Clasificación y detección de tópicos en Twitter: Caso de estudio Elecciones Presidenciales Colombia 2022

Cesar Ramírez Gómez, *Pontificia Universidad Javeriana*, Luis Gabriel Moreno-Sandoval, *Doctor en Ingeniería, Pontificia Universidad Javeriana*

***Abstract*—En esta investigación, se lleva a cabo un análisis de las temáticas discutidas en Colombia en la red social Twitter durante el mes de junio de 2022. Para ello, se implementan dos estructuras concatenadas basadas en procesamiento de lenguaje natural. La primera, se desarrollaron modelos supervisado aplicando técnicas de aprendizaje de máquina (Naive Bayes, regresión logística y XGBoost) y Transformers (Bert) con el fin de detectar temas globales, como política, economía y cultura. En cuanto a la segunda estructura, se emplearon modelos no supervisados para generar tópicos de temas utilizando la estructura Bertopic, ajustando los hiperparametros del algoritmo UMAP y HDBSCAN. Los resultados experimentales del modelo supervisado arrojan una precisión del 0,81%, un Recall del 0,851% y un F1 Score de 0,872. Por otro lado, el modelo no supervisado Bert obtiene un rendimiento de 0,445 en CV y -1,03 en UMASS. Esta investigación demuestra la capacidad de analizar y comprender las opiniones expresadas en la sociedad digital, así como identificar y segmentar los temas más relevantes. Su aplicación abarca desde modelos predictivos de noticias hasta publicidad dirigida y análisis político.**

***Abstract*—In this research, an analysis of the topics discussed in Colombia on the social network Twitter during the month of June 2022 is carried out. For this purpose, two concatenated structures based on natural language processing are implemented. The first one, based on machine learning techniques and Transformers in order to detect global topics, such as politics, economy and culture. As for the second structure, unsupervised models are employed to generate topic topics using the Bertopic structure, adjusting the hyperparameters of the UMAP and HDBSCAN algorithm. The experimental results of the supervised model yield an accuracy of 0.81% , a Recall of 0.85% and an F1 Score of 0.872. On the other hand, the unsupervised Bert model obtained a performance of 0.445 in CV and -1.03 in UMASS. This research demonstrates the ability to analyze and understand the opinions expressed in the digital society, as well as to identify and segment the most relevant topics. Its application ranges from predictive news models to targeted advertising and political analysis.**

***Index Terms*—Redes sociales digitales, inteligencia artificial, procesamiento de lenguaje natural, Twitter, cluster.**

I. Introducción

La comunicación es un elemento fundamental en todas las sociedades humanas, las cuales, en muchas circunstancias, se ven limitadas por el tiempo y el espacio. El desarrollo tecnológico ha reducido estas barreras, permitiendo que los mensajes lleguen a más personas en menos tiempo. Las redes sociales digitales ha revolucionado la forma de comunicarnos ofreciendo canales de comunicación rápidos y masivos, que democratizan el acceso a la comunicación. En la actualidad, más del 50% de la población global cuenta con acceso a Internet y alrededor del 60 % de la población utiliza redes sociales digitales. Aunque las redes sociales digitales ofrecen numerosos beneficios, también presentan riesgos significativos, como adicción, ambigüedad en el desarrollo personal, violencia de género en línea y ciberacoso, los cuales presentan características distintas a la violencia fuera de línea. En vista de la vasta cantidad de información disponible en Internet y las plataformas de redes sociales en línea, se ha vuelto crucial desarrollar herramientas computacionales basadas en inteligencia artificial para procesar rápidamente la información relevante que afecta a la sociedad. En este proyecto, se desarrollan técnicas de aprendizaje automático para la clasificación de temas y generación de tópicos de temas basadas en BERTopic. Además, se desarrolla una interfaz gráfica que permite explorar los datos en el tiempo.

Esta investigación busca identificar las temáticas de las que se habló en Colombia en la red social Twitter para la categoría política en el mes de junio del 2022 a través del desarrollo de dos estructuras concatenadas, en la primera, se detectan temas globales, como política, economía o cultura, empleando modelos supervisados para la detección de temas a partir de técnicas de aprendizaje de máquina (Naive Bayes, regresión logística y XGBoost), basados en Transformers (Bert) que permiten identificarlos. Para la segunda estructura, se emplearon modelos no supervisados para la generación de tópicos de temas usando la estructura Bertopic basado en Transformers, donde se sintonizó el modelo modificando los hiperparametros del algoritmo UMAP y HDBSCAN en busca de las mejores métricas. Se realizaron 54 variaciones del experimento. Finalmente, con el modelo sintonizado se define una línea de tiempo para visualizar la creación, desarrollo y conclusión de tópicos.

Esta investigación muestra la capacidad de analizar y comprender las opiniones expresadas en la sociedad digital, así como la posibilidad de identificar y segmentar los temas más relevantes. Su aplicación abarca desde modelos predictivos de noticias hasta publicidad dirigida y análisis político.

II. Marco teórico

La base de todas las instituciones humanas radica en la comunicación al implicar relaciones sociales por medio de las cuales se genera una concentración de individuos homogéneos en cuanto a creencias, paradigmas, cultura y cosmovisiones [1]. Los canales de comunicación están limitados por tiempo

(envío-recepción del mensaje) y espacio (distancia entre los interlocutores). El desarrollo tecnológico ha reducido las brechas en la comunicación, permitiendo que los mensajes lleguen a más personas en un menor tiempo [2]. La llegada de las redes sociales digitales marcó un punto de inflexión, facilitando canales de comunicación rápidos, masivos y horizontales entre usuarios, democratizando el acceso a la comunicación, lo cual permitió compartir y recibir información de manera instantánea y a gran escala [3] [4]. Ahora bien, esta forma de comunicación se caracteriza por un gran volumen de datos que combina estructuras escritas y audiovisuales [5], en donde, generalmente, el individuo consume y proporciona información personal o grupal sin ningún proceso de selección que permita ratificar la veracidad o el sentido ético de los datos [6] .

Al no estar restringida esta nueva forma de comunicación digital por distancias demográficas, el rango de conexiones es global, lo que permite el intercambio social y cultural de forma indiscriminada para los usuarios. Así, las redes sociales digitales satisfacen la necesidad fundamental de comunicación en el ser humano [7].En este sentido, la representación de la integración de usuarios se logra mediante la utilización de sociogramas, los cuales consisten en puntos interconectados por líneas que simbolizan las relaciones existentes. El núcleo de una red social puede ser heterogéneo [8], pero comparándolo con métodos tradicionales de comunicación, el poder de los integrantes no representa el eje central de las sociedades construidas, sino el número de usuarios que las integran e interactúan en ellas [9]. Cuantos más miembros tiene una comunidad de usuarios, mayor valor tiene para un miembro pertenecer a ella [4][8].

Actualmente, más de la mitad de la población mundial, específicamente el 64,4 %, se encuentra conectada a Internet, lo que equivale a aproximadamente 5160 millones de usuarios [10]. En cuanto a las redes sociales, se estima que hay alrededor de 4760 millones de usuarios en todo el mundo, lo que representa cerca del 60% de la población mundial. Meta, la empresa matriz de Facebook, reporta tener 2 958 000 millones de usuarios activos mensuales, lo que equivale a casi el 37% de la población mundial [11]. Por su parte, YouTube atrae a más de 2500 millones de usuarios mensuales, según cifras de sus recursos publicitarios. Instagram también ha experimentado un crecimiento significativo y cuenta con 2000 millones de usuarios activos mensuales. WhatsApp, propiedad de Meta, atrae a 2000 millones de usuarios activos diarios, lo que sugiere que su cifra de usuarios mensuales es aún mayor [11]. En el quinto lugar se encuentra WeChat, con más de 1300 millones de usuarios activos mensuales.

El consumo exponencial de la información hallada en Internet, como noticias, blogs, artículos científicos, archivos multimedia y redes sociales, ha generado la necesidad de procesar en tiempo mínimo la información relevante que afecta a la sociedad. Ante esta problemática, nuevas herramientas computacionales basadas en inteligencia artificial son las indicadas para procesar la basta cantidad de datos.

A. *Detección de temas*

La detección de temas es un conjunto de técnicas que tiene como objetivo identificar y comprender los temas o tópicos principales en un grupo de textos [12]. Con la explosión de datos en la web, las redes sociales digitales y otras fuentes, resulta difícil y costoso analizar manualmente todo el contenido. Mediante la aplicación de técnicas de procesamiento de lenguaje natural (NLP) se pueden identificar automáticamente los temas principales en un conjunto de documentos, facilitando la búsqueda, el filtrado y la recuperación de información relevante [43]. Otro aspecto importante es la capacidad de detectar temas emergentes o tendencias en tiempo real. La información en línea se actualiza constantemente y las técnicas de detección de temas permiten identificar rápidamente nuevos tópicos que están ganando relevancia en un determinado dominio. Esto puede ser valioso para la detección de noticias o eventos importantes, así como para la identificación de crisis o problemas potenciales en la opinión pública [13].

A continuación se presentan los Modelos basados en aprendizaje de máquina.

1) Modelo de Naive Bayes: algoritmo de clasificación basado en el teorema de Bayes y la suposición de independencia entre las características. Se calcula la probabilidad de que un ejemplo pertenezca a una clase específica dadas sus características. Durante la clasificación, se selecciona la clase con la mayor probabilidad como la etiqueta de clasificación. Emplea estimadores de máxima verosimilitud o técnicas de suavizado, como el suavizado de Laplace, para calcular las probabilidades condicionales y las probabilidades previas a partir de los datos de entrenamiento [14]. Aunque es una suposición simplificada, el modelo puede ser efectivo y eficiente en la clasificación de datos.

2) Regresión logística: modelo empleado para resolver problemas de clasificación binaria. Calcula la probabilidad de pertenencia a alguna de las categorías. Utiliza una función logística para mapear una combinación lineal de las características del ejemplo en un rango entre 0 y 1 [15]. Los coeficientes del modelo se ajustan durante el entrenamiento para maximizar la verosimilitud de los datos de entrenamiento. Durante la clasificación, si la probabilidad estimada supera un umbral, se asigna la etiqueta de la clase positiva; de lo contrario, se asigna la etiqueta de la clase negativa [16].

3) XGBoost: basado en árboles de decisión aplicado para problemas de clasificación como de regresión. Emplea un enfoque de boosting en el cual se construyen árboles de decisión secuencialmente de manera iterativa. Cada árbol se ajusta a los residuos del árbol anterior, lo que permite corregir los errores cometidos en las predicciones previas [17]. Matemáticamente, se optimiza una función objetivo que combina la pérdida del modelo y una regularización para evitar el sobre ajuste.

A continuación se presenta el modelo basado en Transformers:

4) BERT: modelo de lenguaje basado en redes neuronales llamadas Transformers [18]. Usa una arquitectura de codificación que se compone de múltiples capas de atención y transformación. La atención es un mecanismo clave en el cual se calculan las ponderaciones de importancia para cada palabra o token en función de su relación con las demás palabras en

la secuencia. La representación de cada palabra se obtiene combinando información de dos direcciones: de izquierda a derecha y de derecha a izquierda, gracias a la atención bidireccional que permite que cada palabra tenga acceso a la información contextual tanto anterior como posterior a ella [19]. Esta propiedad bidireccional mejora la comprensión del contexto y ayuda a capturar las dependencias entre las palabras.

BERT aprende a asignar un vector de representación a cada palabra o token en función de su contexto circundante. Estos vectores, también conocidos como embeddings, son números reales de alta dimensionalidad que capturan información semántica y sintáctica de las palabras. Durante el entrenamiento, BERT utiliza un enfoque de aprendizaje supervisado donde se ajustan los pesos (parámetros) de la red neuronal para minimizar una función de pérdida. Esto se logra mediante algoritmos de optimización, como el descenso del gradiente que actualiza los parámetros de la red en función del gradiente de la función de pérdida [20]. Una vez que BERT ha sido preentrenado con un corpus de texto grande, se puede utilizar para diversas tareas de procesamiento de lenguaje natural. Para adaptarlo a una tarea específica se agrega una capa adicional de clasificación o generación de texto y se ajustan los pesos finos específicos para esa tarea. Esto se realiza mediante el entrenamiento supervisado con datos etiquetados [21].

5) Métricas de evaluación para modelos supervisados: Se emplean las siguientes métricas para los modelos supervisados mencionados anteriormente:

• Precisión: proporción de instancias positivas correctamente clasificadas sobre el total de instancias clasificadas como positivas [22]. Se calcula dividiendo el número de verdaderos positivos (TP) entre la suma de verdaderos positivos y falsos positivos (FP).

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

• Recall (sensibilidad): proporción de instancias positivas correctamente clasificadas sobre el total de instancias verdaderamente positivas [22]. Se calcula dividiendo el número de verdaderos positivos (TP) entre la suma de verdaderos positivos y falsos negativos (FN).

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

• F1 score: combina la precisión y el recall en un solo valor, proporcionando un equilibrio entre ambas métricas [22]. Se calcula mediante la media armónica de la precisión y el recall.

$$F1score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3}$$

B. Generación de tópicos de temas

Este grupo de técnicas de aprendizaje automático identifica patrones ocultos en un texto mediante la agrupación de palabras que presentan una mayor frecuencia con relación a un tema, se utiliza para identificar tópicos latentes en los textos como, por ejemplo, los discursos de políticos, tendencias en las redes sociales o las conversaciones de grupos de investigación [23]. La generación de tópicos de temas se lleva a cabo mediante métodos basados en estadística frecuentista o en probabilidad bayesiana. La medición de coherencia se realiza mediante diferentes técnicas, como el juicio humano o el análisis cuantitativo [24]. El resultado del modelado de temas está estrechamente relacionado con el vocabulario que se maneja en el entorno del texto y, por tanto, se considera un proceso generativo.

A continuación se presentan las técnicas más representativas:

1) Análisis de la Semántica Latente (LSA): La técnica utilizada para disminuir la dimensionalidad de una matriz término-documento, que indica la frecuencia de cada término en cada texto, se basa en la descomposición de valores singulares (SVD) [25]. Primero, se construye una representación vectorial de los documentos basada en esta matriz. Luego, se aplica SVD para descomponer la matriz en términos latentes, documentos latentes y valores singulares [26]. Los valores singulares determinan la información conservada en la matriz reducida. Las dimensiones latentes representan características semánticas inferidas de la coocurrencia de términos en los documentos [27]. La matriz de términos latentes representa los términos y documentos en un espacio semántico reducido, donde la similitud se mide mediante la distancia coseno [28]. Esto permite identificar patrones y relaciones semánticas ocultas en el corpus, realizando tareas como recuperación de información, clasificación de documentos y generación automática de resúmenes.

2) Asignación latente de Dirichlet (LDA): Método de indexación y recuperación automática de información que permite la identificación de los términos y los documentos en espacios de dimensionalidad reducida [29]. La técnica utiliza la representación vectorial de los documentos, basada en la frecuencia de un término, y aplica una proyección lineal para reducir la dimensionalidad del espacio original. De esta manera, se obtienen los tópicos básicos de los documentos. El proceso comienza con la elección de una distribución de tópicos para cada documento y una distribución de palabras para cada tópico. Luego, para cada palabra en cada documento, se selecciona un tópico a partir de la distribución de tópicos y, posteriormente, se selecciona una palabra para esa posición en el documento a partir de la distribución de palabras del tópico seleccionado [30].

La idea detrás de LDA es que los tópicos latentes sean responsables de la coocurrencia de palabras en los documentos. Si dos palabras tienden a aparecer juntas con frecuencia, entonces es probable que pertenezcan al mismo tópico [31].

3) BERTOPIC: Este es un modelo que extrae tópicos de temas coherentes en documentos mediante modelos de lenguaje natural basados en transformadores preentrenados que generan representaciones incrustadas de los documentos [32]. Estas incrustaciones se agrupan y luego se producen representaciones temáticas utilizando un enfoque de TF-IDF basado en clases. Su estructura consta de tres pasos. En primer lugar, cada documento se convierte en su representación incrustada utilizando un modelo de lenguaje preentrenado. Posteriormente, se disminuye la dimensionalidad de las embeddings resultantes con el objetivo de mejorar el proceso de

clustering. Por último, se extraen las representaciones de temas a partir de los grupos de documentos utilizando una variación basada en clases de TF-IDF [32] [33].

BERTopic se ha desarrollado como un modelo de tema que amplía el enfoque de incrustación de clúster mediante el uso de modelos de lenguaje de última generación y la aplicación de un procedimiento TF-IDF basado en clases para generar representaciones temáticas. Esta separación del proceso de agrupamiento de documentos y generación de representaciones temáticas proporciona una mayor flexibilidad y facilidad de uso [34]. Una ventaja de BERTopic es que no es necesario especificar previamente la cantidad de temas, ya que puede extraer la cantidad de temas descritos en los documentos.

a) Agrupación de documentos:

Se convierten los documentos en una representación numérica empleando modelos de sentence-transformers, optimizados para la similitud semántica. Esto permite crear representaciones espacio-vectoriales en las cuales estructuras semánticamente iguales componen un tema [32].

Existen dos modelos predeterminados:

• ”all-MiniLM-L6-v2”: modelo en inglés entrenado específicamente para tareas de similitud semántica [27].

• ”paraphrase-multilingual-MiniLM-L12-v2”: modelo multilingüe para más de 50 idiomas.

b) Reducción de dimensionalidad:

Emplea una técnica de reducción de dimensionalidad de los datos, en este caso UMAP (Uniform Manifold Approximation and Projection) [35] algoritmo altamente flexible y no lineal. El propósito central es capturar la compleja y variada estructura de los datos al encontrar una representación de dimensiones reducidas que conserve la estructura topológica esencial de dicho conjunto de datos.

El algoritmo UMAP se construye a partir de dos fases principales que permiten su funcionamiento. En la primera, se realiza la construcción de una representación topológica difusa. Este proceso implica calcular las fuerzas de pertenencia de conjuntos borrosos para los vecinos más cercanos de cada punto en el espacio de alta dimensión [36]. Debido a que las fuerzas de pertenencia de los conjuntos borrosos disminuyen rápidamente a medida que los puntos se alejan, solo es necesario calcular estas fuerzas para los vecinos más cercanos, lo que reduce significativamente la complejidad computacional.

En la segunda fase se lleva a cabo la optimización de la incrustación de baja dimensión. Se utiliza el descenso de gradiente estocástico, que es un método de optimización eficiente. Se utiliza una aproximación suave de la función de fuerza de membresía en la incrustación de baja dimensión. Esta aproximación se elige de una familia de curvas versátil y conveniente. Además, se aplica el muestreo negativo similar al utilizado en algoritmos como word2vec y LargeVis. La incrustación de baja dimensión se inicia utilizando técnicas de incrustación espectral basadas en el laplaciano de la representación topológica, lo que ayuda a obtener una representación inicial de alta calidad.

Al combinar estas dos fases, el algoritmo UMAP logra ser rápido y escalable, al tiempo que se fundamenta en una sólida teoría matemática. Su objetivo principal es preservar la estructura topológica esencial de los datos en una representación de baja dimensión [37].

Los hiperparámetros con los cuales se experimentaron fueron:

• Número de vecinos (n neighbors): determina cómo se equilibra la importancia de la estructura local y global en los datos. Se logra al limitar el tamaño del vecindario local que UMAP considerará al intentar aprender la estructura compleja de los datos. Cuando los valores son bajos, se enfoca en estructuras locales, mientras que al aumentar los valores, examina vecindarios más extensos para capturar la complejidad de los datos, aunque a costa de perder algunos detalles finos en pos de una visión más general.

• Distancia mínima (min dist): distancia mínima que se permite entre puntos en la representación de baja dimensión. Valores bajos generarán incrustaciones más densas. Esto puede ser útil si se desea resaltar la formación de grupos o una estructura topológica más detallada.

• Número de componentes (n components): Determina la dimensionalidad del espacio de dimensiones reducidas en el que se incrustarán los datos. A diferencia de otros algoritmos de visualización como t SNE, UMAP es eficiente en términos de escalabilidad en la dimensión de incrustación, lo que significa que se puede utilizar para visualizaciones más allá de las dos o tres dimensiones.

• Métrica: controla cómo se calcula la distancia en el espacio ambiental de los datos de entrada. Se define para todos los experimentos la métrica angular y de correlación Coseno.

c) Agrupación (Clustering):

Se realiza un proceso de agrupamiento para identificar grupos de incrustaciones similares y extraer tópicos de temas. Se emplea HDBSCAN (Density-Based Spatial Clustering of Applications with Noise) [38] que es un algoritmo de agrupamiento que se basa en el concepto de densidad para identificar agrupamientos en conjuntos de datos.

A diferencia de otros algoritmos de agrupamiento, HDBSCAN no se define el número de clusters de antemano, lo cual lo hace especialmente útil en conjuntos de datos donde la estructura de los clusters no es conocida previamente. HDBSCAN es capaz de descubrir clusters de diferentes estructuras identificando puntos de datos que no pertenecen a ningún cluster, conocidos como ”ruido” [39]. El algoritmo HDBSCAN funciona de la siguiente manera [40] [41]:

• Calcula la densidad de cada punto de datos en función de la distancia de sus vecinos más cercanos. Esta medida de densidad se utiliza para determinar la importancia relativa de cada punto en el total de datos.

• Utiliza la información de densidad para construir un árbol jerárquico de agrupamientos llamado Árbol de estabilidad.

• Aplica un algoritmo de corte en el Árbol de estabilidad para extraer un agrupamiento plano. Este corte se realiza considerando la estabilidad de los agrupamientos en diferentes niveles del árbol. Los agrupamientos más estables se conservan como clusters, mientras que los agrupamientos menos estables se consideran ruido.

• La estabilidad de los agrupamientos se determina mediante un coeficiente llamado Coeficiente de estabilidad. Este coeficiente combina la densidad de los puntos y la distancia relativa entre ellos para medir cuán estables son los agrupamientos en

diferentes niveles del Árbol de estabilidad.

4) Bolsa de palabras: La representación de bolsa de palabras se construye mediante el cálculo de la frecuencia de aparición de cada palabra clave en cada grupo [42]. Esta está a nivel de grupo y no a nivel de documento, lo que significa que se centra en las palabras clave relevantes para cada tema o grupo en lugar de considerar todo el contenido del documento.

5) Representación tópica: Para establecer la representación temática por tema se emplea la técnica C-TF-IDF (Class-based Term Frequency-Inverse Document Frequency) variante del algoritmo TF-IDF, adaptado para su uso con la técnica de agrupación de temas [43]. Cada grupo se trata como una única clase o categoría en lugar de considerar documentos individuales. Se busca determinar qué palabras son características de un grupo en particular y no tanto de los otros grupos.

En C-TF-IDF se genera una representación basada en clases al combinar todos los documentos de un grupo en un solo documento representativo. Luego, se calcula la frecuencia de cada palabra clave dentro de esa clase específica. Esta representación se normaliza en L1 para considerar las diferencias en los tamaños de los grupos [44].

Para obtener puntuaciones de importancia por palabra en cada clase se realiza una transformación utilizando el logaritmo de uno más el número promedio de palabras por clase, dividido entre la frecuencia de la palabra clave en todas las clases. Al agregar uno dentro del logaritmo, se garantiza que los valores sean positivos [32]. Finalmente, se multiplica la representación de la frecuencia de término (TF) por la representación inversa de la frecuencia de documento (IDF) para obtener la puntuación de importancia por palabra en cada clase.

$$W_{X,C} = \|tf_{X,C}\| * log\left(1 + \frac{A}{f_x}\right) \tag{4}$$

$$\|tf_{X,C}\| = frecuencia\ de\ la\ palabra\ X\ en\ la\ clase\ C \tag{5}$$

$$f_x = frecuencia\ de\ la\ palabra\ X\ en\ todas\ las\ clases \tag{6}$$

$$A = promediodepalabrasporclase \tag{7}$$

6) Modelado dinámico - temporal de tópicos: Técnica de Bertopic que analiza cómo los temas cambian a lo largo del tiempo, para esto, calcula la representación de los temas en cada paso de tiempo sin tener que ejecutar el modelo completo secuencialmente. Primero, se sintoniza BERTopic sin considerar el aspecto temporal utilizando la representación global se capturan los temas principales que probablemente estén presentes en diferentes intervalos de tiempo. A continuación, se calcula la representación c-TF-IDF para cada tema en cada paso de tiempo proporcionando una representación específica del tema.

Se ajustan dos parámetros: ajuste global y ajuste evolutivo. El primer parámetro se obtiene al promediar la representación c-TF-IDF del tema en el paso de tiempo t con la representación global. Esto permite que la representación de cada tema se desplace proporcionalmente hacia la representación global manteniendo algunas de las palabras específicas. Mientras que el ajuste evolutivo implica promediar la representación c-TF-IDF del tema en el paso de tiempo t con la representación c-TF-IDF en el paso de tiempo t-1. Esto se realiza para cada representación de tema, lo que permite que las representaciones evolucionen a lo largo del tiempo.

7) Métricas de evaluación para modelos no supervisados: Se emplean las siguientes variables cuantitativas enfocadas en análisis no supervisado de temas:

• Medición intrínseca C UMASS: se utiliza para evaluar la coherencia de un conjunto de palabras en un contexto específico. Esta métrica compara una palabra con las palabras precedentes y posteriores en un conjunto ordenado de palabras [45]. Utiliza una función de puntuación por pares que se basa en la probabilidad de logaritmo condicional empírico con un suavizado de recuento para evitar el cálculo del logaritmo de cero.Esta función de medida no es simétrica y es creciente en función de la probabilidad empírica. Al ser una función creciente de la probabilidad empírica no es simétrica en p(wj—wi)p(wj—wi), donde el wi es más común que el wj, siendo las palabras ordenadas por la frecuencia decreciente p(w—k)p(w—k). Cuando UMass se acerca a cero, indica una coherencia óptima, y fluctúa a ambos lados de cero.

• Vector de coherencia CV: medida utilizada para evaluar la coherencia de un conjunto de palabras clave o términos. Se basa en la segmentación de las palabras clave y utiliza medidas de confirmación indirecta y la similitud coseno. Para calcular el CV, i) se recopilan los recuentos de co-ocurrencia de las palabras clave mediante una ventana deslizante (estos recuentos se utilizan para calcular el NPMI de cada palabra clave en relación con las demás, lo que resulta en un conjunto de vectores, uno por cada palabra clave) ii) se realiza la segmentación de las palabras clave, lo que implica calcular la similitud entre cada vector de palabras clave y la suma de todos los vectores [46] [47], esta similitud se mide utilizando el coseno. La coherencia se calcula como el promedio de estas similitudes. El CV proporciona una medida de la coherencia temática en el conjunto de palabras clave. Un valor de CV cercano a 1 indica una coherencia óptima.

III. Metodología de la investigación

Con el objetivo de desarrollar un sistema de detección de tópicos a lo largo del tiempo, se han creado dos módulos. El primero de ellos utiliza modelos supervisados para detectar los tópicos y determinar si pertenecen a las categorías de cultura, deportes, economía, política, tecnosfera o vida. Para su desarrollo, se han implementado distintas etapas, como: preprocesamiento de datos, experimentación y selección del modelo más adecuado. En cuanto al segundo módulo, se emplearon modelos no supervisados para la generación de tópicos de temas mediante la implementación de un algoritmo de cluster automático. A su vez este módulo incluye etapas de preprocesamiento de los datos, experimentación, selección del modelo y aplicación en líneas de tiempo. El Pipeline de la implementación de esta investigación es la siguiente: se obtienen los datos de la plataforma Twitter [48], correspondientes al país Colombia durante el año 2023., Posteriormente, se determina la categoría a la que pertenece cada tweet utilizando técnicas de aprendizaje supervisado. A continuación, se agrupan los tweets por categoría para identificar la generación de

temas utilizando el modelo Bertopic ajustado. Finalmente, se presentan los resultados en una línea de tiempo interactiva que resalta las palabras clave en cada período de tiempo.

A. Preprocesamiento

Se realizaron dos etapas de preprocesamiento de texto para el idioma español [49] aplicado en la detección de temas y la generación de tópicos de temas. La primera etapa se encargó de limpiar el texto de diversos elementos no deseados. El alcance de esta etapa incluye la eliminación de: emojis, menciones de usuarios, hashtags, enlaces web, acentos y caracteres especiales. También se eliminan números, saltos de línea y espacios en blanco adicionales. Este proceso de limpieza se realiza utilizando expresiones regulares y funciones proporcionadas por las bibliotecas re [50], demoji [51], unidecode [52] y nltk [53]. La segunda etapa es la lematización, proceso de reducción de las palabras a su forma base o lema. Esto implica expresar las palabras en infinitivo, lo que permite reducir la variación morfológica y simplificar el análisis del texto, esto se logró utilizando la librería spaCy [54] en el idioma español. El proceso de pre procesamiento se utiliza en la etapa de entrenamiento de los modelos supervisados y no supervisados, además de ser usado en la etapa de producción del Pipeline en los tuits en una línea temporal.

B. Base de datos modelo supervisado

Para entrenar el modelo supervisado de temas, se utilizó como fuente de datos la página web de noticias El Tiempo [55], diario colombiano establecido el 30 de enero de 1911 por Alfonso Villegas Restrepo. Actualidad posee una amplia circulación en Colombia en la actualidad, este periódico constituyo una base de datos invaluable para el presente estudio.

Las noticias recopiladas de El Tiempo se agruparon en las siguientes clases: deportes, política, cultura, vida, economía y tecnosfera. Se abarcaron los años comprendidos entre el 2020y el 2023, obteniendo los siguientes volúmenes de datos por año:

- 2020: 3487 noticias
- 2021: 8010 noticias
- 2022: 8887 noticias
- 2023: 2116 noticias

En el transcurso de los años considerados, se observó que la categoría de deportes presentó la mayor cantidad de noticias acumuladas alcanzando un porcentaje del 47,99%. En segundo lugar, se ubicó la categoría de cultura con un 34,07%, seguida de política en tercer lugar con un 7,70%. La categoría de economía ocupó el cuarto puesto, representando un 5,16% de las noticias acumuladas. En quinto lugar, se encontró la categoría de vida, con un porcentaje de 4,01%. Por último, la tecnosfera registró el porcentaje más bajo de noticias acumuladas, con un 1,07%.

Fig. 1. Gráfica de noticias por años

A continuación, se presenta el análisis de la frecuencia de palabras en todas las clases, expuesto en la tabla 1:

TABLE I
ESTADÍSTICAS DE PALABRAS POR CLASE

Clase	Mínimo	Promedio	Máximo
Cultura	20	316	2986
Deportes	19	204	2525
Economía	31	309	1760
Política	15	291	3145
Tecnosfera	50	300	1227
Vida	38	379	2561

En relación con la frecuencia promedio, es notable que la clase Vida se distingue por presentar un valor más elevado en comparación con las demás categorías. Esto implica que la temática de Vida engloba un espectro lingüístico más amplio dentro de las noticias analizadas. Al analizar, la frecuencia de las palabras por clase, se encuentra que: en la categoría de Cultura, la palabra más recurrente es ”poder”, con 17,366 apariciones. En la clase de Deportes, la palabra más repetida es ”partido”, con 16,486 ocurrencias. En Economía, se destaca la palabra ”poder”, con 2,578 apariciones. En cuanto a la clase de Política, la palabra más frecuente es ”Petro”, con 3,568 ocurrencias. En la categoría de Tecnosfera, la palabra más repetida es ”nuevo”, con 385 apariciones. Por último, en la clase de Vida, la palabra ”más” se presenta con mayor frecuencia, con 3,487 apariciones. Se examina el comportamiento de las noticias durante el año 2022, desglosado por meses, a través de una representación gráfica. Se observa que las categorías predominantes a lo largo de todo el período son deportes y cultura, las cuales muestran una intersección que culmina en el mes de junio.

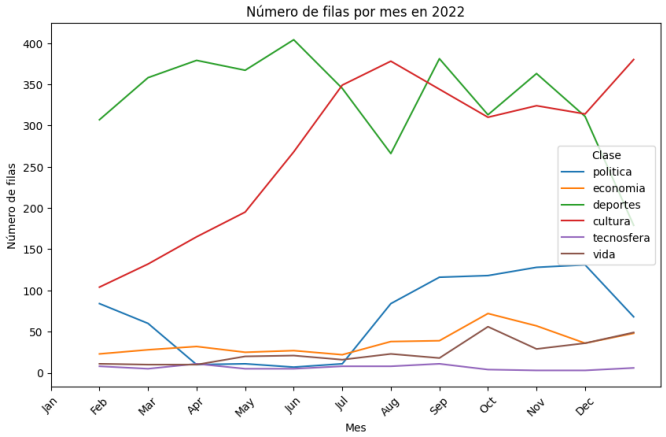

Fig. 2. Comportamiento noticias en el 2022

C. Balanceo de clases

El conjunto de datos original presentaba un desequilibrio significativo, con una gran cantidad de observaciones en las clases ”deportes” y ”cultura”, mientras que las clases ”vida” y ”tecnosfera” contaban con un número muy reducido de observaciones; la cantidad de observaciones para el periodo evaluado fue: deportes (10,798), cultura (7,665), política (1,733), economía (1,160), vida (903) y tecnosfera (241). Para abordar este estudio, se aplicaron dos técnicas de muestreo: SMOTE y RandomUnderSampler. SMOTE (Synthetic Minority Oversampling Technique) [56] consiste en una técnica de sobremuestreo que genera nuevas observaciones sintéticas para la clase minoritaria a partir de las existentes. Por otro lado, Random Under Sampler [57] es una técnica de submuestreo que reduce la cantidad de observaciones en la clase mayoritaria mediante la eliminación aleatoria de algunas observaciones. Se estableció un total de 1000 observaciones por clase para la etapa de balanceo.

IV. EXPERIMENTACIÓN Y RESULTADOS

A. Experimentación de modelos para la detección de temas

Se aplicaron los siguientes métodos de aprendizaje automático: Naive Bayes, regresión logística, XGBoost y el modelo basado en Transformers Bert con la finalidad de comparar su rendimiento en las métricas de precisión, recall y F1 score.

El conjunto de datos se dividió en un 80% para el entrenamiento, lo que equivale a 4200 datos, y un 20% para las pruebas, lo cual representa 1800 datos. Para crear la bolsa de palabras, se seleccionaron las 10,000 palabras más relevantes del texto, junto con monogramas, bigramas y trigramas.

a) Modelos de aprendizaje automático:

• Naive Bayes: se definió un algoritmo de clasificación con distribución multinomial.

• Regresión logística: se crea un objeto de regresión logística utilizando el solucionador ’liblinear’.

• XGBoost : se crea un objeto de clasificador utilizando la clase XGBRFClassifier de la biblioteca XGBoost. Los hiperparámetros definidos son: Tasa de aprendizaje utilizada por el algoritmo (0.1). Número de iteraciones (1000). Profundidad máxima de cada árbol (10). Proporción de características a considerar en cada árbol (0.7).

b) Modelo Transformers Bert:

Se utilizó un codificador universal de oraciones llamado “universal-sentence-encoder-cmlm/multilingual-basemodelo”, capaz de trabajar con más de 100 idiomas, en este se seleccionó el idioma español. Para convertir el texto en vectores de alta dimensión se cargaron las capas de preprocesamiento y codificación desde TensorFlow Hub creando una función sencilla para obtener las representaciones del texto de entrada.

Se estructuró un modelo que consta de las capas de preprocesamiento y codificación, seguidas de una capa de eliminación aleatoria “dropout” con un valor de 0.2 y una capa densa con una función de activación Softmax. Se entreno el modelo durante 20 épocas, paralelamente se utilizó la función “EarlyStopping” para monitorear la pérdida de validación durante el entrenamiento, si la métrica no mejora durante al menos 3 épocas, el entrenamiento se detiene y se restauran los pesos de la época en la que la pérdida de validación presento el mejor valor.

La tabla II se exponen los resultados de los modelos:

TABLE II
RESULTADOS DE LOS MODELOS

Modelo	Precisión (%)	Recall (%)	F1 Score
Naive bayes	0,89	0,88	0,89
Regresión logística	0,91	0,91	0,91
XGBoost	0,86	0,86	0,86
Bert	0,87	0,87	0,88

Como se observa, los resultados presentan un buen desempeño en las tres métricas, sin embargo, la regresión logística obtiene el mejor resultado y XGBoost el menor rendimiento. No obstante, los modelos de aprendizaje automático presentan un desequilibrio en las métricas para cada clase; esto se debe a que su estructura no puede capturar eficientemente la distribución lingüística, especialmente en las clases con valores más bajos, como ”vida” y ”tecnosfera”, que tienen una generación sintética en la etapa de balanceo de clases. El modelo Bert muestra una estabilidad en las métricas individuales para cada clase. A continuación, se presentan gráficas de pérdida, balance de recall, balance de precisión y balance de F1 Score durante las 20 épocas del entrenamiento para este último modelo.

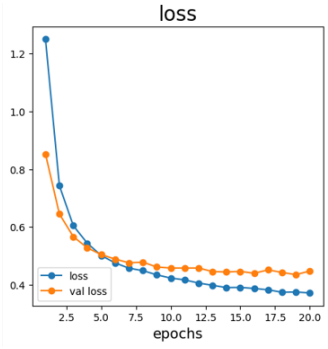

Fig. 3. Perdida por épocas

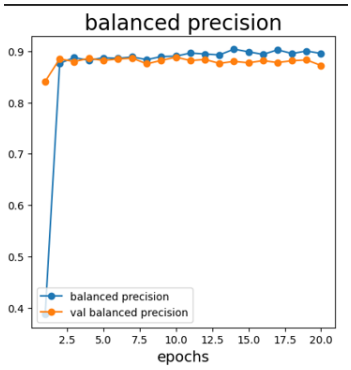

Fig. 4. Balance de precisión

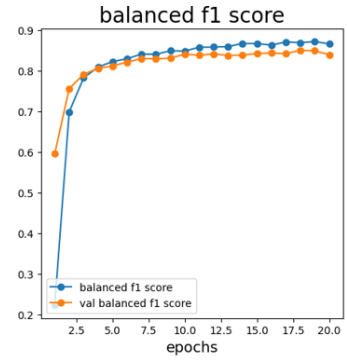

Fig. 5. F1 score

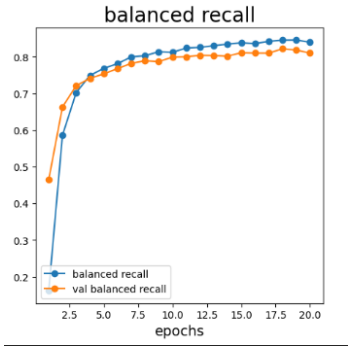

Fig. 6. Balance de Recall

Se evidencia como el rendimiento de todas las métricas se mantiene en un valor constante desde la época 13.

En conclusión, se seleccionó el modelo Bert debido a su equilibrio en las métricas de precisión, recall y puntuación F1 tanto globales como en cada clase individual. Este modelo muestra una falta de sesgo entre las diferentes clases, un aspecto es importante para el análisis imparcial de los datos. Además, en trabajos futuros, se puede mejorar su rendimiento al cargar modelos preentrenados desde TensorFlow Hub y ajustarlos según las necesidades específicas, como la clasificación, incrustaciones, arquitectura, idioma, entre otros.

B. Aplicación modelo supervisado

Utilizando el modelo BERT, se clasificó una base de datos compuesta por 1,801,635 tuits de usuarios colombianos durante el mes de junio de 2022. Se generó la siguiente frecuencia por categorías: cultura (704,947 tuits), política (361,315 tuits), deportes (243,031 tuits), vida (222,994 tuits), tecnosfera (156,864 tuits) y economía (112,484 tuits).

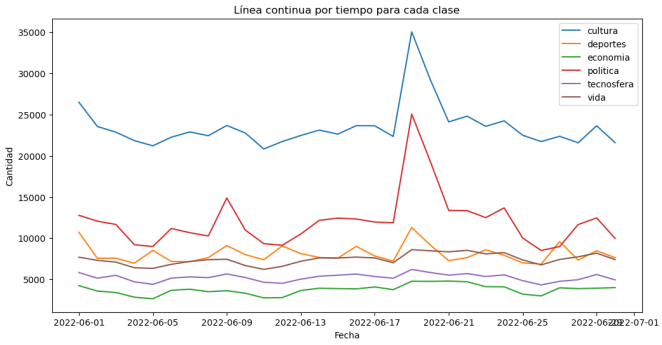

Fig. 7. Comportamiento Tuits en el 2022

Los resultados se mantienen congruentes al considerar la distribución de noticias y tuits generados, siendo las mismas clases predominantes en ambos experimentos.

C. Experimentación para la Generación de tópicos de temas

Durante el proceso de construcción del modelo se evaluaron los siguientes hiperparametros que componen la librería Bertopic comparando las variables CV y UMASS. El conjunto de datos está constituido por de tuits relacionados al país Colombia en el mes de junio y la clase política, a este conjunto de datos se le aplicó la técnica de pre procesamiento y lematización. A continuación, se describen los parámetros empleados durante la experimentación:

I. Agrupación de documentos: se emplea el modelo "paraphrase-multilingual-MiniLM-L12-v2" para el idioma español en todos los experimentos.

II. Reducción de dimensionalidad: se establecen los hiperparametros del algoritmo UMAP en las siguientes combinaciones:

a. Número de vecinos: 10, 15 y 20.
b. Distancia mínima: 0.0 y 0.1.
c. Numero de componentes: 2, 5 y 8.
d. Métrica: se define Coseno.

III. Agrupación: se configura el algoritmo HDBSCAN modificando el tamaño mínimo del cluster en 10, 15 y 20.

IV. Bolsa de palabras: se emplea la librería Count Vectorizer [58] para el idioma español.

V. Numero de temas: "Auto" determina independientemente el número máximo de temas basado en el algoritmo HDBSACN para todos los experimentos

VI. N gramas: se define la detección de monogramas, bigramas y trigramas para todos los experimentos.

Se realizaron 54 experimentos con los siguientes resultados agrupados por el tamaño mínimo del cluster:

Tamaño mínimo del cluster = 10

TABLE III
RESULTADOS TAMAÑO MÍNIMO DEL CLUSTER = 10

#	n_neighbors	n_components	min_dist	CV	UMASS
1	10	2	0	0,439	-0,695
2	10	2	0,1	0,442	-0,993
3	10	5	0	0,415	-0,65
4	10	5	0,1	0,424	-1,044
5	10	8	0	0,426	-1,04
6	10	8	0,1	0,437	-1,028
7	15	2	0	0,415	-1,045
8	15	2	0,1	0,435	-0,932
9	15	5	0	0,395	-1,063
10	15	5	0,1	0,445	-1,03
11	15	8	0	0,434	-1,066
12	15	8	0,1	0,424	-1,004
13	20	2	0	0,414	-1,032
14	20	2	0,1	0,426	-0,984
15	20	5	0	0,45	-0,44
16	20	5	0,1	0,389	-0,638
17	20	8	0	0,395	-0,714
18	20	8	0,1	0,439	-1,021

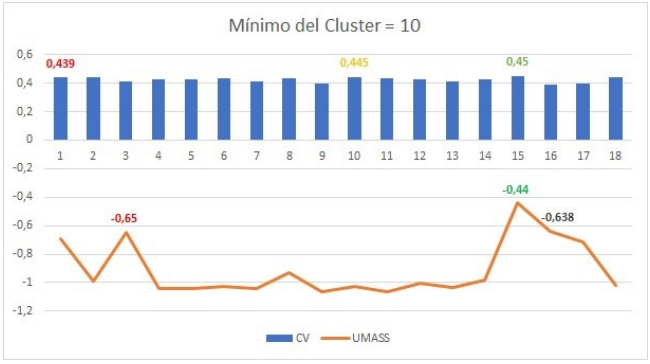

Fig. 8. Comportamiento Tamaño mínimo del cluster = 10

La variación de la métrica CV se mantiene en un rango estrecho entre 0.389 y 0.45. Los valores de la métrica UMASS presentaron mayores variaciones más ampliamente, oscilando entre -1.066 y -0.44.A continuación, se presentan los mejores tres experimentos:

TABLE IV
MEJORES RESULTADOS MINIMO CLUSTER 10

Ranking	# experimento
Mejor experimento	15
Segundo mejor experimento	2
Tercer mejor experimento	14

El experimento número 15 obtuvo el valor más alto tanto en la métrica CV (0.45) como en la métrica UMASS (-0.44). Se resalta la combinación del número de componentes igual a 2, distancia mínima en 0,1 y un tamaño mínimo de cluster igual a 10, los cuales generan un buen desempeño en las métricas combinadas.

Tamaño mínimo del cluster = 15

TABLE V
TAMAÑO MÍNIMO DEL CLUSTER Y RESULTADOS DE EXPERIMENTOS

#	n_neighbors	n_components	min_dist	CV	UMASS
19	10	2	0,0	0,362	-1,119
20	10	2	0,1	0,393	-1,008
21	10	5	0,0	0,386	-1,135
22	10	5	0,1	0,398	-1,088
23	10	8	0,0	0,373	-1,092
24	10	8	0,1	0,376	-1,084
25	15	2	0,0	0,398	-1,051
26	15	2	0,1	0,398	-0,964
27	15	5	0,0	0,376	-1,062
28	15	5	0,1	0,403	-1,024
29	15	8	0,0	0,393	-1,095
30	15	8	0,1	0,422	-1,058
31	20	2	0,0	0,385	-1,051
32	20	2	0,1	0,396	-0,984
33	20	5	0,0	0,400	-1,091
34	20	5	0,1	0,396	-0,617
35	20	8	0,0	0,404	-0,444
36	20	8	0,1	0,376	-0,690

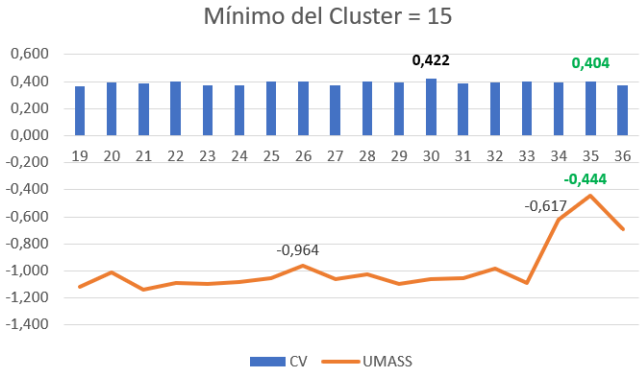

Fig. 9. Comportamiento Tamaño mínimo del cluster = 15

Para el siguiente experimento, La variación de la métrica CV se mantuvo en un rango estrecho entre 0,362 y 0.422, Mientras que, Los valores de la métrica UMASS presentaron mayores variaciones, oscilando entre -1.135 y -0.44. A continuación, se presenta los mejores tres experimentos: El

TABLE VI
MEJORES RESULTADOS MINIMO CLUSTER 15

Ranking	# experimento
Mejor experimento	35
Segundo mejor experimento	30
Tercer mejor experimento	34

experimento número 35 obtuvo el valor más alto tanto en la métrica CV (0.404) como en la métrica UMASS (-0.444). Se resalta la combinación del número de componentes igual a 8 y un tamaño mínimo de cluster igual a 15 generando un buen desempeño en las métricas combinadas.

Tamaño mínimo del cluster = 20

TABLE VII
TAMAÑO MÍNIMO DEL CLUSTER Y RESULTADOS DE EXPERIMENTOS

#	n_neighbors	n_components	min_dist	CV	UMASS
37	10	2	0,0	0,349	-1,094
38	10	2	0,1	0,362	-1,085
39	10	5	0,0	0,375	-1,119
40	10	5	0,1	0,365	-1,017
41	10	8	0,0	0,366	-1,115
42	10	8	0,1	0,364	-1,045
43	15	2	0,0	0,363	-1,107
44	15	2	0,1	0,361	-0,989
45	15	5	0,0	0,364	-1,077
46	15	5	0,1	0,385	-0,858
47	15	8	0,0	0,364	-1,051
48	15	8	0,1	0,388	-0,992
49	20	2	0,0	0,348	-1,089
50	20	2	0,1	0,359	-0,960
51	20	5	0,0	0,393	-0,579
52	20	5	0,1	0,380	-0,684
53	20	8	0,0	0,380	-0,735
54	20	8	0,1	0,391	-1,008

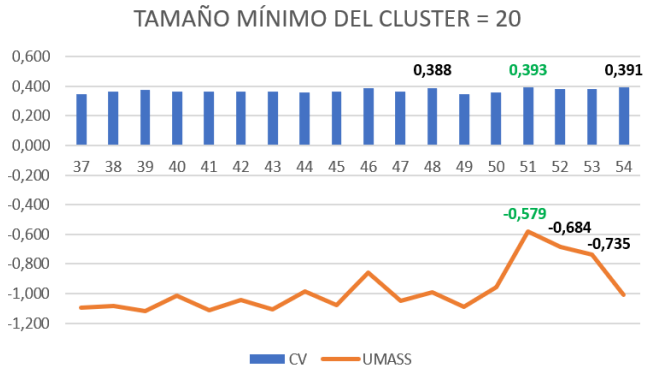

Fig. 10. Comportamiento Tamaño mínimo del cluster = 20

Para el experimento con un cluster con un tamaño superior a 20, La variación de la métrica CV se mantuvo en un rango estrecho entre 0,348 y 0,393. Los valores de la métrica UMASS presentaron mayores variaciones, oscilando entre -0,579 y -1,119 A continuación, se presenta los mejores experimentos:

TABLE VIII
MEJORES RESULTADOS MINIMO CLUSTER 20

Ranking	# Experimento
Mejor experimento	51
Segundo mejor experimento	53
Tercer mejor experimento	54

El experimento número 51 obtuvo el valor más alto tanto en la métrica CV (0.393) como en la métrica UMASS (-0.579). Se resalta la combinación del número de vecinos igual a 20, el número de componentes igual a 8 y un tamaño mínimo de cluster igual a 20 generando buen desempeño en las métricas combinadas.

El experimento con el mejor resultado entre las 54 combinaciones es el número 15. En este experimento se utilizó un tamaño mínimo de cluster igual a 20, un número de vecinos igual a 20, un número de componentes igual a 5 y una distancia mínima igual a 0. Los valores obtenidos fueron un CV de 0.45 y un UMASS de -0.44. Estos parámetros garantizan que los temas generados automáticamente por el modelo son coherentes, las palabras clave extraídas son altamente relevantes y representativas del contenido analizado.

D. Aplicación modelo Bert sintonizado

Durante el mes de junio de 2022, se llevó a cabo el entrenamiento del modelo mediante la generación automática de temas relacionados con la clase política. Como resultado automático, se obtuvieron un total de 752 tópicos. A continuación, se presenta la distribución de estos temas: Los temas más populares, según la cantidad de tuits generados, fueron los siguientes:

- Tópicos 1: 8000 tuits
- Tópicos 0: 7886 tuits
- Tópicos 2: 6663 tuits
- Tópicos 5: 6190 tuits
- Tópicos 4: 5968 tuits

Por otro lado, se identificaron los temas menos frecuentes, los cuales tuvieron una menor cantidad de tuits:

- Tópicos 699: 27 tuits
- Tópicos 717: 28 tuits
- Tópicos 611: 34 tuits
- Tópicos 593: 38 tuits
- Tópicos 513: 39 tuits

La distribución de los tópicos muestra una notable variación en la cantidad de tuits asociados, evidenciando la existencia de temas con mayor y menor participación. Algunos tópicos alcanzaron una alta frecuencia de tuits, con valores superiores a 5,000, mientras que otros presentaron una frecuencia mínima, con menos de 50 tuits asociados. En términos estadísticos, se observa una media de frecuencia de 480 tuits, una mediana de 224.0 tuits y una moda de 146 tuits. Estos valores indican una dispersión significativa en la frecuencia de los tuits, con una presencia destacada de temas con alta frecuencia y la existencia de temas menos comunes.

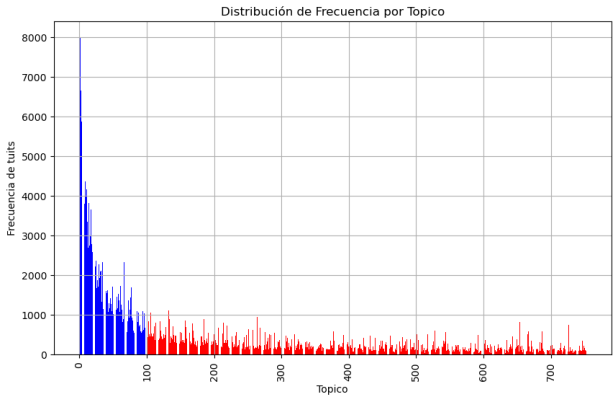

Fig. 11. Distribución de frecuencia por topico

La utilización de técnicas de incrustación c-TF-IDF y UMAP posibilita la generación de representaciones visuales bidimensionales interactivas para los temas estudiados. Estas representaciones visuales permiten un análisis exhaustivo y dinámico de la estructura entre los temas presentada a continuación:

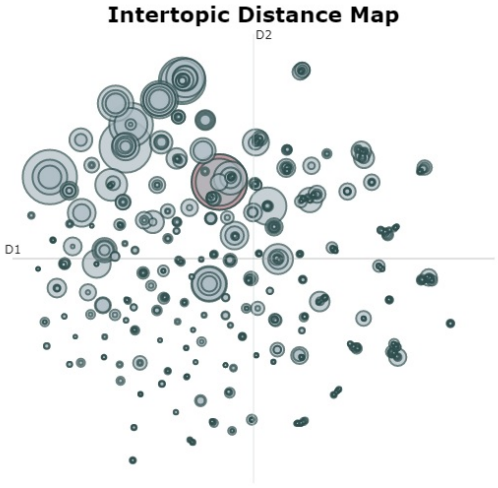

Fig. 12. Mapa de distancia intertopica

El modelo entrenado exhibe una notable capacidad para detectar una gran modularidad en los temas políticos analizados. Esto se refleja en la presencia de numerosos tópicos con una frecuencia baja, lo cual indica que el modelo se enfoca en las representaciones individuales de los temas en lugar de considerar una representación global.

E. Aplicación modelo Bert sintonizado

Con el modelo de tópicos general establecido se continua con la representación en cada paso en el tiempo. Para garantizar una visualización y rendimiento óptimos, se configura el modelo únicamente con los 40 tópicos de mayor frecuencia. Esta elección se basa en la recomendación del autor de la biblioteca Bertopic [32].

Se activan los parámetros de ”afinación global” y ”afinación evolutiva”. Los resultados se representan en la siguiente línea de tiempo donde el número de tuits se encuentra en el eje y, las fechas en el eje x.

Los nombres generados por Bertopic se determinan según la frecuencia de las palabras y se representan mediante colores distintivos en la siguiente gráfica.

Fig. 13. Nombre tópicos en el mes de junio del 2022

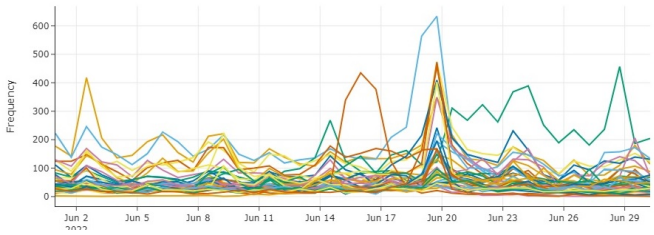

Fig. 14. Tópicos en el mes de junio del 2022

La gráfica presenta un comportamiento dinámico en la frecuencia de tuits para los tópicos analizados. Se destacan el tópico 0 denominado ”fajardo bogota sergio claudia”, y el tópico 1 denominado ”votar votar petro voto petro” con una alta frecuencia a lo largo de toda la línea temporal. Un cambio frecuencial en la mayoría de los tópicos se evidencia a medida que se acerca la fecha del 19 de junio pasado el mediodía, esto subrayan la naturaleza cambiante y dinámica de los tópicos analizados. En este día se llevó a cabo la segunda vuelta de las elecciones presidenciales, en las cuales participaron los candidatos Gustavo Petro y su compañera de fórmula Francia Márquez, quienes obtuvieron el 50,44% de los votos, y Rodolfo Hernández y su fórmula vicepresidencial Marelen Castillo, quienes obtuvieron el 47,31% de los votos [59].

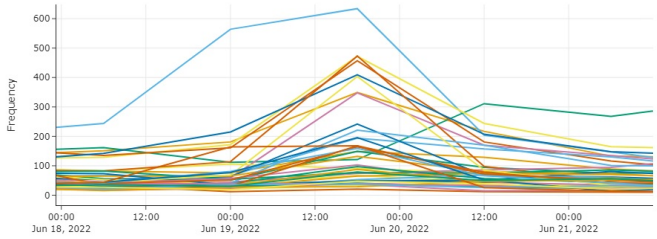

Fig. 15. Tópicos ampliados del 18 al 21 de junio

Para comprender su naturaleza evolutiva, se selecciona el tópico 1 como caso de estudio examinando su trayectoria temporal. Inicialmente, se caracteriza por una variabilidad en la frecuencia de tuits, oscilando entre 100 y 230, lo que representa su etapa inicial. A partir del 17 de junio, experimenta un incremento alcanzando su punto máximo el día 19 de junio a las 18 horas.

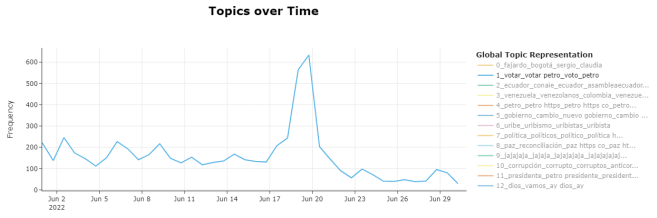

Fig. 16. Comportamiento tópico 1 en el mes de junio del 2022

Posteriormente, se observa un declive abrupto llegando a aproximadamente 50 tuits con una tendencia descendente.

El día 29 de junio, la frecuencia disminuye aún más, aproximándose a cero, lo que podría indicar el final del tema.

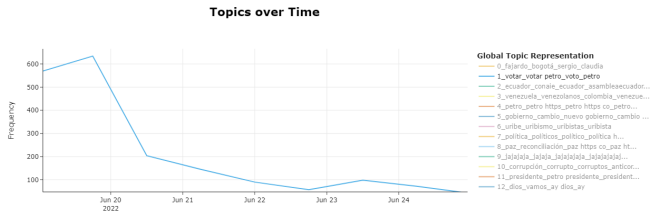

Fig. 17. Finalización tópico 1 en el mes de junio del 2022

F. Detección de comunidades

La detección de comunidades en base a la granularidad de los tópicos detectados proporciona una visión más profunda de cómo los usuarios se agrupan y se relacionan en función de sus intereses y propósitos compartidos. La estructura lingüística y la frecuencia de los tuits son indicadores clave para identificar estas comunidades. Es importante tener en cuenta que las comunidades no son estáticas, sino dinámicas, lo que significa que pueden experimentar cambios en la cantidad de miembros y en sus características a lo largo del tiempo. En relación con los diferentes puntos de vista y posturas que pueden existir dentro de una comunidad, un enfoque prometedor es el análisis no supervisado de sentimientos. Este enfoque permite examinar las opiniones y emociones expresadas en los mensajes de los usuarios, lo que puede revelar discrepancias, similitudes o afinidades en términos de actitudes y perspectivas. En un estudio se descubrieron un total de 752 comunidades utilizando una estructura granular de tópicos. Estas comunidades abarcan una amplia variedad de temas como centro de su comunicación, lo cual refleja la diversidad de intereses presentes en la plataforma. La figura 11 muestra las primeras 100 comunidades destacadas por su alta frecuencia de tuits, lo que indica su relevancia y participación en la plataforma. Cabe destacar que las comunidades siguientes, al tener una frecuencia más baja en contenido, pueden tener menos influencia o fuerza en comparación con las primeras.

V. Conclusiones

Los experimentos llevados a cabo para la detección de temas utilizando técnicas de aprendizaje automático como Naive Bayes, regresión logística y XGBoost, así como basados en Transformers, demuestran un desempeño satisfactorio en las tres métricas evaluadas con un promedio en precisión de 88,25%, recall de 88% y f1 de 0,885. Se destaca la regresión logística con los resultados más favorables, mientras que XGBoost presentó el rendimiento más bajo. A pesar de esto, los modelos de aprendizaje automático exhiben una desigualdad en las métricas para cada clase, debido a la incapacidad de su estructura para capturar de manera eficiente la distribución lingüística, especialmente en las clases con valores más bajos como ”vida” y ”tecnosfera” lo que se identifica como sesgo. Estas clases experimentan una generación sintética durante la etapa de balanceo de clases.

En relación con la detección de tópicos de temas, se han logrado resultados prometedores mediante el uso de la biblioteca Bertopic. Se realizaron 54 experimentos con el objetivo de determinar los hiperparametros óptimos de las bibliotecas HDBSCAN y UMAP. El experimento número 15 se destacó al ofrecer los mejores resultados entre todas las combinaciones evaluadas. En dicho experimento, se establecieron parámetros específicos, como un tamaño mínimo de clúster de 20, un número de vecinos de 20, un número de componentes de 5 y una distancia mínima de 0. Los valores obtenidos para (CV) y UMASS fueron de 0.45 y -0.44, respectivamente. Al aplicar estos parámetros a los tweets generados por los ciudadanos colombianos durante el mes de junio de 2022, se logró agrupar un total de 752 tópicos. El modelo entrenado demostró una destacada capacidad para identificar una alta modularidad en los temas políticos analizados, como se evidencia en la presencia de múltiples tópicos con una frecuencia baja. Sin embargo, es importante destacar que los ajustes actuales del modelo generan una detección de tópicos excesivamente detallada, lo que dificulta la comprensión global de los textos analizados.

VI. Trabajos futuros

En la fase de clasificación de temas, se sugiere aumentar la cantidad de datos por clase, centrándose especialmente en las clases minoritarias, como ”vida” y ”tecnosfera”, con el objetivo de lograr modelos más equilibrados en sus métricas y evitar posibles sesgos causados por el desequilibrio de clases. Además, se plantea la exploración de variaciones en la estructura de Bert, específicamente adaptadas al idioma español, como BETTO.

En cuanto a la etapa de detección de tópicos, se recomienda ampliar los hiperparámetros del modelo Bertopic para mejorar la comprensión de los temas, reduciendo así el número de tópicos generados. Se propone realizar un análisis exhaustivo que considere la granularidad deseada por parte del usuario y la necesidad de la tarea en cuestión. De esta manera, se busca lograr una representación más completa y comprensible de los tópicos presentes en los textos.

References

[1] R. Williams, “Tecnologıas de la comunicación e instituciones sociales,” *Historia de la comunicación*, vol. 2, pp. 181–209, 1992.

[2] R. V. L. Taylor, “Recesión: La comunicación: De los orıgenes a internet,” *PAAKAT: Revista de Tecnologıa y Sociedad*, no. 4, 2012.

[3] G. Vélez, “Exploración de las relaciones entre redes sociales y comunicación,” *Razón y palabra*, no. 61, 2008.

[4] E. Puertas, L. G. Moreno Sandoval, ·. Redondo, J. Alvarado, and A. Pomares Quimbaya, “Detection of sociolinguistic features in digital social networks for the detection of communities,” *Cognitive Computation*, vol. 13, p. 20, Mar. 2021. DOI: 10.1007/s12559-021-09818-9.

[5] L. G. Moreno Sandoval, A. Pomares Quimbaya, and J. Alvarado, "Celebrity profiling through linguistic analysis of digital social networks," *Computational Social Networks*, vol. 8, Aug. 2021. DOI: 10.1186/s40649-021-00097-w.

[6] J. M. Flores Vivar, "Nuevos modelos de comunicación, perfiles y tendencias en las redes sociales," *Comunicar: Revista Cientıfica de Comunicación y Educación*, vol. 17, no. 33, pp. 73–81, 2009.

[7] M. B. Pulido, Á. D. Soto, F. M. Lozano, and W. Q. Peña, "Redes sociales y relaciones digitales, una comunicación que supera el cara a cara," *Revista internacional de pedagogıa e innovación educativa*, vol. 1, no. 1, pp. 123–148, 2021.

[8] L. G. Moreno Sandoval and A. Pomares Quimbaya, "Hybrid onion layered system for the analysis of collective subjectivity in social networks," *IEEE Access*, vol. PP, pp. 1–1, Oct. 2022. DOI: 10.1109/ACCESS.2022.3217467.

[9] W. Huang, S.-H. Hong, and P. Eades, "Effects of sociogram drawing conventions and edge crossings in social network visualization.," *J. Graph Algorithms Appl.*, vol. 11, no. 2, pp. 397–429, 2007.

[10] *Ranking mundial de redes sociales por número de usuarios*, https://es.statista.com/estadisticas/600712/ranking-mundial-de-redes-sociales-por-numero-de-usuarios/, Accedido el 19 de mayo, 2023.

[11] Datareportal, *Digital 2023 global overview report*, https://datareportal.com/reports/digital-2023-global-overview-report, 2023, 2023.

[12] K. Hashimoto, G. Kontonatsios, M. Miwa, and S. Ananiadou, "Topic detection using paragraph vectors to support active learning in systematic reviews," *Journal of biomedical informatics*, vol. 62, pp. 59–65, 2016.

[13] M. Boukes, "Social network sites and acquiring current affairs knowledge: The impact of twitter and facebook usage on learning about the news," *Journal of Information Technology & Politics*, vol. 16, no. 1, pp. 36–51, 2019.

[14] I. B. A. Peling, I. N. Arnawan, I. P. A. Arthawan, and I. G. N. Janardana, "Implementation of data mining to predict period of students study using naive bayes algorithm," *Int. J. Eng. Emerg. Technol*, vol. 2, no. 1, p. 53, 2017.

[15] L. A. U. López, M. T. M. Valdivia, M. Á. G. Cumbreras, and A. J. O. Martos, "Detección automática de spam utilizando regresión logıstica bayesiana," *Procesamiento del Lenguaje Natural*, no. 35, pp. 127–133, 2005.

[16] M. K. Liao-Li, J. M. Celaya-Padilla, C. E. Galván-Tejada, *et al.*, "Detección de nefropatıa como complicación en pacientes diabéticos de tipo ii mediante el uso de la regresión logıstica,"

[17] X. Lin, "Sentiment analysis of e-commerce customer reviews based on natural language processing," in *Proceedings of the 2020 2nd International Conference on Big Data and Artificial Intelligence*, 2020, pp. 32–36.

[18] "Descripción general de los modelos basados en transformadores para tareas de pnl," in *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*.

[19] L. Tunstall, L. Von Werra, and T. Wolf, *Natural language processing with transformers*. " O'Reilly Media, Inc.", 2022.

[20] I. Tenney, D. Das, and E. Pavlick, *Bert rediscovers the classical nlp pipeline*, 2019. arXiv: 1905.05950 [cs.CL].

[21] H. Choi, J. Kim, S. Joe, and Y. Gwon, "Evaluation of bert and albert sentence embedding performance on downstream nlp tasks," in *2020 25th International conference on pattern recognition (ICPR)*, IEEE, 2021, pp. 5482–5487.

[22] D. Lévano and F. E. C. León, "Discriminación de masas mamográficas mediante k-nearest neighbor y atributos birads," *Revista cientıfica de sistemas e informática*, vol. 2, no. 1, pp. 1–12, 2022.

[23] P. R. Q. Nina and F. B. Q. Cahuana, "Clasificación de texto con nlp en tweets relacionados con desastres naturales," *Innovación y Software*, vol. 4, no. 1, pp. 198–203, 2023.

[24] J. G. Suntaxi Recalde, "Modelo de detección de discurso de odio en ecuador mediante clasificación supervisada de tweets y técnicas de nlp.," M.S. thesis, Quito: EPN, 2022., 2022.

[25] P. Das and N. Sultana, "Sentiment analysis on comments in bengali language using text mining machine learning approach," Cited by: 0, 2022. DOI: 10.1109/I2CT54291.2022.9825373. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85135618368&doi=10.1109%2fI2CT54291.2022.9825373&partnerID=40&md5=0bdcd6df0f3ec776a9625a6949b47a42.

[26] N. Jung and G. Lee, "Automated classification of building information modeling (bim) case studies by bim use based on natural language processing (nlp) and unsupervised learning," *Advanced Engineering Informatics*, vol. 41, 2019. DOI: 10.1016/j.aei.2019.04.007. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85064926758&doi=10.1016%2fj.aei.2019.04.007&partnerID=40&md5=995dcabbc2c2a887eba0f70a5f036fb9.

[27] S. H. Mohammed and S. Al-augby, "Lsa & lda topic modeling classification: Comparison study on e-books," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 19, no. 1, pp. 353–362, 2020.

[28] R. Alghamdi and K. Alfalqi, "A survey of topic modeling in text mining," *Int. J. Adv. Comput. Sci. Appl.(IJACSA)*, vol. 6, no. 1, 2015.

[29] J. Caballero Villalobos, J. E. Enciso Agudelo, *et al.*, "Mineria de texto aplicado en preguntas abiertas sobre evaluación docente: Enfoque de modelado de tópicos con lda,"

[30] A. Srivastav and S. Singh, "Proposed model for context topic identification of english and hindi news article through lda approach with nlp technique," *Journal of The Institution of Engineers (India): Series B*, pp. 1–7, 2022.

[31] Z. Tong and H. Zhang, "A text mining research based on lda topic modelling," in *International conference on computer science, engineering and information technology*, 2016, pp. 201–210.

[32] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[33] A. Abuzayed and H. Al-Khalifa, "Bert for arabic topic modeling: An experimental study on bertopic technique," *Procedia computer science*, vol. 189, pp. 191–194, 2021.

[34] R. Egger and J. Yu, "A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts," *Frontiers in sociology*, vol. 7, 2022.

[35] UMAP, *UMAP visualization documentation*, UMAP Visualization Documentation, 2018. [Online]. Available: https://umap-learn.readthedocs.io/en/latest/plotting.html.

[36] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[37] K. Bothmer and T. Schlippe, "Investigating natural language processing techniques for a recommendation system to support employers, job seekers and educational institutions," in *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners' and Doctoral Consortium: 23rd International Conference, AIED 2022, Durham, UK, July 27–31, 2022, Proceedings, Part II*, Springer, 2022, pp. 449–452.

[38] R. J. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Advances in Knowledge Discovery and Data Mining: 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part II 17*, Springer, 2013, pp. 160–172.

[39] J. Collard, "Lsi and dbscan: Natural language processing for sociolinguistic analysis," 2015.

[40] D. Birant and A. Kut, "St-dbscan: An algorithm for clustering spatial–temporal data," *Data & knowledge engineering*, vol. 60, no. 1, pp. 208–221, 2007.

[41] K. Khan, S. U. Rehman, K. Aziz, S. Fong, and S. Sarasvady, "Dbscan: Past, present and future," in *The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)*, 2014, pp. 232–238. DOI: 10.1109/ICADIWT.2014.6814687.

[42] A. J. Rawat, S. Ghildiyal, and A. K. Dixit, "Topic modelling of legal documents using nlp and bidirectional encoder representations from transformers," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 28, no. 3, pp. 1749–1755, 2022.

[43] A. Udupa, K. Adarsh, A. Aravinda, N. H. Godihal, and N. Kayarvizhy, "An exploratory analysis of gsdmm and bertopic on short text topic modelling," in *2022 Fourth International Conference on Cognitive Computing and Information Processing (CCIP)*, IEEE, 2022, pp. 1–9.

[44] M. Zhou, Y. Kong, and J. Lin, "Financial topic modeling based on the bert-lda embedding," in *2022 IEEE 20th International Conference on Industrial Informatics (INDIN)*, IEEE, 2022, pp. 495–500.

[45] J. H. Rincón Ruiz *et al.*, "Estudio comparativo de técnicas tradicionales del modelado de tópicos frente a redes neuronales artificiales tomando como contexto el discurso digital del autor en la red social twitter y otras publicaciones," 2021.

[46] I. Guillén-Pacho, C. Badenes-Olmedo, and O. Corcho, "Dynamic topic modelling for exploring the scientific literature on coronavirus: An unsupervised labelling technique," 2023.

[47] S. V. Raju, B. K. Bolla, D. K. Nayak, and J. Kh, "Topic modelling on consumer financial protection bureau data: An approach using bert based embeddings," in *2022 IEEE 7th International conference for Convergence in Technology (I2CT)*, IEEE, 2022, pp. 1–6.

[48] *Twitter*, https://twitter.com/?lang=es.

[49] L. G. Moreno Sandoval, A. Pomares Quimbaya, C. Gutiérrez, J. Pachón, and D. Ramírez, "Comparación de métodos de análisis de sentimientos en comunidades de habla hispana," Sep. 2022. DOI: 10.26507/paper.2367.

[50] *re - Operaciones de expresiones regulares*, Sitio web de la documentación de Python, 2023. [Online]. Available: https://docs.python.org/es/3/library/re.html.

[51] *Emojipedia - Libros*, Sitio web de Emojipedia, 2023. [Online]. Available: https://emojipedia.org/es/libros/.

[52] *Unidecode - pypi*, https://pypi.org/project/Unidecode/, 2023.

[53] Steven Bird, Edward Loper, and Ewan Klein, *Natural Language Toolkit (NLTK)*, https://www.nltk.org/, 2023.

[54] Matthew Honnibal, Ines Montani, and spaCy Contributors, *spaCy*, https://spacy.io/, 2023.

[55] *El tiempo*, https://www.eltiempo.com/, 2023.

[56] Imbalanced-learn, *Imbalanced-learn - smote*, https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html, 2023.

[57] Imbalanced-learn, *Imbalanced-learn - randomundersampler*, https://imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.RandomUnderSampler.html, 2023.

[58] scikit-learn Development Team, *CountVectorizer*, scikit-learn documentation, 2023. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html.

[59] C. N. E. (CNE), *Elecciones presidenciales de colombia 2022*, https://www.cne.gov.co/elecciones/elecciones-2022/presidencia-2022, 2023.