Preprocesamiento

En esta sección, se describen las técnicas y métodos utilizados para preparar los datos antes de aplicar las técnicas de reducción de dimensionalidad y algoritmos de agrupación. El objetivo del preprocesamiento es transformar los datos brutos en un formato adecuado, asegurando que sean consistentes, completos y relevantes para el análisis posterior.

Transformaciones de las variables de tiempo

De acuerdo con el análisis exploratorio realizado, la variable ‘first_seen_at’ puede ser eliminada del conjunto de datos ya que puede ser representada por la variable ‘creation_user’.

Para evitar problemas de sesgo, sobreajuste o multicolinealidad en el modelo, se crearán nuevas variables basadas en las variables temporales originales. Estas nuevas variables se describen a continuación:

- **antiquity**: Calculada como la diferencia entre la fecha de creación del usuario y la fecha actual del sistema.
- **time_between_visits**: Calculada como la diferencia entre las fechas 'last_seen_at' y 'previous_visit_at'.
- **time_between_posts**: Calculada como la diferencia entre las fechas 'last_posted_at' y 'first_post_created_at'.

Finalmente, se obtienen los siguientes gráficos de distribución de densidad para las nuevas variables creadas en el conjunto de datos; es importante aclarar que las variables originales fueron eliminadas del conjunto de datos una vez se crearon las descritas previamente.

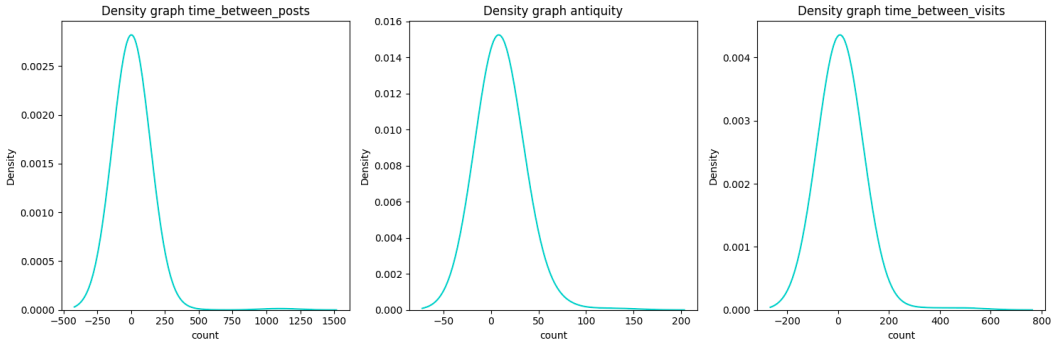

Figura 19: Gráficos de distribución de densidad variables de tiempo

Los gráficos de distribución de densidad, al igual que las otras variables numéricas muestran un sesgo muy marcado hacia la derecha en todas las variables.

Transformaciones de variables categóricas

De acuerdo con el análisis exploratorio realizado previamente, se eliminarán las siguientes variables categóricas debido a su baja variabilidad, lo que indica que no aportan información significativa al modelo: (moderator, manual_locked_trust_level, location, badge_granted_title, mobile, flags_ignored, developer y software).

Con las dos variables restantes, 'trust_level' y 'badge_name', se realiza un balance de las clases que conforman estas variables, ya que un desbalance significativo puede afectar negativamente el rendimiento del modelo. Un análisis detallado de la distribución de clases asegura que todas las categorías estén adecuadamente representadas, evitando así posibles sesgos y mejorando la capacidad del modelo para generalizar.

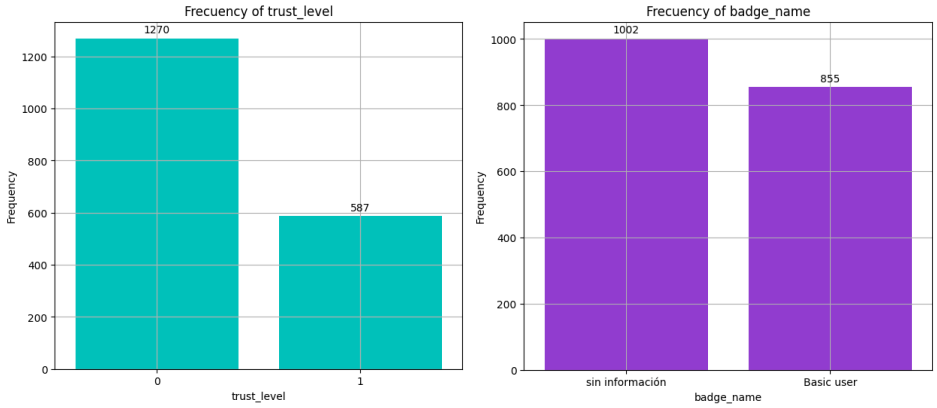

Figura 20: Balance de variables categóricas

Las variables siguen presentando un desbalance; pero no tan considerable como al inicio. Finalmente, se realiza un reemplazo de la clase 'sin información' por el valor 0 y la clase 'Basic user' por el valor 1 para la variable 'bagde_name'; para manejar la variable como valores numéricos.

Transformaciones variables numéricas

Dado que las columnas numéricas presentan valores atípicos y parecen seguir una distribución normal, se realizan pruebas de normalidad Shapiro-Wilk para validar su distribución y poder determinar que método de imputación de valores atípicos.

La prueba de Shapiro-Wilk se utiliza para evaluar la hipótesis de normalidad de una muestra de datos. Específicamente, esta prueba verifica la siguiente hipótesis:

- **Hipótesis nula (H_0)**: La muestra proviene de una distribución normal.
- **Hipótesis alternativa (H_1)**: La muestra no proviene de una distribución normal.

.

Al aplicar la prueba a cada una de las columnas numéricas presentes en el conjunto de datos se obtiene que ninguna sigue una distribución normal.

De acuerdo con los resultados ninguna de las variables numéricas sigue una distribución normal; por lo que, para la detección de valores atípicos se usará el método Z-Score modificado, ya que cuando los datos son asimétricos o no se distribuyen de forma normal podemos utilizar el Z-score modificado, también conocido como MAD-Z-Score. Este, a diferencia del Z-score, utiliza la mediana y la desviación absoluta mediana (MAD en inglés) en lugar de la media y la desviación estándar con el fin de evitar el efecto de los outliers sobre estas dos últimas medidas. [9].

Al graficar los diagramas de cajas para las variables numéricas, se observa que todavía hay presencia de valores atípicos, aunque en menor cantidad que antes de la imputación:

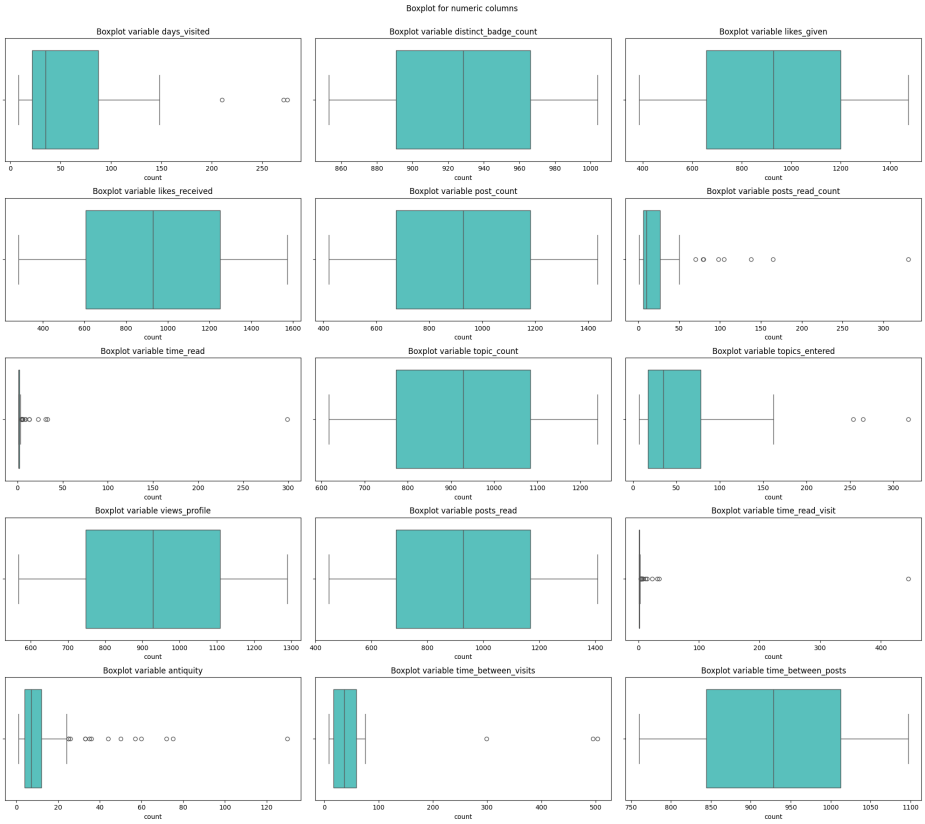

Figura 21: Boxplot variables numéricas post imputación

En los gráficos, se observa como en la mayoría de variables numéricas se logran eliminar los valores atípicos; persistiendo todavía algunos en las variables: days_visited, post_read_count, time_read, topics_entered, posts_read, time_read_visit, antiquity y time_between_visits.

Al observar ahora los gráficos de distribución de densidad, se puede apreciar que ya no presentan sesgos tan marcados hacia la derecha y que la distribución parece más simétrica:

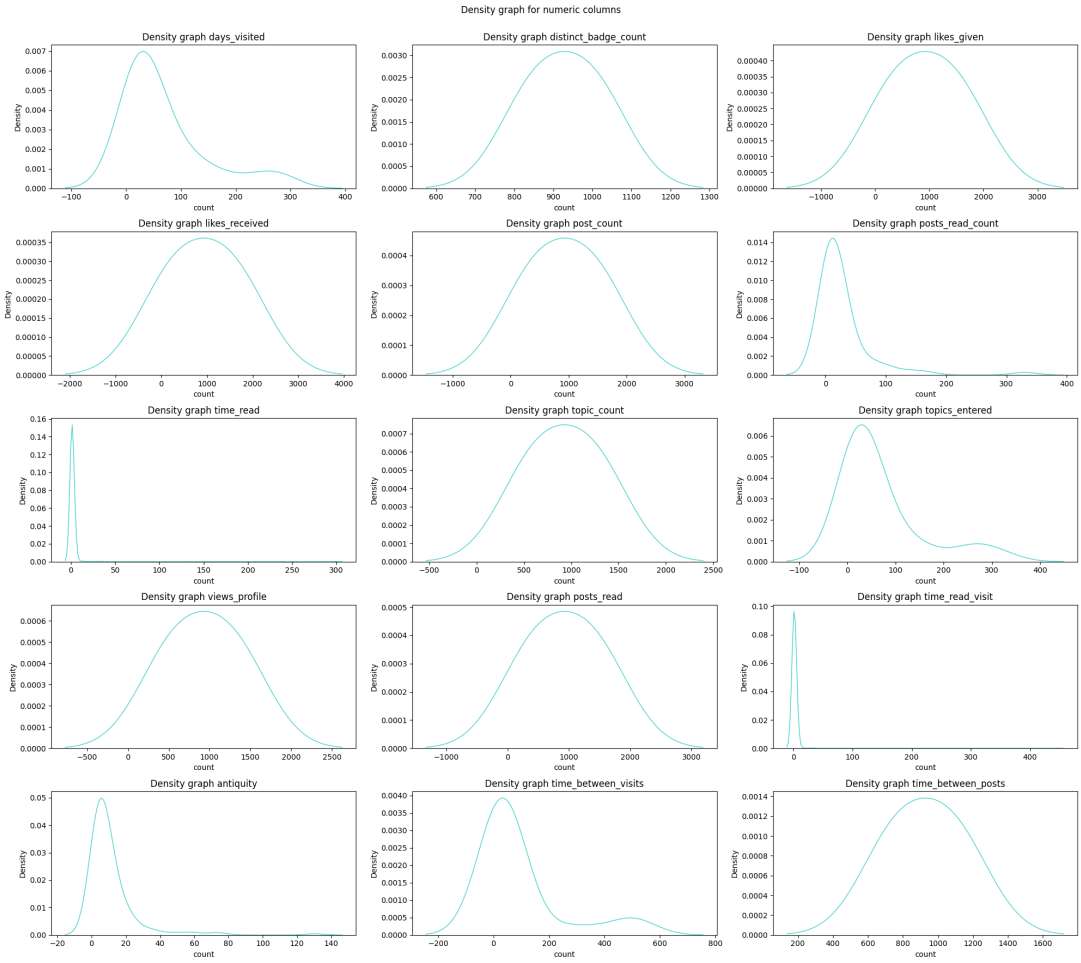

Figura 22: Gráficos de distribución de densidad de variables numéricas post imputación

Finalmente, después de realizar las transformaciones necesarias en todas las variables del conjunto de datos y asegurar que todas sean numéricas, se procede a escalar el conjunto de datos. Este paso es crucial para garantizar que todas las características tengan la misma oportunidad de influir en el resultado final.

Se usará la técnica **MinMaxScaler** la cual escala los datos para que estén en el rango de [0, 1]. Es útil cuando la distribución de los datos no es gaussiana o cuando se desea preservar las relaciones entre los valores de las variables. Sin embargo, es sensible a los valores atípicos porque pueden sesgar el rango de los datos. [10]

Por último, se toma una muestra de 100 registros del conjunto de datos original para realizar pruebas posteriormente de los modelos seleccionados con registros desconocidos por el modelo.

Algoritmos de reducción de dimensionalidad

El objetivo de esta sección es aplicar y comparar diversos algoritmos de reducción de la dimensionalidad con el fin de seleccionar la más adecuada según dos criterios principales: la interpretación de la variabilidad de los datos y el porcentaje de pérdida de información.

La reducción de la dimensionalidad de las características del conjunto de datos; la cual es una forma de convertir un conjunto de datos de dimensiones elevadas en un conjunto de datos de dimensiones menores, asegurando que la información que proporciona es similar en ambos casos. [11]

Al evaluar las siguientes diferentes técnicas de reducción de dimensionalidad se obtienen los siguientes resultados para los errores MSE:

Tabla 1: Resultados técnicas reducción de dimensionalidad

Técnica	Error
PCA	0.20218
Kernel PCA	0.11472
Sparse PCA	0.25927

Y los siguientes gráficos de matriz de dispersión y distribución de densidad de los 10 componentes principales:

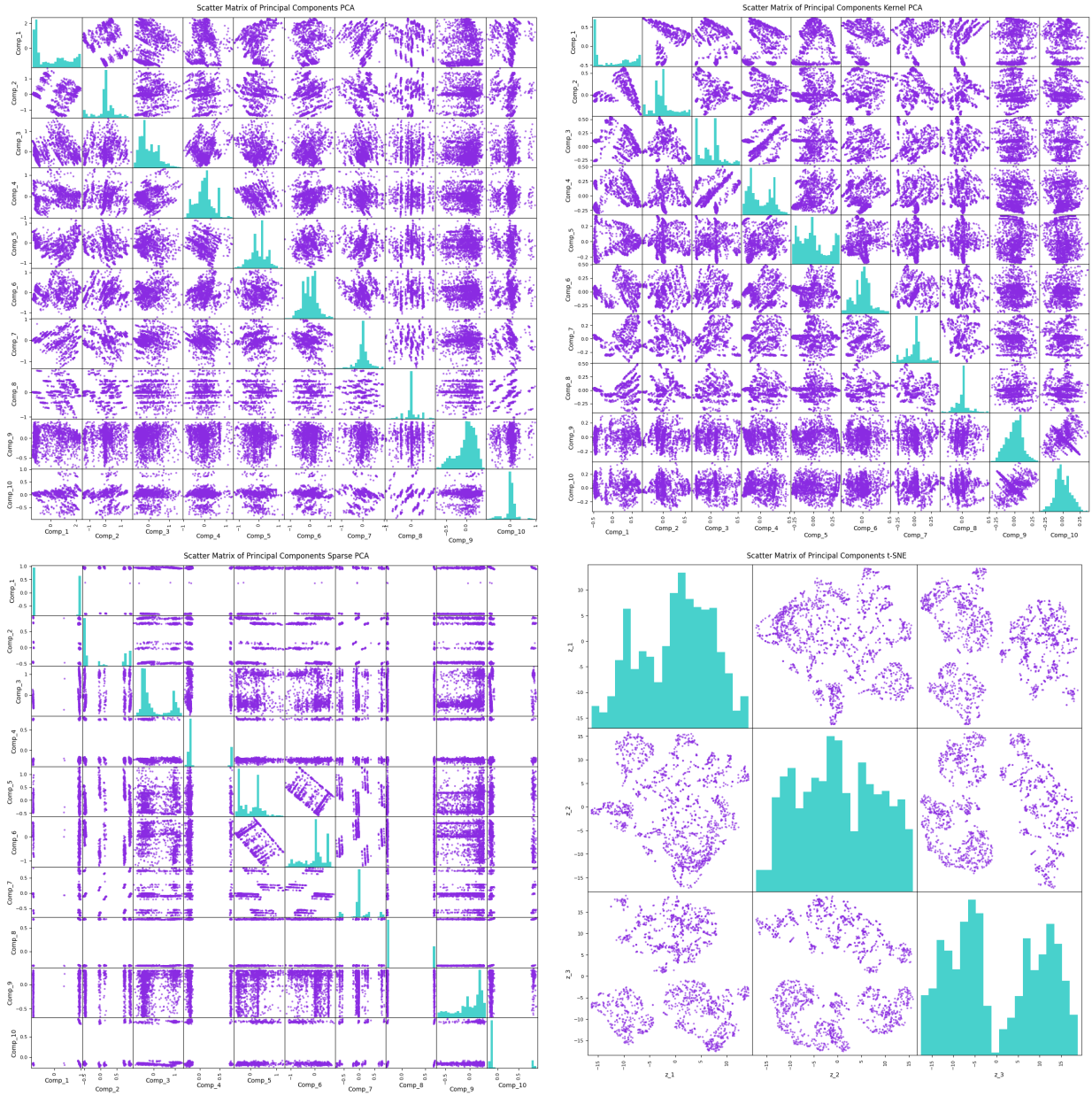

Figura 23: Matriz de dispersión técnicas de reducción de dimensionalidad

De acuerdo con los resultados obtenidos para el error promedio de cada una de las técnicas aplicadas, el método con el que se obtiene menor pérdida de información es **Sparse PCA** con 0.0.08484; sin embargo, el comportamiento de las distribuciones sigue patrones unimodales o bimodales, por lo que el parámetro de esparcimiento puede generar cierta pérdida de información.

Los resultados del análisis de reducción de dimensionalidad mostraron que **Kernel PCA** obtuvo el segundo menor error promedio 0.10556 entre los métodos evaluados, lo que sugiere una menor

pérdida de información. Adicionalmente, los gráficos de densidad parecen tratar de seguir más una distribución normal que Sparse PCA, teniendo en cuenta que también presenta comportamientos unimodales y bimodales; de igual forma no se logra ver una separación clara entre diferentes grupos, ya que los puntos parecen estar más dispersos a lo largo de las líneas horizontales.

Si bien t-SNE redujo la dimensionalidad a 3 componentes, los gráficos de dispersión no muestran una clara agrupación de los datos; por lo que no se opta por este algoritmo.

De acuerdo con los resultados, Kernel PCA parece ser una buena opción para reducir la dimensionalidad de los datos, especialmente si la complejidad computacional no es una gran limitación; ya que, muestra una mayor separación y distribución de los puntos en un espacio más continuo y curvado.

De acuerdo con lo anterior, se selecciona el método kernel PCA como el que mejor reduce la dimensionalidad de mis datos:

```
KernelPCA
KernelPCA(fit_inverse_transform=True, gamma=0.2, kernel='rbf', n_components=10)
```

Análisis de sentimientos

Realizar un análisis de sentimiento en los textos de la columna 'excerpt', usando la librería VADER (Valence Aware Dictionary and sEntiment Reasoner) y luego clasifica cada texto como positivo, negativo o neutral. Vader viene de “Valence Aware Dictionary and sEntiment Reasoner” y es la librería que usa Python para el análisis de sentimientos.

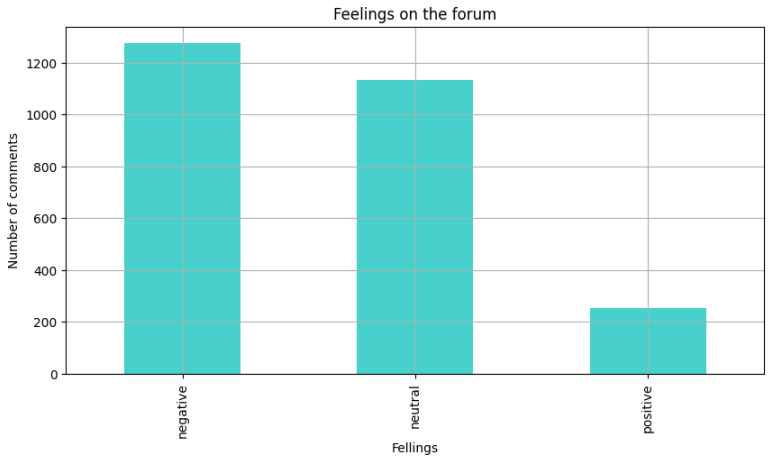

Figura 24: Análisis de sentimientos en el foro

El gráfico muestra un foro donde predominan los comentarios neutros y negativos, sugiriendo un ambiente principalmente enfocado en la resolución de problemas o la discusión de temas serios, con una limitada expresión de emociones positivas.

Se convierten los valores a numéricos los sentimientos: negative (0), neutral (1) y positivo (2). Se usa CountVectorizer para convertir las publicaciones en datos numéricos que pueden ser utilizados por algoritmos de aprendizaje automático.

Finalmente, se divide el conjunto de datos en entrenamiento y prueba.

4.3.Modelos

En esta sección se realiza el entrenamiento de diferentes modelos de agrupamiento (DBSCAN, HDBSCAN, Espectral, Jerárquico y Kmeans) para identificar posibles segmentaciones de usuarios dentro del foro. De igual forma, se busca realizar la validación de cada uno de acuerdo con el resultado de los coeficientes de Calinski-Harabasz, Silhouette y Davies-Bouldin para encontrar la mejor combinación de hiperpárametros para cada modelo de agrupación evaluados.

Para definir cada uno de los modelos, se llevó a cabo una búsqueda de hiperpárametros para cada distancia posible. Posteriormente, se evaluaron los modelos resultantes y se seleccionó el mejor para cada algoritmo, considerando la combinación de hiperpárametros que optimizara su

rendimiento. Se obtienen los siguientes mejores modelos basados en los hiperpárametros mostrados por los coeficientes descritos:

Tabla 2: Selección de los mejores modelos de agrupación

Modelo	Basado en la Métrica	Clases
`best_model_DBSCAN_euclidian = DBSCAN(eps = 0.4, min_samples = 20, metric = 'euclidean')`	Silueta (0.3930) Calinski-Harabasz (1221.5912)	2
`best_model_HDBSCAN_euclidian = HDBSCAN(min_cluster_size = 5, min_samples = 20, cluster_selection_epsilon = 0.4, metric = 'euclidean', alpha = 1.0)`	Silueta (0.3930) Calinski-Harabasz (1221.5912)	2
`best_model_spectral_ch = SpectralClustering(n_clusters = 2, eigen_solver = 'arpack', n_components = 2, n_init = 10, gamma = 0.01, affinity='rbf', eigen_tol = 1e-3, assign_labels='kmeans')`	Davies-Bouldin (0.9115)	2
`best_hierarchical_model= AgglomerativeClustering(n_clusters = 2, connectivity = A, linkage='ward')`	Calinski-Harabasz (1221.5912)	2
`best_kmeans_model= KMeans(n_clusters=2, random_state=42)`	Calinski-Harabasz (1222.3794)	2

4.4.Métricas

Para calcular las métricas para evaluar los modelos de agrupación se usa la librería metrics de sklearn:

```
from sklearn.metrics import silhouette_score
from sklearn.metrics import calinski_harabasz_score
from sklearn.metrics import davies_bouldin_score
```

```
labels = model.labels_
silhouette = silhouette_score(data_prep_reducted, labels)
calinski_harabasz = calinski_harabasz_score(data_prep_reducted, labels)
davies_bouldin = davies_bouldin_score(data_prep_reducted, labels)
```

Figura 25: Métricas de modelos de agrupación

Para calcular las métricas para evaluar los modelos de clasificación se usa la librería metrics de sklearn:

```
from sklearn.metrics import (precision_score, recall_score, f1_score, classification_report,
                             accuracy_score)
# Calcular métricas
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred, average='weighted')
recall = recall_score(y_test, y_pred, average='weighted')
f1 = f1_score(y_test, y_pred, average='weighted')
confusion = confusion_matrix(y_test, y_pred)
classification_rep = classification_report(y_test, y_pred)
```

Figura 26: Métricas de modelos de clasificación

Para calcular las métricas de negocio se realizan los siguientes cálculos:

Tasa de participación:

$$\text{engagement\_rate} = \frac{\#\ usuarios\ participativos}{total\ de\ usuarios} x100$$

Tasa de interacción:

$$\text{interaction\_rate} = \frac{\text{total\_actions}}{\#\ de\ usuarios} x100$$

Donde:

$$\begin{aligned}\text{total\_actions} = \sum(&\text{likes\_given} + \text{likes\_received} + \text{post\_count} + \text{posts\_read\_count} \\ &+ \text{topic\_count} + \text{topics\_entered} + \text{time\_read} + \text{views\_profile} \\ &+ \text{posts\_read})\ \text{para}\ \text{classes\_predicted} = '\text{participatory\_users}'\end{aligned}$$

Tasa de retención:

$$\text{retention\_rate} = \frac{\#\ usuarios\ que\ regresaron}{\#\ usuarios\ activos\ iniciales} x100$$

Donde:

- # usuarios que regresaron: El número de usuarios que cumplen con el criterio de retorno; en este caso, tiempo entre visitas menor o igual que el umbral de retención (0.5).
- # usuarios activos iniciales**:** El número total de usuarios considerados inicialmente activos (en este caso, los usuarios clasificados como "participatory_users").

Proporción de sentimientos neutros:

$$\text{neutral\_proportion} = \frac{\#\ sentimientos\ neutros}{Prediciones\ totales} x100$$

Proporción de sentimientos positivos:

$$\text{positive\_proportion} = \frac{\#\ sentimientos\ positivos}{Prediciones\ totales} x100$$

5. Metodología

5.1. Baseline

Al aplicar le modelo seleccionado previamente **K-means** sobre el conjunto de datos original completo, se obtienen los siguientes resultados:

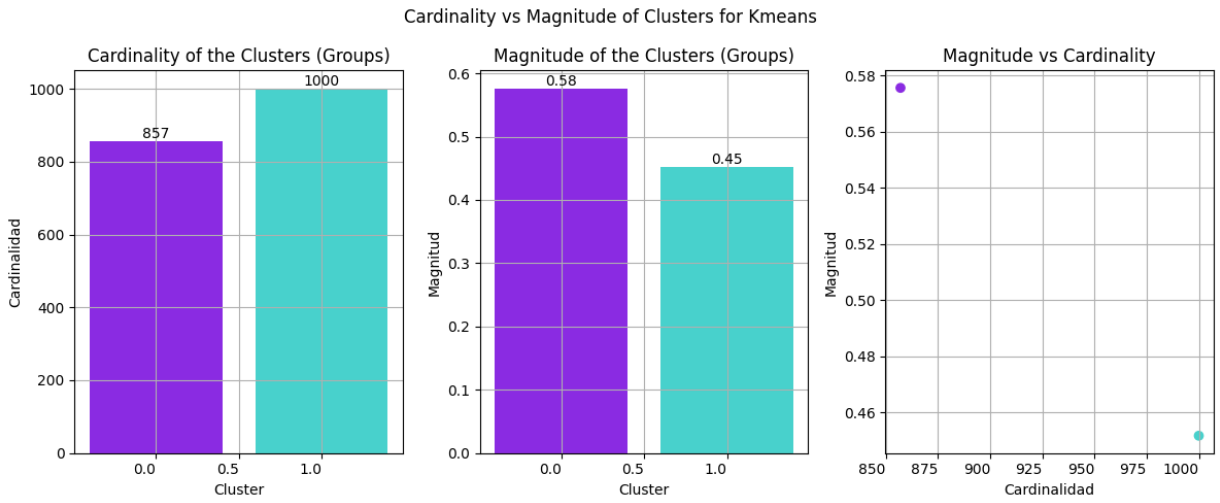

Figura 27: Cardinalidad y magnitud modelo Kmeans

De acuerdo con el gráfico se observa que:

- El primer gráfico muestra la cantidad de puntos de datos que pertenecen a cada grupo. El grupo violeta 0 tiene una cardinalidad de 857, mientras que el grupo turquesa 1 tiene una cardinalidad de 1000 aproximadamente.

- El segundo gráfico muestra la magnitud de cada grupo; la cual se calcula como la distancia promedio de todos los puntos dentro del clúster a su centroide. El grupo violeta 0 tiene una magnitud de 0.58 mayor que la del grupo turquesa 1 de 0.45. La magnitud presenta un desbalance menos adecuado que la cardinalidad.
- El tercer gráfico representa la relación entre la magnitud y la cardinalidad de cada grupo. Se puede ver que el grupo turquesa 1 tiene una mayor magnitud y una menor cardinalidad que el grupo violeta 0.

En resumen, estos gráficos sugieren que el algoritmo K-means ha encontrado dos grupos con diferentes características. El grupo turquesa (1) es más pequeño en cantidad de datos, pero tiene una mayor magnitud. El grupo violeta (0) es más grande en cantidad de datos pero tiene una menor magnitud.

Los coeficientes obtenidos con esta primera iteración son:

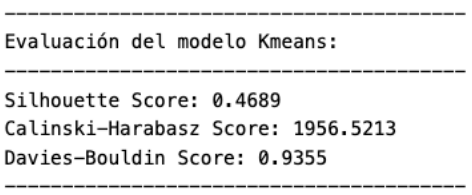

Figura 28: Métricas primera iteración

De acuerdo con estos resultados obtenidos se puede afirmar que el coeficiente de Silhouette los clústeres están razonablemente bien separados y que los puntos están bien agrupados dentro de sus propios grupos. Con respecto al valor obtenido en el coeficiente de Calinski-Harabasz, que es bastante alto, sugiere que los grupos están bien definidos y que la dispersión entre los grupos es baja en comparación con la dispersión entre grupos. El coeficiente de Davies-Boulding es medianamente bueno, indicando que los grupos están bien separados y que la distancia intra-grupos es pequeña.

En el análisis de sentimientos se obtiene los siguientes resultados:

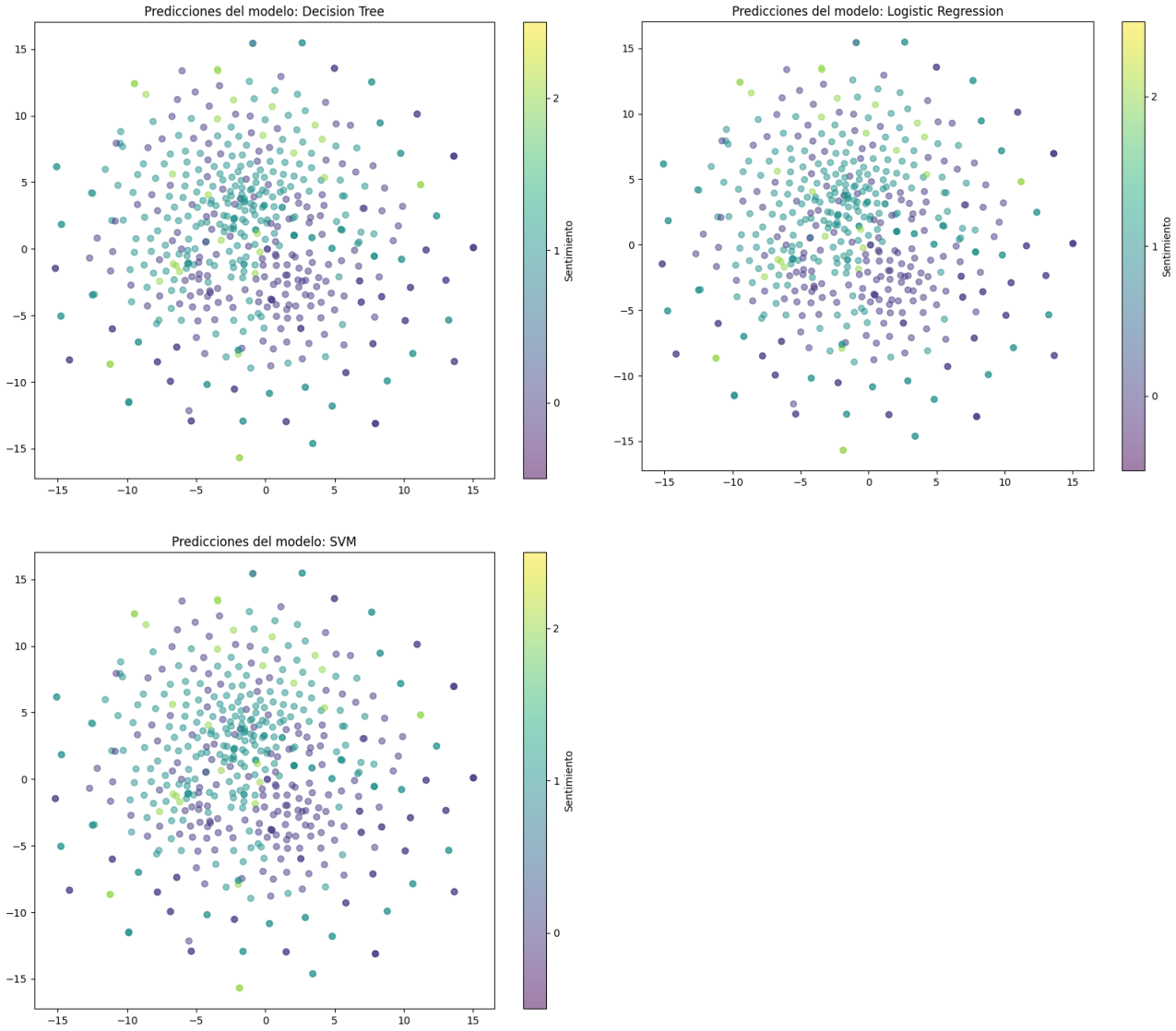

Figura 29: Predicciones de los modelos

Los resultados gráficos de los tres modelos (Regresión Logística, Árbol de Decisión y SVM) muestran una tendencia a predecir con mayor frecuencia sentimientos negativos y neutros. Una de las causas es el desbalanceo en las categorías de sentimientos presentes en los datos de entrenamiento.

El Árbol de decisión parece separar las clases de sentimiento de forma más definida que la Regresión Logística, sin embargo, podría ser susceptible al sobreajuste. El SVM presenta una dispersión de colores intermedia, lo que sugiere una separación de clases menos definida que el Árbol de Decisión pero más definida que la Regresión Logística.

5.2. Validación

Se evalúan los cinco (5) modelos seleccionados previamente para escoger el que mejor se adapta al conjunto de datos. Para esto, se comparan las métricas Silueta, Calinski-Harabasz y Davies-Bouldin, se visualiza el balance de las clases y se utilizan gráficos de radar para identificar los principales componentes que contribuyen a cada clase. Finalmente, se realiza un análisis del comportamiento del modelo utilizando una muestra de 100 registros extraídos del conjunto de datos original en la etapa de preparación.

De acuerdo con lo evaluado, inicialmente se descartan los modelos DBSCAN y HDBSCAN ya que al generar predicciones sobre la muestra de datos clasifico como ruido 57% y el 35% de las muestras; lo que indica que estos modelos son mejores para detectar anomalías que para clasificar grupos.

Una vez descartados estos dos modelos, se evalúan las métricas que se obtienen con el conjunto de datos original menos las 100 muestras extraídas:

Tabla 3: Métricas modelos conjunto de datos original

Modelo	Silhouette	Calinski-Harabasz	Davies-Bouldin	Clases
SpectralClustering	0.3911	1215.6539	1.1776	Clase 1: 808 Clase 0: 949
AgglomerativeClustering	0.3930	1221.5912	1.1755	Clase 1: 810 Clase 0: 947
KMeans	0.3930	1221.5912	1.1750	Clase 1: 812 Clase 0: 945

Los tres algoritmos tienen métricas muy similares, lo que sugiere que su performance es comparable. AgglomerativeClustering y K-Means tienen el mejor valor de Calinski-Harabasz, lo que podría indicar que la separación entre grupos es ligeramente mejor que con SpectralClustering.

Kmeans tiene el mejor valor de Davies-Bouldin, lo que podría indicar que la separación entre grupos es ligeramente mejor que con los otros dos métodos. Todos los algoritmos encontraron dos clases relativamente equilibradas.

Para las métricas obtenidas con los datos de muestra tomados previamente, obtenemos que:

Tabla 4: Métricas modelos muestra 100 registros

Modelo	Silhouette	Calinski-Harabasz	Davies-Bouldin	Clases
SpectralClustering	0.3755	62.9010	1.2222	Clase 1: 45 Clase 0: 55
AgglomerativeClustering	0.3555	57.9253	1.2378	Clase 1: 45 Clase 0: 55
KMeans	0.3795	63.2812	1.2226	Clase 1: 45 Clase 0: 55

De acuerdo con estos resultados se obtiene que:

- KMeans muestra ligeramente la mejor performance en Silhouette y Calinski-Harabasz, pero las diferencias son muy pequeñas.

- SpectralClustering tiene ligeramente la mejor performance en Davies-Bouldin, pero la diferencia es también muy pequeña.

- Los tres algoritmos tienen un rendimiento similar, con diferencias mínimas en las métricas.

Teniendo en cuenta todos los resultados evaluados a lo largo de este notebook, se determina que el algoritmo **K-means** es el modelo más adecuado para este conjunto de datos.

Para los modelos de clasificación se realizan las siguientes validaciones:

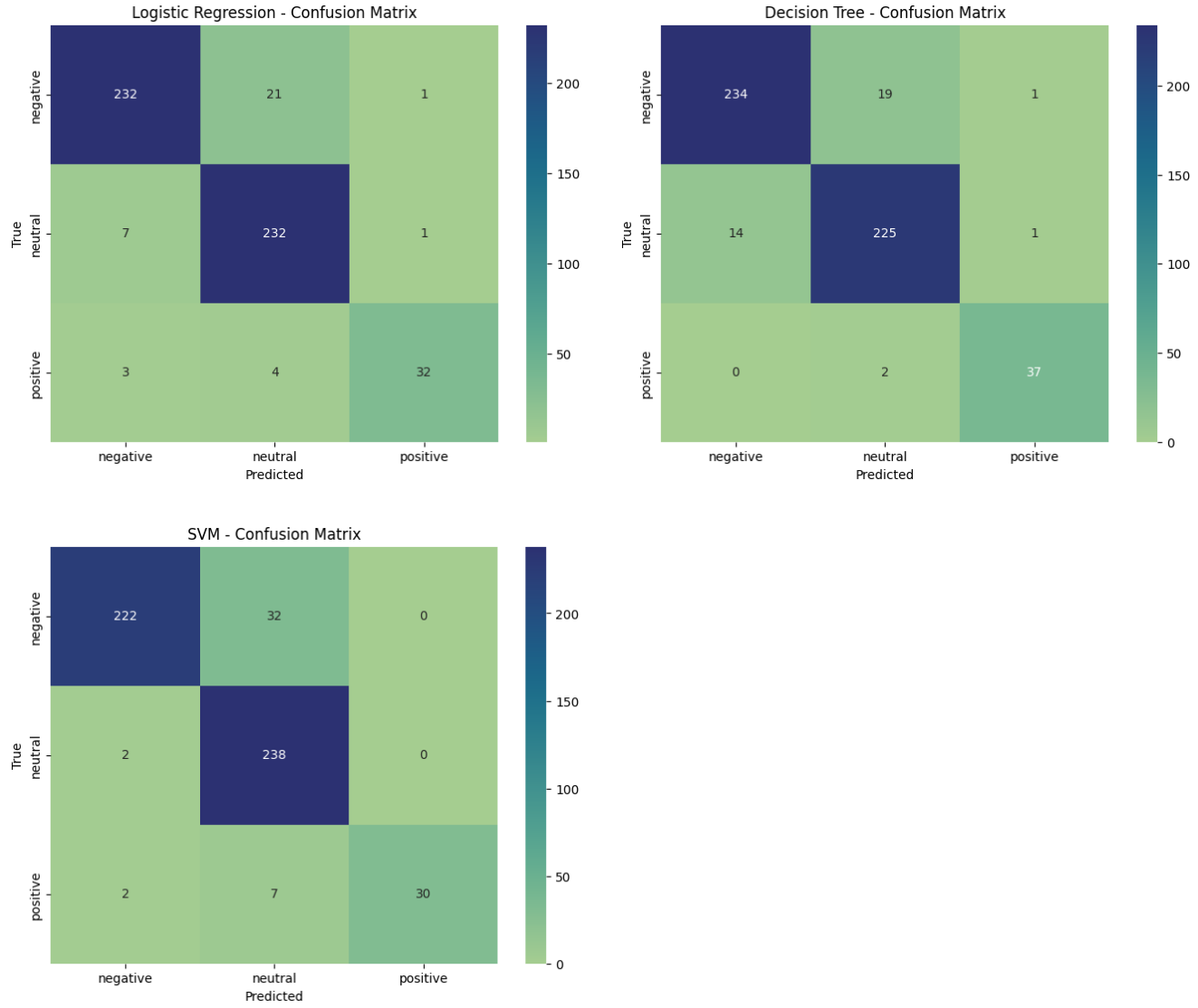

Figura 30: Matriz de confusión de los modelos

El análisis de las matrices de confusión revela que los tres modelos (Regresión Logística, Árbol de Decisión y SVM) presentan un rendimiento superior en la clasificación de sentimientos negativos y neutros, con un bajo desempeño en la clasificación de sentimientos positivos.

El modelo de Regresión Logística tiene una mayor precisión en la clasificación de comentarios negativos y neutrales, mientras que el Árbol de Decisión y el SVM muestran una mejor precisión en la clasificación de comentarios positivos. Sin embargo, todos los modelos presentan dificultades para clasificar correctamente los comentarios positivos, con una menor cantidad de predicciones correctas para esta categoría.

5.3. Iteraciones y evolución

Para optimizar el análisis de los datos, se llevó a cabo un proceso de cuatros etapas:

- Primero, se evaluaron diferentes técnicas de reducción de dimensionalidad (PCA, Kernel PCA, Sparse PCA y t-SNE) para identificar la más efectiva en la transformación del conjunto de datos.
- Segundo, se experimentó con distintos modelos de agrupamiento (DBSCAN, HDBSCAN, Espectral, Jerárquico y K-means) con el objetivo de encontrar aquel que mejor categorizara los datos en clases distintas, priorizando la optimización de las métricas de evaluación. Para esto se usó la muestra de 100 registros obtenida en la etapa de preparación de datos.
- En tercer lugar, se lleva a cabo un análisis de patrones utilizando el modelo K-means seleccionado. Este análisis permite identificar las características que definen cada grupo y, por lo tanto, renombrar las clases de acuerdo a sus propiedades distintivas.
- Finalmente, se evalúan las métricas de negocio usando el conjunto de datos original y un conjunto de datos sintético generado a partir de una red neuronal GAN (Generative Adversarial Network).

Para el análisis de sentimientos, se realizaron las siguientes etapas:

- Primero, se realiza el análisis de sentimientos sobre las publicaciones realizadas en el foro de discusión.
- Segundo, se evalúan los modelos de clasificación (árboles de decisión, SVM y regresión logística y se selecciona el mejor de acuerdo a los resultados obtenidos en las diferentes métricas.

5.4 Herramientas

- Se extrajeron los datos utilizando PostgreSQL y consultas SQL para exportar la información necesaria para el estudio.
- Para implementar los modelos de agrupación, se utilizó la librería Scikit-learn en Python. Esta librería proporciona algoritmos de clustering como K-means y DBSCAN. Los datos

fueron manipulados y analizados con Pandas, y la visualización de los resultados se realizó con Matplotlib y Seaborn.

- Se utilizó Jupyter Notebook como entorno de trabajo para ejecutar el código y visualizar los resultados de los modelos de agrupación.

6. Resultados y discusión

Al aplicar del modelo Kmeans, se realiza una análisis de patrones para identificar las características que definen a cada grupo específicamente:

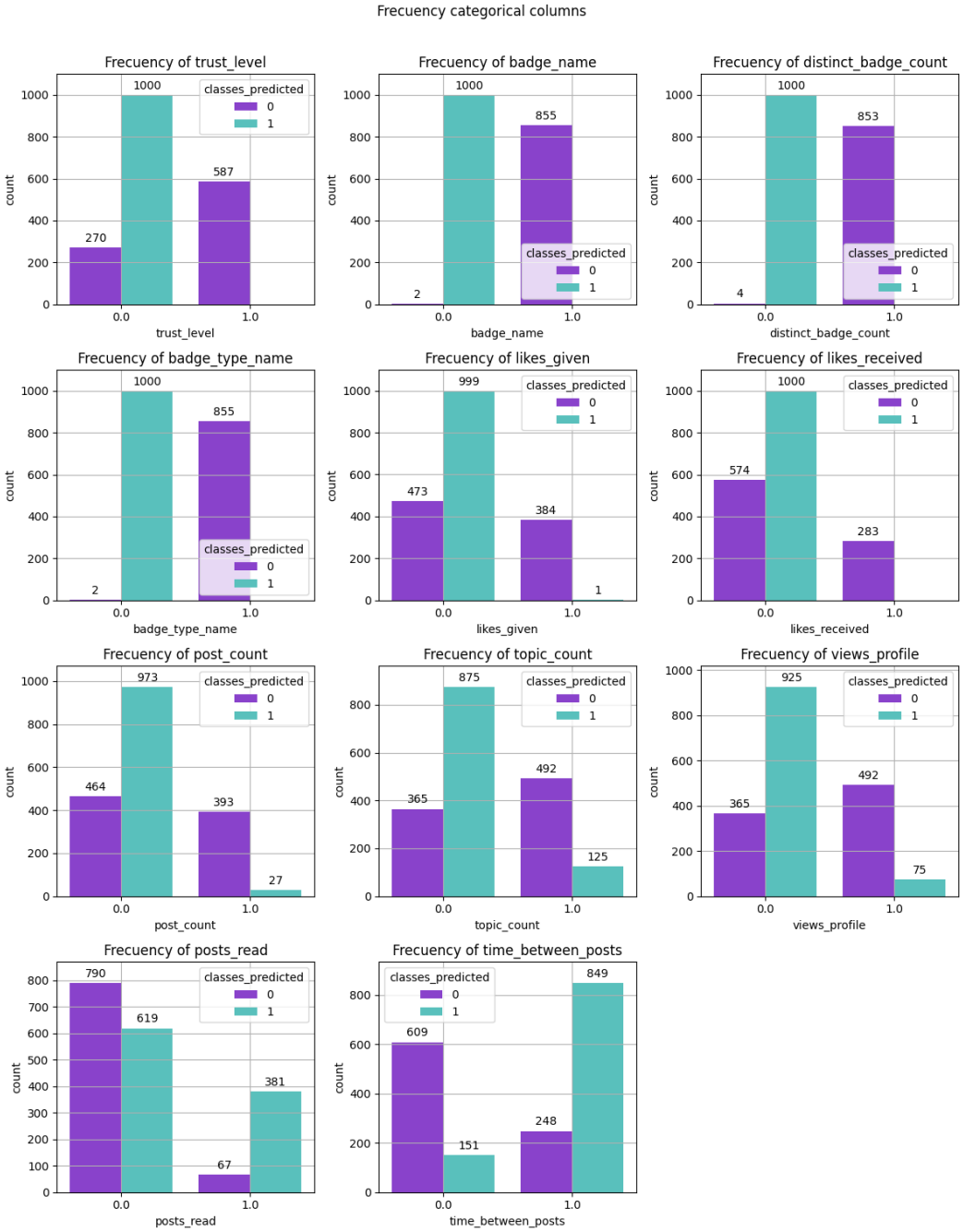

Figura 31: Gráficos de frecuencia columnas categóricas

En los gráficos anteriores se observa que:

- La clase violeta (0) esta predominantemente representada por usuarios con alto nivel de confianza, que poseen medallas distintivas y participan activamente en el foro, otorgando y recibiendo "me gusta", creando temas y leyendo publicaciones con frecuencia.

- La clase turquesa (1) esta mayormente integrada por usuarios con bajo nivel de confianza, que no poseen medallas distintivas y son menos activos en el foro. Estos usuarios tienen menor frecuencia de publicación de mensajes, de lectura de publicaciones, y otorgan/reciben menos "me gusta".

Se observa que las características como "likes_given", "likes_received" y "post_count" no son predictores muy útiles para la clasificación, ya que ambas clases muestran un comportamiento similar en relación a estas características.

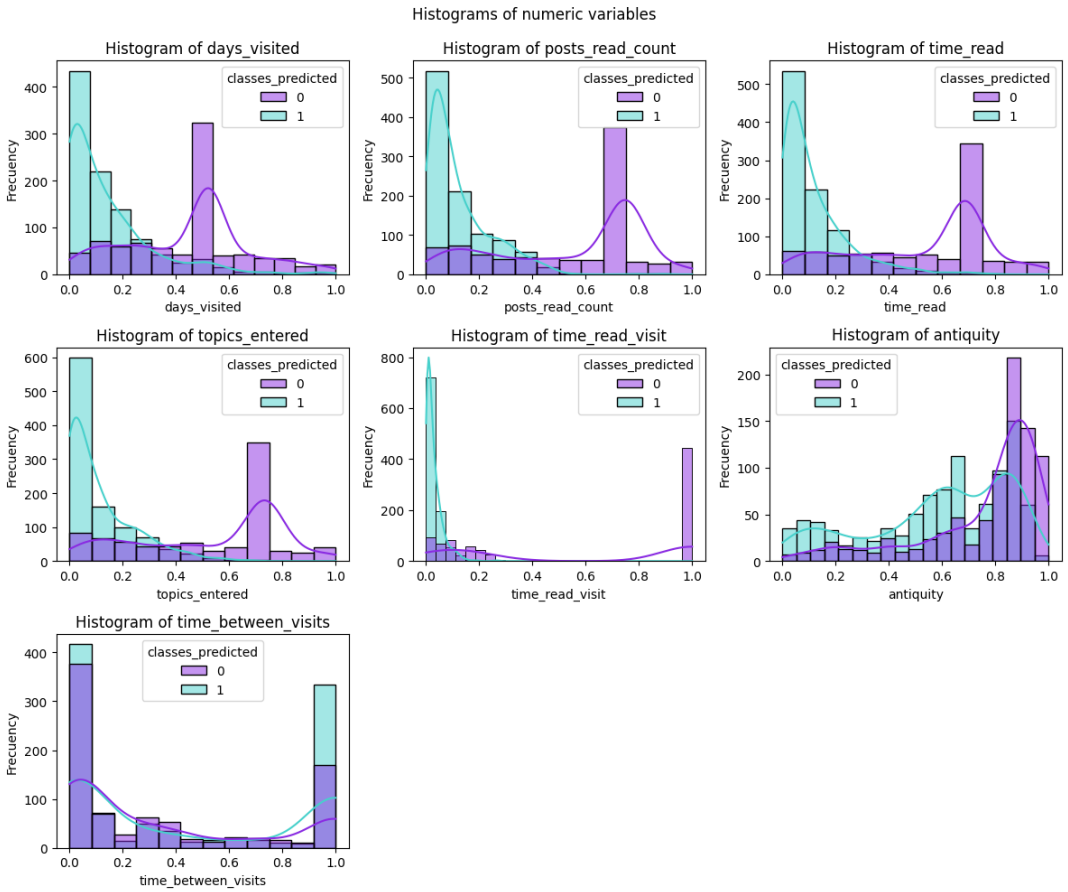

Figura 32: Histogramas variables numéricas

El análisis de los gráficos de distribución de las variables "días visitados", "publicaciones leídas", "tiempo de lectura", "temas ingresados" y "tiempo de lectura durante la visita" muestra un patrón consistente: la clase turquesa (1) presenta una mayor frecuencia de usuarios con actividad menos

intensa, caracterizada por periodos de visita más cortos, menos publicaciones leídas, menor tiempo de lectura y menor cantidad de temas ingresados. En contraste, la clase violeta (0) muestra una mayor frecuencia de usuarios con actividad más intensa, con periodos de visita más largos, mayor lectura de publicaciones, mayor tiempo de lectura y mayor cantidad de temas ingresados. La variable "antigüedad" no parece ser un factor determinante para distinguir las clases. La clase turquesa (1) también presenta una mayor frecuencia de usuarios con un menor tiempo entre visitas, mientras que la clase violeta (0) muestra un mayor número de usuarios con un tiempo más extenso entre visitas.

El análisis de agrupamiento con K-means identificó dos grupos distintos de usuarios en el foro de discusión: usuarios activos y usuarios inactivos. Los usuarios activos se caracterizan por ser más comprometidos e interactuar en el foro de discusión de forma activa otorgando más likes, leyendo más publicaciones y participando en más temas. Por otro lado, los usuarios menos comprometidos tienen una interacción más esporádica y se centran en un menor número de actividades. Por lo anterior, la clase violeta (0) puede renombrarse como 'usuarios participativos', mientras que la clase turquesa (1) puede nombrarse como 'usuarios observadores'.

El modelo de árbol de decisión puede identificar comentarios con sentimientos positivo con mayor precisión. Esto permite destacar comentarios positivos para crear una atmósfera más positiva en el foro, identificar a los usuarios que contribuyen con comentarios positivos y de valor para la comunidad y priorizar las publicaciones positivas para crear un feed más atractivo y útil para los usuarios.

Analizar los comentarios negativos puede proporcionar información valiosa sobre los temas que generan mayor frustración en los usuarios. Con esta información se podría identificar áreas que requieren mejoras o atención, mejorar la asistencia a los usuarios con problemas técnicos y responder a sus necesidades con mayor eficacia e identificar potenciales problemas que podrían afectar la satisfacción del usuario.

6.1. Métricas

Aplicando el algoritmo K-means sobre el conjunto de datos original completo de 1857 registros y sobre el conjunto de datos sintéticos generados a partir de la red neuronal GAN, se obtienen las siguientes métricas:

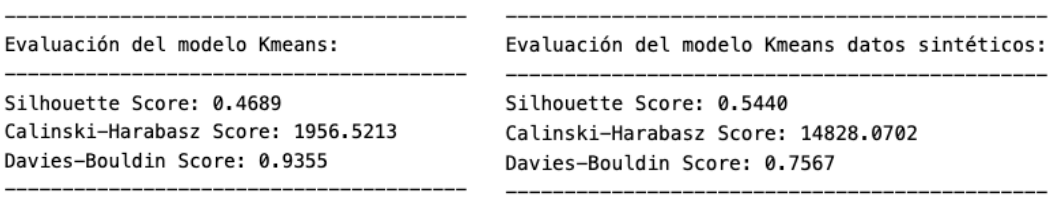

Figura 33: Métricas datos originales y datos sintéticos

De acuerdo con estos resultados obtenidos se puede decir que de acuerdo con el coeficiente de Silhouette los grupos están razonablemente bien separados y que los puntos están bien agrupados dentro de sus propios grupos. De acuerdo con el valor obtenido en el coeficiente de Calinski-Harabasz, que es bastante alto, sugiere que los clústeres están bien definidos y que la dispersión entre los clústeres es baja en comparación con la dispersión entre grupos. El coeficiente de Davies-Boulding es medianamente bueno, indicando que los grupos están bien separados y que la distancia intra-grupo es pequeña.

Para las métricas de negocio se obtienen los siguientes resultados:

Tabla 5: Métricas de negocio

Métrica de negocio	Conjunto de datos original	Conjunto de datos sintético
Tasa de participación	46.15%	36.18%
Tasa de interacción	4 interacciones	3 interacciones
Tasa de retención	70.83%	62.82%

Los resultados del análisis muestran que el modelo de agrupamiento identifica una alta proporción de usuarios participativos tanto en los datos originales como en los sintéticos, aunque con una diferencia notable: en los datos originales se clasificó como participativos un 46.15% de los usuarios, mientras que en los sintéticos solo un 36.18%. A pesar de esta diferencia, la tasa de retención de usuarios participativos en los datos sintéticos (62.82%) se acerca a la observada en los

datos originales (70.83%), lo cual indica que el modelo de datos sintéticos puede capturar de manera eficiente las características que impulsan la retención. En cuanto a la interacción, los usuarios participativos en los datos originales realizan un promedio de 4 acciones, mientras que en los sintéticos realizan 3 acciones, lo que sugiere que los usuarios generados sintéticamente podrían ser ligeramente menos activos. En general, los resultados obtenidos con los datos sintéticos parecen reflejar con precisión las características clave de los usuarios participativos en los datos originales, lo que sugiere que el modelo es útil para explorar diferentes escenarios y estrategias para mejorar la participación y la retención de usuarios.

El resultado de las métricas evaluadas es el siguiente:

```
Modelo: Logistic Regression
Accuracy: 0.9306
Precision: 0.9322
Recall: 0.9306
F1 Score: 0.9303
Matriz de Confusión:
[[232  21   1]
 [  7 232   1]
 [  3   4  32]]
Informe de Clasificación:
              precision    recall  f1-score   support

           0       0.96      0.91      0.94       254
           1       0.90      0.97      0.93       240
           2       0.94      0.82      0.88        39

    accuracy                           0.93       533
   macro avg       0.93      0.90      0.92       533
weighted avg       0.93      0.93      0.93       533
```

```
Modelo: Decision Tree
Accuracy: 0.9306
Precision: 0.9309
Recall: 0.9306
F1 Score: 0.9306
Matriz de Confusión:
[[234  19   1]
 [ 14 225   1]
 [  0   2  37]]
Informe de Clasificación:
              precision    recall  f1-score   support

           0       0.94      0.92      0.93       254
           1       0.91      0.94      0.93       240
           2       0.95      0.95      0.95        39

    accuracy                           0.93       533
   macro avg       0.94      0.94      0.94       533
weighted avg       0.93      0.93      0.93       533
```

```
Modelo: SVM
Accuracy: 0.9193
Precision: 0.9282
Recall: 0.9193
F1 Score: 0.9190
Matriz de Confusión:
[[222  32   0]
 [  2 238   0]
 [  2   7  30]]
Informe de Clasificación:
              precision    recall  f1-score   support

           0       0.98      0.87      0.93       254
           1       0.86      0.99      0.92       240
           2       1.00      0.77      0.87        39

    accuracy                           0.92       533
   macro avg       0.95      0.88      0.91       533
weighted avg       0.93      0.92      0.92       533
```

Figura 34: Métricas de modelos de clasificación

De acuerdo con las métricas obtenidas en los modelos:

*Accuracy***:** Los tres modelos tienen un accuracy similar, alrededor de 0.93.

F1-Score**:** El Árbol de Decisión tiene un F1-score ligeramente superior (0.9306) que la Regresión Logística (0.9303) y el SVM (0.9190).

Rendimiento en Positivos**:** El Árbol de Decisión tiene un mejor recall para la clase positiva (0.95) que la Regresión Logística (0.82) y el SVM (0.77).

En este caso, el Árbol de Decisión parece ser el mejor modelo para tu problema de análisis de sentimientos, ya que tiene un F1-score ligeramente más alto y un mejor recall para la clase positiva. Esto sugiere que el Árbol de Decisión es más preciso en la clasificación de comentarios positivos, lo cual es importante si el objetivo es identificar comentarios positivos correctamente.

6.2. Evaluación cualitativa

La proporción de usuarios participativos identificados en los datos sintéticos (36.18%), es menor que la del conjunto original (46.15%). Esto sugiere que el modelo podría no estar aprendiendo todas las características claves que definen los usuarios participativos.

La tasa de retención en los datos sintéticos (62.82%) es cercana a la del conjunto original (70.83%).

Los resultados del análisis de agrupamiento son útiles para comprender las diferentes clases de usuarios en el foro, permitiendo la implementación de estrategias de marketing y engagement personalizadas.

El análisis de los datos originales y sintéticos sugiere que, a pesar de una tasa de participación moderadamente alta, el foro presenta una interacción relativamente baja, posiblemente debido a su reciente creación (un año). Para aumentar la participación, se recomienda destacar temas principales, implementar encuestas y preguntas para fomentar la interacción entre usuarios, y personalizar notificaciones y correos electrónicos.

Para aumentar la interacción, se propone un sistema de puntos de fidelización que recompense a los usuarios por sus acciones (likes, comentarios, publicaciones).

En cuanto a la retención, se observa una alta tasa en los datos originales (70.83%), lo que sugiere que el foro ofrece contenido de calidad y una buena experiencia de usuario. Los datos sintéticos, con un 62.82% de retención, corroboran la buena calidad del foro, aunque con una ligera disminución.

El bajo sentimiento positivo (7%) dentro del foro puede ser por la naturaleza intrínseca de los problemas discutidos (que son inherententemente desafíos o problemas presentados durante el desarrollo). Además los usuarios tienden a comentar más cuando se están enfrentando a problemas que cuando tienen sentimientos positivos.

Casi la mitad de las publicaciones (46%) tienen un sentimiento neutro. Esto es típico en foros técnicos donde las publicaciones son informativas, centradas en la descripción de problemas, pasos de solución, o discusiones técnicas sin una carga emocional fuerte. Un alto porcentaje de publicaciones neutras es esperable y positivo en un entorno técnico, ya que sugiere que las discusiones están orientadas a la resolución de problemas y la transferencia de conocimiento, en lugar de emociones extremas.

Casi la mitad de las publicaciones (47%) tienen un sentimiento negativo. En el contexto de un foro de discusión técnica, esto puede reflejar la frustración de los usuarios con los problemas que enfrentan, errores persistentes, o dificultades en la implementación de soluciones. Un alto porcentaje de sentimientos negativos puede ser preocupante ya que podría indicar que muchos usuarios están enfrentando problemas serios o persistentes en su trabajo de desarrollo. También puede reflejar la falta de soluciones efectivas o respuestas satisfactorias en el foro.

6.2. Consideraciones de producción

Monitoreo del desempeño: Implementar un sistema de monitoreo que supervise en tiempo real las métricas clave del modelo, como la precisión, la tasa de retención, la interacción y la tasa de usuarios participativos. Implementar un sistema de análisis de errores que permita identificar las causas de los errores del modelo, especialmente aquellos que afectan el rendimiento de las métricas de negocio.

Integración con Streams de Datos: Integrar el modelo con flujos de datos en tiempo real para que pueda procesar información actualizada sobre el comportamiento de los usuarios, las interacciones en el foro y el contenido generado. También es posible implementar una plataforma de procesamiento de datos en streaming (como Apache Kafka o Apache Flink) para gestionar la ingesta y el procesamiento de datos en tiempo real.

Servicios en la nube: Aprovechar los servicios de machine learning en la nube (como Amazon SageMarker, Azure Machine Learning o Google Cloud AI Platform) para simplificar la implementación, el entrenamiento y el despliegue del modelo.

Mantenimiento y actualización: Implementar un proceso para actualizar el modelo de forma regular, utilizando nuevos datos y ajustando sus parámetros para mejorar su precisión y rendimiento.

Referencias

[1] León Guzmán, E. (s.f.). Métricas para la validación de Clustering. Recuperado de https://disi.unal.edu.co/~eleonguz/cursos/mda/presentaciones/validacion_Clustering.pdf.

[2] Fasttrack Company. (s.f.). Calidad de los conglomerados: control de calidad, evaluación de la calidad del análisis de conglomerados. Recuperado de https://fastercapital.com/es/contenido/Calidad-de-los-conglomerados--control-de-calidad--evaluacion-de-la-calidad-del-analisis-de-conglomerados.html

[3] XLSTAT. (s.f.). DBSCAN: Density-Based Spatial Clustering of Applications with Noise. XLSTAT. https://www.xlstat.com/es/soluciones/funciones/dbscan-density-based-spatial-clustering-of-applications-with-noise#:~:text=DBSCAN%20significa%20Agrupamiento%20espacial%20basado,basados%20%E2%80%8B%E2%80%8Ben%20densidad

[4] LinkedIn. (s.f.). How do you compare the performance and scalability of HDBSCAN? LinkedIn. https://www.linkedin.com/advice/1/how-do-you-compare-performance-scalability-hdbscan?lang=es&originalSubdomain=es#:~:text=HDBSCAN%20significa%20Agrupaci%C3%B3n%20espacial%20basada,jer%C3%A1rquica%20de%20aplicaciones%20con%20ruido

[5] Agrupamiento espectral. (2023, 3 de junio). En *Wikipedia, la enciclopedia libre*. https://es.wikipedia.org/wiki/Agrupamiento_espectral

[6] Gallardo, J. A. (s.f.). *Cluster Analysis*. [PDF]. Universidad de Granada. https://www.ugr.es/~gallardo/pdf/cluster-3

[7] Universidad de Oviedo. (s.f.). *K-means*. Universidad de Oviedo. https://www.unioviedo.es/compnum/laboratorios_py/kmeans/kmeans.html

[8] Salcedo, P. C. (s.f.). *Modelo de regresión logística*. En *Estimación de la ocurrencia de incidencias en declaraciones de pólizas de importación*. Universidad Nacional Mayor de San Marcos. Recuperado de https://sisbib.unmsm.edu.pe/bibvirtualdata/tesis/basic/salcedo_pc/enpdf/cap2.pdf

[9] UNIR. (s.f.). Árboles de decisión: qué son y cuál es su uso en Big Data. Recuperado de https://www.unir.net/ingenieria/revista/arboles-de-decision/

[10] MathWorks. (s.f.). Support Vector Machine (SVM). Recuperado el 11 de junio de 2024, de https://la.mathworks.com/discovery/support-vector-machine.html

[11] Machine Learning para Todos. (s.f.). Tratamiento de Clases Desbalanceadas. Recuperado de https://machinelearningparatodos.com/tratamiento-de-clases-desbalanceadas/

[12] Casdelg, M. (2021, 17 de febrero). Cómo identificar y tratar outliers con Python. Medium. https://medium.com/@martacasdelg/c%C3%B3mo-identificar-y-tratar-outliers-con-python-bf7dd530fc3

[13] Prompt. (s/f). Métricas de evaluación de modelos. Recuperado de https://prompt.uno/aprendizaje-automatico/metricas-de-evaluacion-de-modelos/

[14] Chandradip. (2020, December 6). *MinMaxScaler*. Medium. https://medium.com/@chandradip93/minmaxscaler-7ee697b9e89

[15] Softtek. (s.f.). La reducción de dimensionalidad en el machine learning. Recuperado de https://blog.softtek.com/es/la-reduccion-de-dimensionalidad-en-el-machine-learning#:~:text=Por%20tanto%2C%20la%20t%C3%A9cnica%20de,en%20similar%20en%20ambos%20casos

[16] KeepCoding. (s. f.). ¿Qué es el Kernel PCA y ejercicios de aplicación?. Recuperado de https://keepcoding.io/blog/que-es-el-kernel-pca-y-ejercicios-de-aplicacion/