Maestría en

Ciencia de Datos y Máquinas de Aprendizaje mención en Inteligencia Artificial

Trabajo de investigación previo a la obtención del título de

Magíster en Ciencia de Datos y Máquinas de Aprendizaje

con mención en Inteligencia Artificial

AUTORES:

FRANK PALADINES SHIRLEY MARIBEL

GARCIA ORTIZ JOSELIN SOFIA

PINEDA FERNANDEZ DE CORDOVA PEDRO JOSE

POLANCO FRANCO JUAN FERNANDO

TUTORES:

CORTES LOPEZ ALEJANDRO

Análisis Comparativo de Algoritmos de Agrupamiento para Identificar Patrones de Opinión en Redes Sociales: Estudio de Tweets en Español

Quito, noviembre 2024

Certificación de autoría

Nosotros, Pedro José Pineda Fernández de Córdova, Shirley Maribel Frank Paladines, Joselin Sofía García Ortiz, Juan Fernando Polanco Franco**,** declaramos bajo juramento que el trabajo aquí descrito es de nuestra autoría; que no ha sido presentado anteriormente para ningún grado o calificación profesional y que se ha consultado la bibliografía detallada.

Cedemos nuestros derechos de propiedad intelectual a la Universidad Internacional del Ecuador (UIDE), para que sea publicado y divulgado en internet, según lo establecido en la Ley de Propiedad Intelectual, su reglamento y demás disposiciones legales.

Firma del graduando
Pedro José Pineda Fernández de Córdova

Firma del graduando
Shirley Maribel Frank Paladines

Firma del graduando
Juan Fernando Polanco Franco

Firma del graduando
Joselin Sofía García Ortiz

Autorización de Derechos de Propiedad Intelectual

Nosotros, Pedro José Pineda Fernández de Córdova, Shirley Maribel Frank Paladines, Joselin Sofía García Ortiz , Juan Fernando Polanco Franco, en calidad de autores del trabajo de investigación titulado *Análisis Comparativo de Algoritmos de Agrupamiento para Identificar Patrones de Opinión en Redes Sociales: Estudio de Tweets en Español*, autorizamos a la Universidad Internacional del Ecuador (UIDE) para hacer uso de todos los contenidos que nos pertenecen o de parte de los que contiene esta obra, con fines estrictamente académicos o de investigación. Los derechos que como autores nos corresponden, lo establecido en los artículos 5, 6, 8, 19 y demás pertinentes de la Ley de Propiedad Intelectual y su Reglamento en Ecuador.

D. M. Quito, Diciembre 2024

Firma del graduando
edro José Pineda Fernández de Córdova

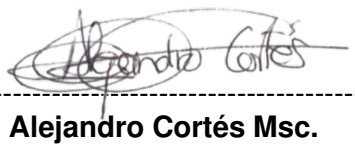
Firma del graduando
Shirley Maribel Frank Paladines

Firma del graduando
Juan Fernando Polanco Franco

Firma del graduando
Joselin Sofía García Ortiz

Aprobación de dirección y coordinación del programa

Nosotros, **Alejandro Cortés e Iván Reyes,** declaramos que los Pedro José Pineda Fernández de Córdova, Shirley Maribel Frank Paladines, Joselin Sofía García Ortiz, Juan Fernando Polanco Franco, son los autores exclusivos de la presente investigación y que ésta es original, auténtica y personal de ellos.



Alejandro Cortés Msc.
Director de la
Maestría en Ciencia de datos y Maquinas de Aprendizaje mención Inteligencia Artificial EIG

Iván Reyes Ch. Mgtr.
Coordinador de la
Coordinador de Posgrados Escuela de Ciencia de la Computación UIDE

DEDICATORIA

Dedico esta tesis principalmente a Dios, por regalarme siempre un día más de vida y darme las fuerzas para continuar, especialmente en los días más difíciles.

Y finalmente a mis padres, por darme su amor y apoyo incondicional durante esta etapa de mi vida y siempre motivarme a buscar más retos que me ayuden a crecer como una persona de bien.

Juan Fernando Polanco Franco

Dedico esta tesis a Dios por ser mi fortaleza e inspiración. A mis padres Miguel y Mercedes por su amor, ejemplo de dedicación y esfuerzo, también a mis hermanos por estar presentes en cada etapa de mi vida, apoyándome y motivándome para ser la mejor.

Joselin Sofía García Ortiz

A mis hijos, mis verdaderos motores, cada paso dado ha sido motivado por su bienestar y su felicidad, y este trabajo es, en definitiva, un homenaje a ustedes. Mi gratitud y amor por siempre.

Shirley Maribel Frank Paladines

A Beethoven, Vladi, Linda, Veneno, Pinky y Cejas, que sigan moviendo sus colitas felices en el cielo de los perritos.

Pedro José Pineda Fernández de Córdova

AGRADECIMIENTOS

Quiero comenzar agradecimiento a mis compañeros de tesis, cuya compañía y amistad me ayudaron en los momentos más duros de este trabajo. Cada uno de ustedes contribuyó a que este proceso fuera algo significativo en mi vida.

A la Universidad Internacional del Ecuador, por abrirme las puertas a un mundo de nuevos conocimientos y de crecimiento profesional. Aprecio profundamente su dedicación para brindarme un ambiente de aprendizaje de calidad.

Finalmente quiero agradecer a mi familia por su apoyo incondicional, especialmente a mis padres. Ustedes fueron y serán siempre el motor que me ayuda a cruzar todos los obstáculos que la vida me presenta.

Juan Fernando Polanco Franco

Agradezco a todas las personas que me acompañaron durante este camino, cuya presencia y apoyo hicieron posible la culminación de esta nueva etapa. A Dios, por brindarme sabiduría y motivación para alcanzar mis objetivos. A mis padres, por ser mi fuente diaria de inspiración y amor. Y a mis hermanos, por ser no solo mis mejores amigos, sino también un ejemplo constante de perseverancia y esfuerzo en la vida.

Joselin Sofía Garcia Ortiz

Quiero expresar mi más sincero agradecimiento a todas las personas que me han acompañado a lo largo de este proceso. A mis padres, por su amor incondicional y apoyo constante. A mis hijos, por su alegría y motivación, quienes han sido mi mayor fuente de inspiración y fortaleza. A mis compañeros por el apoyo, colaboración y amistad durante todo este desafío. A mis profesores y tutores, por su orientación en cada etapa de este trabajo. Y a todos aquellos que, de alguna manera, contribuyeron a que hoy pueda presentar este logro. Gracias de corazón.

Shirley Maribel Frank Paladines

A Laura Portocarrero, por ser mi motor que dobla el espacio para alcanzar las estrellas y mi cable a tierra cuando más lo necesito.

A todos los que luchan por democratizar la tecnología y acercarla a la sociedad.

Y, por último, quiero agradecerme a mí mismo citando esta poesía:

"Me agradezco a mí mismo por seguir adelante cuando ya no podía más.
Por ser valiente esas veces que quise salir corriendo.
Por seguir intentando sin rendirme.
Por soñar y amar a pesar de las circunstancias.
Hoy me agradezco, me valoro y me felicito."

Anónimo

Pedro José Pineda Fernández de Córdova

Resumen

Este trabajo presenta un análisis exhaustivo de técnicas de agrupamiento aplicadas a datos textuales, con el objetivo de identificar tendencias temáticas relevantes. Se evaluaron tres algoritmos de clustering (KMeans, DBSCAN y Agglomerative Clustering) en combinación con dos representaciones vectoriales: TF-IDF y Word2Vec. Los resultados muestran que KMeans con Word2Vec sobresale como la metodología más efectiva, logrando clústeres bien definidos con alta cohesión interna y baja superposición. El estudio también identifica limitaciones en la implementación, como el consumo elevado de memoria RAM, y propone pasos futuros para mejorar los resultados, incluyendo la limpieza más profunda del corpus y el uso de algoritmos avanzados. Estas conclusiones son relevantes para aplicaciones como campañas de marketing político, donde la identificación de tendencias temáticas es crucial para la toma de decisiones estratégicas.

Palabras Claves:

Agrupamiento, KMeans, DBSCAN, Agglomerative Clustering, TF-IDF, Word2Vec, campañas de marketing político, análisis de datos textuales, tendencias temáticas.

Abstract

This study provides a comprehensive analysis of clustering techniques applied to textual data to identify relevant thematic trends. Three clustering algorithms (KMeans, DBSCAN, and Agglomerative Clustering) were evaluated using two vector representations: TF-IDF and Word2Vec. Results indicate that KMeans with Word2Vec outperforms other methods, achieving well-defined clusters with high internal cohesion and low overlap. The study also highlights implementation limitations, such as high memory usage, and suggests future improvements, including deeper corpus cleaning and advanced algorithms. These findings are particularly relevant for applications like political marketing campaigns, where thematic trend identification is critical for strategic decision-making.

Keywords:

Clustering, KMeans, DBSCAN, Agglomerative Clustering, TF-IDF, Word2Vec, political marketing campaigns, textual data analysis, thematic trends.

Tabla de Contenidos

Lista de Tablas

Lista de Figuras

Capítulo 1: Introducción

Planteamiento del Problema e Importancia del Estudio

Definición del proyecto

El objetivo de esta investigación es llevar a cabo un análisis comparativo de algoritmos de agrupamiento o clustering aplicados a un conjunto de tweets en español sobre temas políticos. La investigación pretende evaluar la efectividad de diversas técnicas de clustering para identificar comunidades de opinión y patrones subyacentes en las conversaciones, sin depender de etiquetas predefinidas en el conjunto de datos, ya que se utilizan herramientas de aprendizaje no supervisado. Esto permitirá analizar la capacidad de los distintos métodos de agrupamiento para identificar grupos de usuarios que comparten conversaciones sobre tópicos similares, revelando así estructuras de comunidad emergentes.

En la era digital actual, la red social X, anteriormente conocida como Twitter antes de su adquisición por el empresario multimillonario Elon Musk (Jia & Xu, 2022), se ha consolidado como una de las plataformas fundamentales para comunicar nuestros pensamientos y permitir el intercambio de información (Alsina, 2024). Esta red social genera un volumen masivo de datos sobre los comportamientos y características de los usuarios, y es conocida por promover la libertad de expresión, permitiendo a los usuarios expresar sus opiniones sin sentirse censurados. Esto lo convierte en un recurso muy valioso para la investigación académica sobre el comportamiento humano y las tendencias de conversación social (Criado & Villodre, 2018; Piñeiro-Otero, 2023). Analizar datos con estas características resulta esencial para comprender las dinámicas sociales y de comunicación en estas plataformas, ya que facilitan el estudio de tendencias y opiniones que emergen en el entorno digital (Piñeiro-Otero, 2023).

En este contexto, la red social X ha sido objeto de numerosos debates y análisis académicos que se han centrado en su papel como medio de comunicación política para la comunicación de

diversas perspectivas de los usuarios (Jungherr, 2014). La efectividad de esta plataforma para comunicar ideas ha sido tan alta que varias campañas electorales se han ejecutado a lo largo del mundo, tales como las elecciones alemanas de 2009 (Jürgens & Jungherr, 2015), las de los Países Bajos en 2010 (Vergeer & Hermans, 2013), las del Reino Unido en 2010 (Graham et al., 2016), entre muchas otras. Todas estas campañas han seguido un patrón influenciado por los estudios de las elecciones estadounidenses, comenzando con la campaña de Barack Obama en 2008 y observando un cambio significativo durante la campaña de Donald Trump en 2016 (Campos-Domínguez, 2017).

A partir de esta tendencia, X se ha consolidado como un canal clave para la comunicación masiva y la difusión de información de los candidatos, lo cual requiere diversos procesos de medición y de optimización. Incrementando de esta manera la implementación de técnicas de ciencias de datos para identificar los temas de conversación más relevantes y optimizar la estrategia de comunicación (Arcila-Calderón et al., 2017; Bimber, 2014).

El análisis de agrupamiento, o clustering, es una técnica en el análisis de datos que permite identificar patrones y estructuras subyacentes en conjuntos de datos complejos. Esta técnica facilita la detección de comunidades y la comprensión de diferencias comunicativas entre grupos, proporcionando una visión más profunda de cómo se organizan y distribuyen las opiniones en el entorno digital (Avila Perez-Grovas, 2023). En el contexto del marketing político, disponer de esta información resulta crucial para diseñar estrategias adaptadas a las necesidades reales de los votantes, contribuyendo a la creación de campañas más efectivas que respondan a sus intereses ocultos y comportamientos colectivos que pueden ser cruciales para el análisis político y social (Campos-Domínguez & Calvo, 2017).

Para llevar a cabo el objetivo de esta investigación, es necesario enfrentar varios desafíos técnicos, como encontrar un conjunto de datos pertinente y la limpieza de este para asegurar su relevancia, la selección de representaciones numéricas que capturen de manera adecuada las similitudes semánticas entre los tweets y la adaptación de los datos a las distintas técnicas de

agrupamiento no supervisado. A su vez, se debe explorar alternativas cuando estas representaciones no logran reflejar las relaciones esperadas, recurriendo a enfoques que identifican temas subyacentes en el texto para mejorar la coherencia de la agrupación.

Finalmente, la evaluación de la metodología aplicada permitirá comparar el desempeño de las distintas técnicas y ajustar los parámetros necesarios para optimizar la identificación de comunidades dentro de la red social. De este modo, se busca aportar un análisis detallado sobre la dinámica de la conversación política en Twitter, contribuyendo al entendimiento de la influencia de las redes sociales en el discurso público.

Este análisis comparativo no solo permitirá evaluar la precisión y eficiencia de cada algoritmo para identificar comunidades de opinión, sino que también proporcionará una comprensión más profunda de las dinámicas discursivas presentes en la red social X. Esta información es importante para el análisis político desde una perspectiva de ciencia de datos, ya que permite identificar patrones de comportamiento y tendencias de opinión entre diferentes grupos de usuarios. Esto facilita el desarrollo de modelos que capturen de manera más precisa la dinámica de la conversación pública y el impacto de ciertos temas en las redes sociales. En última instancia, la investigación busca contribuir a una mejor comprensión de la interacción entre los actores políticos y la ciudadanía, utilizando los conocimientos obtenidos del análisis de datos para revelar cómo ciertos temas impactan de mayor manera en ciertos segmentos de la población y cómo se propagan las ideas en el entorno digital.

Justificación del Proyecto

La creciente popularidad de las redes sociales ha generado una necesidad urgente de utilizar herramientas de ciencia de datos para extraer información valiosa y aprender de las dinámicas que se desarrollan en estas plataformas. Redes sociales como la plataforma X ofrecen una vasta cantidad de datos sobre los intereses, opiniones y comportamientos de los usuarios, convirtiéndose en un recurso

esencial para entender tendencias sociales y diseñar campañas de marketing político más efectivas. La capacidad de segmentar y analizar estos datos permite a los investigadores identificar patrones y tendencias, brindando una comprensión más profunda de cómo se forman las comunidades digitales y de qué manera se pueden dirigir mejor los esfuerzos de comunicación para conectar con los votantes (Saura García, 2023).

En este contexto, los algoritmos no supervisados son fundamentales, ya que permiten identificar patrones y estructuras subyacentes en los datos sin la necesidad de etiquetas previas. Esto resulta especialmente útil en el análisis de redes sociales, donde la cantidad de datos es inmensa y la variabilidad del lenguaje es alta. La eficiencia de estos algoritmos se mide no solo en términos de precisión, sino también en su capacidad para procesar rápidamente los datos y escalar a conjuntos más grandes.

El uso de técnicas avanzadas de procesamiento de lenguaje natural (NLP) es crucial para enfrentar los retos que plantea el lenguaje informal y dinámico de las redes sociales. Herramientas de NLP, como la vectorización de palabras, permiten transformar el texto en representaciones numéricas que los algoritmos pueden procesar, facilitando la identificación de patrones significativos en el discurso digital. La calidad del preprocesamiento de los datos, que incluye la limpieza y adaptación del conjunto de datos al análisis en español, es fundamental para mejorar el rendimiento de los algoritmos. Una preparación adecuada puede ayudar a los modelos a captar las sutilezas del lenguaje en español, optimizando así su eficacia en la identificación de patrones de opinión.

La implementación de los algoritmos de agrupamiento requiere una evaluación para determinar cuál es más adecuado según la naturaleza de los datos y los objetivos del análisis. A través de un análisis comparativo de diferentes algoritmos, este proyecto busca identificar cuál proporciona la mejor segmentación, no solo en términos de precisión, sino también en su eficiencia y escalabilidad, considerando la complejidad inherente a los datos de las redes sociales.

En última instancia, este proyecto contribuye al avance en el uso de técnicas de ciencia de

datos para el análisis de redes sociales, proporcionando una guía práctica para elegir el mejor enfoque de agrupamiento en función de las características de los datos y los objetivos de análisis. Los resultados obtenidos pueden ser aprovechados para desarrollar herramientas de segmentación de audiencia más precisas y personalizadas, facilitando una mejor interacción entre actores sociales y sus comunidades. Esto no solo tiene implicaciones para la investigación académica, sino que también ofrece un marco robusto para el diseño de estrategias de comunicación política más conectadas con los intereses y necesidades reales de los usuarios.

Alcance del Proyecto

Este proyecto busca realizar una evaluación comparativa de diversas técnicas de agrupamiento (clustering) aplicadas a un conjunto de datos conformado por tweets redactados en español, con el objetivo de identificar patrones de opinión relevantes. El propósito central es determinar cuál de estas técnicas es más efectiva para identificar y analizar las dinámicas de opiniones relacionadas con distintos temas, incluyendo la identificación de comunidades afines a diversos partidos políticos.

Las etapas del proyecto incluyen:

- Selección y preparación del conjunto de datos: Se buscará y seleccionará un conjunto de datos adecuado que permita un análisis profundo de las opiniones de usuarios en redes sociales, específicamente a partir de los tweets publicados en español.
- Implementación de técnicas de representación de texto: Se explorarán diferentes enfoques para la representación de los tweets, como métodos de vectorización y modelos temáticos, asegurando que la semántica del discurso se preserve adecuadamente en español.
- Implementación de técnicas de agrupamiento: Se aplicarán diferentes algoritmos de agrupamiento utilizando herramientas de análisis de datos en Python, adaptando cada

técnica a las características particulares del conjunto de datos y asegurando que los algoritmos identifiquen comunidades de usuarios de manera efectiva.

- Evaluación del rendimiento de los algoritmos: Se medirá la eficiencia de cada técnica de agrupamiento.(Avila Perez-Grovas, 2023; Campos-Domínguez & Calvo, 2017)
- Análisis comparativo de resultados: Se realizará un análisis detallado de los resultados obtenidos, identificando las fortalezas y limitaciones de cada técnica de agrupamiento.
- Visualización de resultados: Se desarrollarán visualizaciones que representen de manera gráfica los grupos formados por los algoritmos, con el fin de facilitar la interpretación y análisis de los patrones de opinión identificados

Objetivos del Proyecto

Objetivo General

Analizar y comparar el desempeño de diferentes algoritmos de agrupamiento para la identificación de patrones de opinión en redes sociales.

Objetivos Específicos

- Encontrar, preprocesar y limpiar un conjunto de datos de tweets en español para la eliminación de ruido, asegurando la representatividad de la información y adaptarlos adecuadamente para el análisis.
- Implementar técnicas de representación de texto, como métodos de vectorización y modelos temáticos, para la preservación de la semántica del discurso en español.
- Implementar diversas técnicas de agrupamiento en un conjunto de datos obtenido de la red social X para la evaluación de los patrones de opinión que mantienen los usuarios.

- Identificar las fortalezas y debilidades de las técnicas de agrupamiento en el contexto del análisis de texto.

Capítulo 2. Marco Teórico

Introducción al Clustering

El clustering se ha convertido en una herramienta fundamental en el análisis exploratorio de datos crudos sin procesar, siendo utilizada en diversas disciplinas, desde las ciencias sociales hasta la medicina y la bioinformática (Shalev-Shwartz & Ben-David, 2014). Los científicos de datos buscan identificar similitudes en grupos significativos de datos relacionados a estos campos. Por ejemplo, los biólogos informáticos agrupan genes en función de las similitudes encontradas en diferentes experimentos (Peng, 2008; Quintana et al., 2003; Tari et al., 2009), como también los marketeros agrupan a los posibles clientes según sus perfiles para fines de campañas de marketing dirigidas (Mahdiraji et al., 2019; Quintana, 2003).

En este sentido, el clustering es la tarea de agrupar un conjunto de datos de tal manera que los que presentan características similares terminen en el mismo grupo y los que no, se separen en grupos diferentes. Para explicar de manera más clara el concepto de clustering, podemos abordarlo desde un enfoque matemático, para lo cual consideremos un conjunto de datos compuesto por puntos:

$$X = \{\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n\} \; donde \; \bar{x}_i \in \mathbb{R}^m \qquad (1)$$

Supongamos que podemos identificar un criterio o característica que asocie cada muestra con un grupo específico:

$$g_k = G(\bar{x}_i) \; donde \; k \in \{0, 1, 2, \ldots, t\} \qquad (2)$$

En este contexto, cada grupo se denomina *"cluster"* y el proceso de encontrar la función G se denomina *"clusterización"*. En este ejemplo, no se imponen restricciones a los clusters, pero es necesario establecer un criterio que permita agrupar ciertos elementos y separarlos de otros. Existen diversos algoritmos de clusterización no supervisadas que emplean diferentes estrategias para abordar este desafío, lo que puede resultar en resultados significativamente variados. La función de

agrupamiento genérica se expresa como (Bonaccorso, 2017):

$$\bar{m}_i = F(x_i) \text{ donde } \bar{m}_i = (m_i^0, m_i^1, m_i^2, \ldots, m_i^t) \text{ y } m_i^k \in [0, 1] \quad (3)$$

El vector $\bar{m}_i$ representa la membresía relativa de $\bar{x}_i$, un concepto en el clustering que indica el grado de pertenencia de un dato a una categoría específica. Esto resulta especialmente útil en situaciones donde un elemento puede pertenecer a más de un grupo. Un valor de membresía es una representación numérica que oscila entre 0 y 1, indicando el grado de pertenencia de un valor de entrada a una categoría definida por una función de membresía (Strušnik et al., 2015).

Concepto de Similitud y Disimilitud en Clustering

El desafío en el clustering surge de que los dos objetivos, similitud y disimilitud, pueden a menudo entrar en conflicto. Esto significa que, aunque deseemos agrupar elementos similares, también queremos asegurarnos de que los elementos disímiles no se agrupen (Esteban et al., 2021).

Conflicto entre Similitud y Disimilitud

Desde una perspectiva matemática, la similitud (o proximidad) no es una relación transitiva. Es decir, si dos elementos A y B son similares, y B es similar a un tercer elemento C, esto no implica necesariamente que A y C también sean similares. En cambio, la pertenencia a un clúster se considera una relación de equivalencia y, por tanto, transitiva. Esto significa que si

A pertenece al mismo clúster que B y B al mismo clúster que C, entonces A también debe estar en el mismo clúster que C. Este requisito de transitividad puede hacer que ciertos elementos disímiles terminen agrupados en el mismo clúster, lo que dificulta una separación clara entre grupos (Shalev-Shwartz & Ben-David, 2014).

Para entender este concepto de manera gráfica, consideremos cuatro conjuntos de puntos organizados en forma de cuadrado, como si fueran los vértices. En esta disposición, existen dos tipos posibles de agrupamiento: uno a lo largo del eje vertical y otro en el eje horizontal, ilustrados en la

Agrupamiento Vertical: Aquí, los puntos se agrupan en dos clústeres verticales (por columnas), lo cual puede responder a una similitud en una característica representada en el eje vertical.

- **Agrupamiento Horizontal**: Los puntos se agrupan en dos clústeres horizontales (por filas), lo que representa una similitud en una dimensión distinta, como el eje horizontal.

Ambos agrupamientos son válidos y reflejan diferentes criterios de similitud, mostrando cómo los datos pueden organizarse en formas diversas según se defina la similitud entre elementos. El clustering presenta la dificultad de equilibrar similitud y disimilitud, lo cual puede llevar a soluciones de agrupamiento variadas (Shalev-Shwartz & Ben-David, 2014).

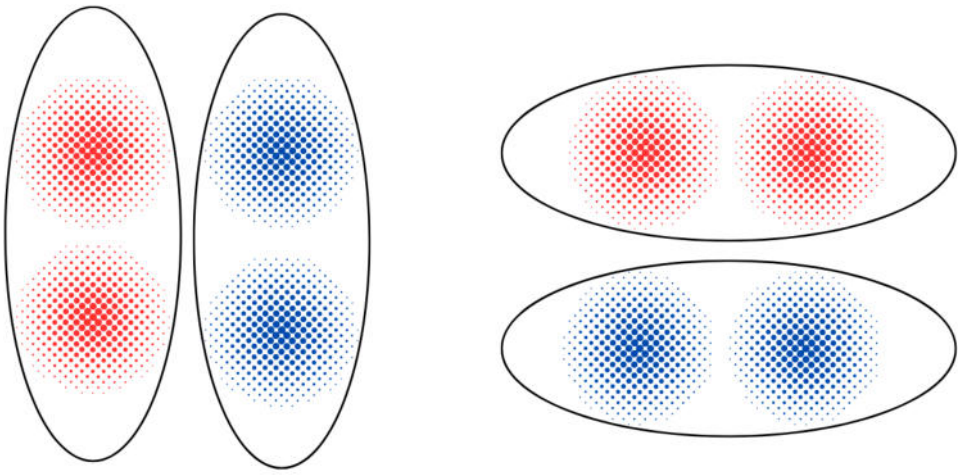
Figura *1*.

- **Agrupamiento Vertical**: Aquí, los puntos se agrupan en dos clústeres verticales (por columnas), lo cual puede responder a una similitud en una característica representada en el eje vertical.
- **Agrupamiento Horizontal**: Los puntos se agrupan en dos clústeres horizontales (por filas), lo que representa una similitud en una dimensión distinta, como el eje horizontal.

Ambos agrupamientos son válidos y reflejan diferentes criterios de similitud, mostrando cómo los datos pueden organizarse en formas diversas según se defina la similitud entre elementos. El clustering presenta la dificultad de equilibrar similitud y disimilitud, lo cual puede llevar a soluciones de agrupamiento variadas (Shalev-Shwartz & Ben-David, 2014).

Figura 1

Agrupamiento Horizontal y Vertical de Datos en Clusters



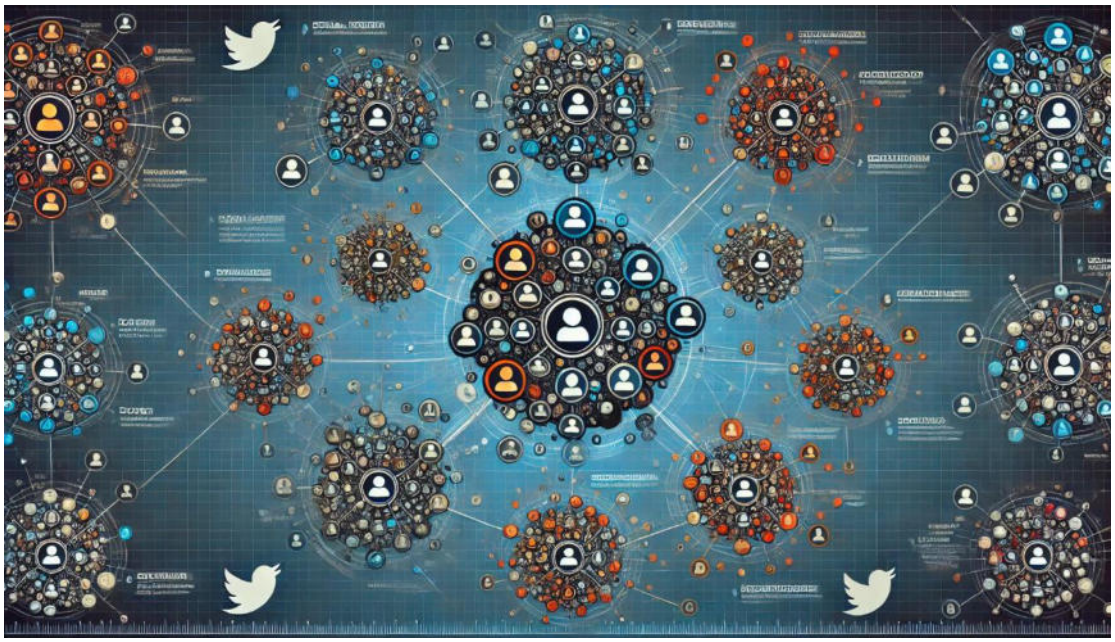
Nota: Representación esquemática del agrupamiento horizontal y vertical de datos. Elaborado por el autor.

Clustering en el Análisis de Redes Sociales

El clustering se ha convertido en una herramienta esencial para el análisis de datos obtenidos por redes sociales, permitiendo identificar comunidades y patrones de comportamiento a partir de la información obtenida por millones de usuarios a diario (García Diéguez, 2014). Esta vasta cantidad de datos tiene un denominador común: el lenguaje. Este instrumento fundamental de comunicación, en el cual todos los seres humanos lo utilizamos para comunicar conocimientos, emociones y opiniones (Yule, 2004). Sin embargo, en el contexto del análisis de datos, el lenguaje presenta un reto considerable. Los algoritmos de agrupamiento deben ser capaces de entender cómo se expresan los usuarios, en qué contexto se generan sus mensajes y qué información está circulando, para poder generar comunidades. En este sentido, el descubrimiento de comunidades y agruparlas en clusters se define intuitivamente como una colección de individuos con patrones de amistad densos internamente y escasos externamente (Mishra et al., 2007).

Figura 2

Visualización de Clustering en Redes Sociales en X



Nota: La imagen representa un modelo de clustering en redes sociales, mostrando cómo los usuarios de Twitter pueden agruparse en comunidades o clusters. Esta visualización se creó a partir de un prompt específico que describe la capacidad de los algoritmos de agrupamiento para identificar patrones de amistad densos internamente y escasos externamente, tal como se define en Mishra et al. (2007). Imagen generada con DALL-E (OpenAI, 2024).

Dado este contexto, el análisis de datos en redes sociales, potenciado por técnicas de clustering, nos ofrece una ventana hacia las interacciones humanas en línea. A medida que los usuarios comparten sus pensamientos y emociones en un entorno digital, el desafío radica en desentrañar el significado detrás de sus palabras y contextos (Velásquez-Gushiken, 2023). La capacidad de los algoritmos de agrupamiento para identificar patrones y comunidades dentro de esta complejidad lingüística no solo enriquece nuestra comprensión de los fenómenos sociales, sino que también proporciona herramientas cruciales para investigadores, mercadólogos y tomadores de decisiones. Por lo tanto, es fundamental continuar desarrollando y refinando estas técnicas, para asegurar que podamos captar y analizar adecuadamente la riqueza del lenguaje que se despliega en el vasto océano de información que es Internet.

Desafíos para encontrar Tendencias Políticas en X

El clustering es una muy buena herramienta en el análisis de redes sociales, especialmente cuando se trata de identificar comunidades y patrones de comportamiento entre millones de usuarios. En plataformas como X (anteriormente Twitter), esta permite explorar temas de conversación, revelando así las dinámicas sociales en torno a temas políticos. Sin embargo, el análisis de datos textuales provenientes de redes sociales presenta una serie de desafíos.

Originalmente concebida para que los usuarios compartieran pensamientos en tweets de hasta 140 caracteres, organizados cronológicamente, X ha evolucionado para permitir la publicación de imágenes, videos y enlaces que complementan el mensaje. Su facilidad para compartir contenido en tiempo real, tanto desde dispositivos móviles como computadoras, genera una gran cantidad de datos no estructurados disponibles para el análisis. No obstante, el estudio de esta información enfrenta grandes retos debido al lenguaje fragmentado y dinámico que se emplea en la plataforma, caracterizado por la brevedad de los mensajes y el uso frecuente de lenguaje de internet, abreviaturas, emojis y hashtags, lo cual dificulta el procesamiento y comprensión del contenido (Velásquez-Gushiken, 2023).

Además, el análisis se complica aún más cuando el contenido incorpora jergas políticas, sarcasmo, emociones y variaciones lingüísticas según la región. Los algoritmos de clustering deben ser capaces de interpretar el contexto, el tono y el significado de cada mensaje para identificar comunidades de opinión reales. La diversidad en los estilos de expresión, adaptados a temas actuales o eventos específicos, requiere una constante adaptación en las técnicas de procesamiento y análisis.

En este contexto, el clustering en redes sociales permite capturar estructuras de comunidad, es decir, grupos de usuarios conectados por temas e intereses comunes, incluidas las discusiones sobre política. Para que sea efectivo, es fundamental superar estos desafíos técnicos y semánticos, de modo que los algoritmos comprendan tanto el contexto lingüístico como la estructura social de los datos.

Técnicas de Preprocesamiento de Datos de Texto

Para analizar datos textuales en redes sociales como Twitter, es necesario preprocesar los datos, transformando los textos crudos en una representación efectiva para que los algoritmos de clustering pueden interpretarlos de manera coherente, lo que mejora la calidad y precisión de los análisis posteriores. Este trabajo aborda dos etapas principales en el preprocesamiento de datos de texto que son:

- La limpieza y preparación de los datos
- La vectorización del texto

Limpieza y Preparación de los Datos

Independientemente del método que utilicemos para obtener tweets para su análisis (API, web scraping o conjunto de datoss sin procesar), es común encontrar una gran cantidad de "ruido" que dificulta el análisis y puede introducir sesgos en los resultados. Este ruido puede distorsionar los resultados al agrupar incorrectamente tweets de diferentes temas o al generar ambigüedades en el análisis. Por ejemplo, si los hashtags populares no son eliminados o gestionados correctamente, pueden agrupar en un mismo cluster tweets que en realidad tratan de temas distintos, creando una representación errónea de las opiniones de los usuarios. Además, elementos como emojis y abreviaturas pueden llevar a interpretaciones equivocadas si el algoritmo no capta su significado original.

Este ruido incluye elementos como:

- Emojis y símbolos: Estos pueden alterar el tono de un mensaje. Por ejemplo, un emoji de risa después de un comentario político podría cambiar el sentimiento interpretado.
- Menciones y hashtags: Aunque los hashtags y menciones pueden ser relevantes, su exceso o uso en contextos irrelevantes puede introducir sesgo en los datos.

- URLs y enlaces: Son comunes en los tweets y, si no se eliminan, pueden confundir el análisis al introducir contenido externo irrelevante.
- Errores y abreviaturas: En redes sociales, los usuarios suelen utilizar abreviaturas y cometer errores ortográficos, lo que complica la interpretación de los datos y puede llevar a una categorización incorrecta.

Vectorización del Texto

Una vez que los datos están limpios, el siguiente paso es convertir el texto en representaciones numéricas mediante la vectorización. A continuación, se analiza tres métodos de vectorización: TF-IDF, Word2Vec y Tweet Tokenizer. Cada uno tiene sus particularidades y es útil en diferentes contextos.

TF-IDF (Term Frequency-Inverse Document Frequency). TF-IDF es una técnica común en el procesamiento de texto para convertir palabras en valores numéricos que reflejan su importancia relativa en un conjunto de documentos. TF-IDF se compone de dos partes:

Term Frequency (TF). La frecuencia de término es el número de veces que una palabra t aparece en un documento d. Matemáticamente, se calcula como:

$$TF(t,d) = \frac{Número\ de\ veces\ que\ t\ aparece\ en\ d}{Total\ de\ palabras\ en\ d} \qquad (4)$$

Este valor nos dice qué tan frecuente es una palabra en un documento específico, normalizando la frecuencia con respecto a la longitud del documento.

Inverse Document Frequency (IDF). La frecuencia inversa de documentos ajusta la relevancia de palabras que son comunes en todos los documentos (como "el", "de", etc.). Se calcula como:

$$IDF(t) = log\left(\frac{N}{1 + Número\ de\ documentos\ que\ contienen\ t}\right) \quad (5)$$

Donde N es el total de documentos en el conjunto. Al tomar el logaritmo de esta relación, se reduce el peso de palabras muy comunes y se aumenta el peso de palabras que son menos frecuentes en la colección completa.

TF-IDF Score. El puntaje TF-IDF de una palabra en un documento se obtiene multiplicando los valores de TF e IDF:

$$TF - IDF(t, d) = TF(t, d) \times IDF(t) \quad (6)$$

Este método se enfoca en la frecuencia de las palabras, dando mayor peso a términos que son únicos en un contexto específico, pero comunes dentro del conjunto de datos. TF-IDF es rápido y efectivo para resaltar palabras clave, pero no captura relaciones semánticas. Así, puede resultar útil en un análisis preliminar donde el contexto no es tan relevante, pero para temas complejos, sus limitaciones semánticas pueden llevar a una interpretación incompleta (Suresh et al., 2023).

Word2Vec. Word2Vec es una técnica basada en redes neuronales que convierte palabras en vectores, capturando las relaciones semánticas y contextuales entre ellas. Utiliza un modelo de red neuronal superficial, para aprender representaciones vectoriales de palabras. Existen dos arquitecturas principales en Word2Vec:

- CBOW (Continuous Bag of Words): Predice una palabra central utilizando su contexto, es decir, las palabras vecinas. La arquitectura del modelo elimina la capa oculta no lineal y comparte la capa de proyección para todas las palabras, promediando sus vectores sin considerar el orden en el historial. También incorpora palabras futuras en el contexto, optimizando la probabilidad de clasificar correctamente la palabra central al observar cuatro palabras anteriores y cuatro palabras futuras como entrada.

- Skip-gram: Predice las palabras de contexto de una palabra dada, maximizando la probabilidad de que una palabra específica esté cerca de su contexto real. A diferencia de CBOW, Skip-gram usa la palabra actual como entrada para clasificar palabras dentro de un rango de contexto. Aunque un rango más amplio mejora la calidad de los vectores, aumenta la complejidad computacional. Para equilibrar esto, el modelo asigna menor peso a las palabras más distantes, ya que suelen estar menos relacionadas con la palabra central.

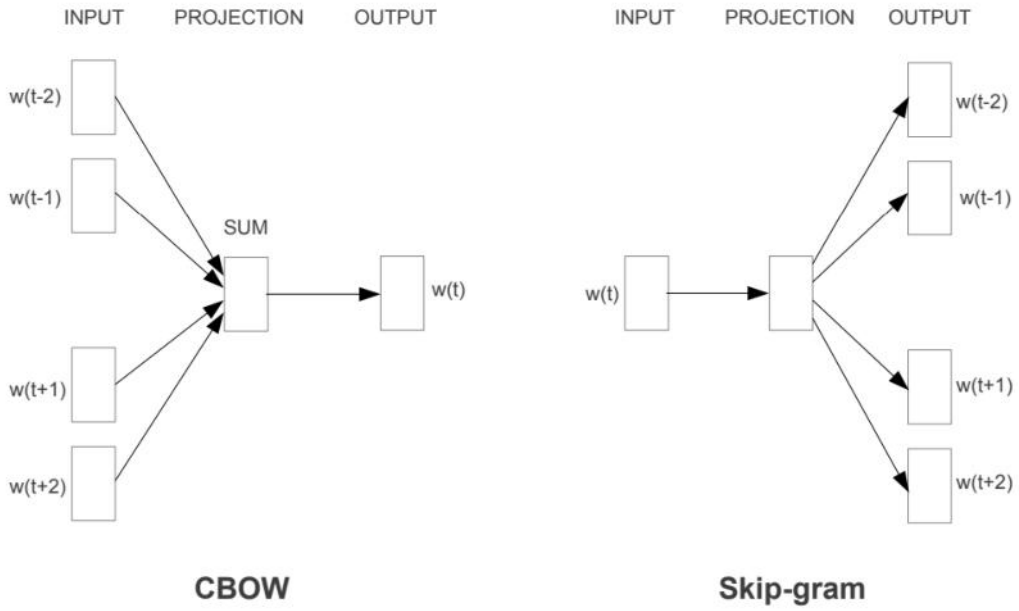
La Figura *3* lustra las arquitecturas de CBOW y Skip-gram. En CBOW, las palabras vecinas de la palabra objetivo (representadas como $w(t-2), w(t-1), w(t+1), w(t+2)$) se proyectan en una capa compartida y se promedian para predecir la palabra central $w(t)$. Este modelo, insensible al orden de las palabras de contexto, funciona como una "bolsa de palabras", enfocándose en la presencia de palabras vecinas en lugar de su secuencia.

En Skip-gram, el proceso es inverso: la palabra central $w(t)$ se utiliza para predecir sus palabras de contexto, también representadas como $w(t-2), w(t-1), w(t+1), w(t+2)$. La palabra central se proyecta en un vector para predecir cada palabra de contexto individualmente. Para mejorar la precisión y reducir el costo computacional, se asigna menor peso a las palabras más distantes, dado que es menos probable que estén directamente relacionadas con la palabra central(Mikolov et al., 2013).

Al finalizar el proceso de entrenamiento con Word2Vec, el resultado son vectores llamados word embeddings, que representan cada palabra en un espacio de menor dimensión. Estos embeddings capturan relaciones semánticas y contextuales entre las palabras, ubicando aquellas con significados similares en posiciones cercanas dentro de este espacio (Almeida & Xexéo, 2019). De este modo, el producto final de Word2Vec no solo son vectores numéricos, sino representaciones enriquecidas que preservan el contexto y significado de las palabras en el corpus original.

Figura 3

Arquitecturas de los modelos CBOW y Skip-gram en Word2Vec.



Nota: La arquitectura CBOW predice la palabra central en función del contexto, mientras que la arquitectura Skip-gram predice las palabras de contexto a partir de la palabra central. Adaptado de (Mikolov et al., 2013)

Tweet Tokenizer. Tweet Tokenizer es una herramienta de la biblioteca NLTK creada para tokenizar tweets de manera efectiva. Dado que los tweets contienen elementos característicos como hashtags, menciones, URLs y emojis, el preprocesamiento de este tipo de texto informal requiere un enfoque especializado. Tweet Tokenizer segmenta cada tweet en "tokens" o unidades significativas, permitiendo analizar cada componente de forma individual y facilitar su preparación para etapas posteriores del análisis (Amin Baybon, 2023; Gupta & Joshi, 2017).

Parámetros Configurables. Para adaptar el proceso de tokenización a las necesidades específicas del análisis, el Tweet Tokenizer incluye varios parámetros útiles (NLTK Project, 2024):

- preserve_case (bool): Controla si se conserva la capitalización del texto. Al activarlo (True), las palabras mantienen su formato original; si está desactivado (False), se convierten a minúsculas, facilitando el análisis uniforme de términos.
- reduce_len (bool): Reduce secuencias largas de caracteres repetidos (como "coooool") a un máximo de tres caracteres (quedando como "cooool") cuando está activado. Esto ayuda a normalizar el texto y evitar problemas con variaciones ortográficas.
- strip_handles (bool): Elimina las menciones de Twitter (@usuario) si está activado (True), lo que puede ser útil si estas no son relevantes para el análisis.
- match_phone_numbers (bool): Permite detectar y tokenizar números de teléfono si está activado (True), una opción útil en contextos en los que los números de teléfono son significativos.

Con el Tweet Tokenizer, un tweet como "#Política @usuario: Increíble resultado en las elecciones 2024! 🎉" se convierte en tokens como ["#Política", "@usuario", "Increíble", "resultado", "en", "las", "elecciones", "2024", " 🎉 "]. Este proceso permite analizar los elementos del tweet de forma granular, eliminando o conservando partes específicas según el propósito del análisis.

Algoritmos de Agrupamiento

Para identificar patrones de opinión y comunidades dentro de un conjunto de datos de tweets, se aplican diversos algoritmos de agrupamiento o clustering. Cada uno de estos algoritmos tiene enfoques, ventajas y limitaciones distintas en el análisis de texto, lo cual permite adaptarlos según las características de los datos y los objetivos específicos del estudio.

Algoritmos de clustering aplicables a texto

K-means. K-means es un algoritmo de clustering basado en particiones, ampliamente usado en el análisis de texto y otras áreas. Su objetivo es agrupar datos en k $clusters$, donde $\boldsymbol{k}$ es el

número de clusters definido previamente. El proceso comienza asignando aleatoriamente los datos a cada cluster y ajustando iterativamente los centroides para minimizar la variación dentro de cada grupo, siendo la distancia euclidiana la métrica de proximidad más común. Dependiendo de los datos, también pueden emplearse otras métricas, como la distancia de Manhattan o la del coseno. El nombre "k-means" se debe a que crea "k" agrupaciones y usa la media de los puntos de cada grupo para definir su centro, llamado centroide. Este centroide representa el valor promedio de todas las características dentro del grupo y actúa como punto de referencia para medir la cercanía de los demás puntos (Ahmad, 2024).

K-means es popular por su simplicidad, velocidad y escalabilidad. Su eficiencia computacional proviene de un proceso iterativo que ajusta los centroides hasta que representan fielmente a sus grupos, lo que permite aplicarlo a grandes conjuntos de datos (Ahmad, 2024).

- **Ventajas**: $K-means$ es eficiente en términos de tiempo de ejecución, especialmente en grandes conjuntos de datos, y es relativamente fácil de interpretar.
- **Desventajas**: Sin embargo, $K-means$ tiende a crear clusters de formas esféricas, lo que limita su efectividad en estructuras complejas o no lineales. Además, requiere definir previamente el número de clusters, lo cual puede ser una desventaja en contextos donde el número de comunidades no es evidente.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise). DBSCAN es un algoritmo de clustering basado en densidad que identifica estructuras de datos no lineales y detecta puntos de ruido, es decir, aquellos que no pertenecen a ningún cluster. En el análisis de tweets, DBSCAN es útil para descubrir comunidades de opinión densamente conectadas, ignorando los tweets aislados o menos relevantes (Deng, 2020).

DBSCAN define los clusters mediante dos parámetros clave: $epsilon$ (ε), que representa el

radio de búsqueda alrededor de cada punto, y $minPoints$, que es el número mínimo de puntos dentro del radio ε para que un punto sea considerado núcleo de un cluster. Con estos criterios, cualquier punto con al menos $minPoints$ vecinos en el radio ε es clasificado como punto central, mientras que aquellos con menos vecinos son considerados puntos de borde o ruido si están aislados (Rehman et al., 2014).

- **Ventajas:** DBSCAN tiene varias ventajas importantes. No requiere definir el número de clusters de antemano, lo cual es útil cuando se desconoce la estructura de los datos. Además, puede identificar clusters de diferentes formas y tamaños, por lo que es ideal para datos con estructuras complejas o no esféricas. Otra ventaja es su capacidad para manejar valores atípicos y ruido, ya que clasifica como "ruido" a los puntos que no cumplen con los criterios de densidad, evitando que estos afecten la precisión de los clusters principales.
- **Desventajas:** DBSCAN es sensible a sus parámetros principales, ε y $minPoints$, que pueden ser difíciles de ajustar, especialmente en conjuntos de datos complejos. Además, su rendimiento disminuye en datos de alta dimensionalidad, como textos vectorizados, ya que la métrica de distancia pierde efectividad y puede dificultar la identificación correcta de clusters.

Agrupamiento jerárquico. El agrupamiento jerárquico es un método de clustering que organiza los datos en una estructura jerárquica, permitiendo explorar la relación de proximidad entre los elementos sin necesidad de definir el número de clusters de antemano, a diferencia de algoritmos como K-means. Este enfoque es especialmente útil cuando se desconoce la estructura de los datos y se desea analizar cómo se agrupan de manera natural. Dentro del clustering jerárquico existen dos enfoques: aglomerativo y divisivo. En el método aglomerativo, cada elemento comienza como un cluster independiente y los clusters más cercanos se fusionan iterativamente hasta formar un único cluster que contiene todos los datos. Este proceso se representa mediante un dendrograma, que ilustra

visualmente cómo se fusionan los clusters en diferentes niveles, mostrando las relaciones de similitud de una manera intuitiva. En contraste, el método divisivo realiza el proceso inverso, iniciando con un único cluster que se divide progresivamente, pero esta variante suele ser menos utilizada debido a su complejidad computacional en comparación con el método aglomerativo (Godoy Viera, 2017; Soria-Olivas et al., 2023).

- **Ventajas.** El clustering aglomerativo tiene varias ventajas. No requiere definir el número de clusters al inicio, lo que lo hace útil para explorar datos desconocidos. Su estructura jerárquica permite visualizar las relaciones entre elementos en distintos niveles mediante un dendrograma, lo cual ayuda a entender mejor las conexiones entre temas o categorías. Esta flexibilidad en la visualización es especialmente útil para datos con formas o tamaños de clusters variados.
- **Desventajas.** El clustering aglomerativo también tiene limitaciones. Es computacionalmente costoso, sobre todo en grandes conjuntos de datos, ya que necesita calcular y actualizar las distancias entre clusters en cada fusión. Además, su rendimiento puede verse afectado en datos de alta dimensionalidad, como los textos vectorizados, debido a la dificultad de seleccionar correctamente la medida de similitud y el método de fusión de clusters adecuados.

En esta sección se presentaron tres algoritmos de clustering ampliamente utilizados para analizar y agrupar datos sin supervisión**:** K-means**,** DBSCAN **y** Agglomerative Clustering**.** Cada uno de estos métodos ofrece un enfoque distinto para detectar patrones en los datos, con fortalezas y limitaciones específicas según el tipo de datos y el contexto de aplicación. La Tabla 1 proporciona una comparación detallada de estos algoritmos, resaltando sus principales características, aplicaciones y limitaciones, para facilitar la elección del método más adecuado según los objetivos del análisis.

Tabla 1
Comparación de algoritmos de clustering: K-means, DBSCAN y Agglomerative Clustering

Característica	K-means	DBSCAN	Clustering jerárquico aglomerativo
Definición de clusters	Debe especificarse el número de clusters al inicio	No requiere definir el número de clusters	No requiere definir el número de clusters
Métrica de distancia	Debe especificarse el número de clusters al inicio	No requiere definir el número de clusters	No requiere definir el número de clusters
Forma de clusters	Tiende a formar clusters esféricos	Maneja clusters de formas irregulares	Puede manejar clusters de diferentes formas
Sensibilidad a valores atípicos	Alta sensibilidad	Robusto frente a valores atípicos	Moderada sensibilidad a valores atípicos
Computación en grandes conjuntos de datos	Escalable y rápido	Moderado, depende del ajuste de parámetros	Costoso, especialmente en grandes conjuntos de datos
Aplicaciones comunes	Análisis de grupos definidos y datos bien separados	Detección de comunidades densas y clusters no lineales	Exploración de relaciones jerárquicas y datos complejos
Visualización	No proporciona visualización jerárquica	No proporciona visualización jerárquica	Dendrograma, permite ver relaciones entre clusters
Ventajas principales	Rapidez y simplicidad	Flexibilidad en la forma de los clusters y robustez frente al ruido	Estructura jerárquica sin necesidad de predefinir clusters
Desventajas principales	Sensible a la elección del número de clusters; requiere clusters esféricos	Sensible a los parámetros ε y minPoints; limita en alta dimensionalidad	Computacionalmente intensivo en grandes conjuntos; difícil en alta dimensionalidad

Nota. Esta tabla compara los tres algoritmos de clustering en términos de configuración, capacidad de manejo de datos complejos, sensibilidad a valores atípicos y aplicaciones comunes, ayudando a seleccionar el más adecuado según el tipo de datos y los objetivos del análisis.

Evaluación de Algoritmos de Agrupamiento

La evaluación de los algoritmos de agrupamiento es crucial para determinar su eficacia en la

identificación de patrones y comunidades en los datos de texto.

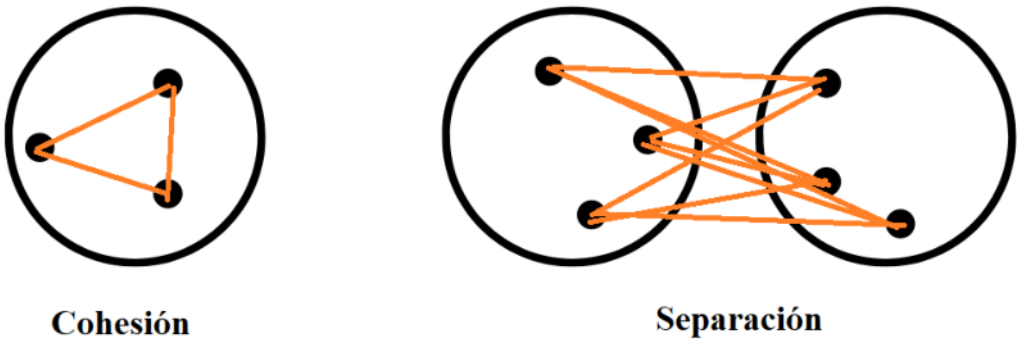
Métricas de Evaluación Interna

Son métricas en las que no se requieren etiquetas originales para evaluar los clústeres. Gira en torno a los siguientes dos tipos de métricas:

- La cohesión dentro de cada clúster.
- Separación entre los distintos clústeres.

Coeficiente de Cohesión y Separación. Estas métricas evaluan la compacidad (cohesión) de los clusters y la separación entre ellos. La cohesión mide qué tan compactos están los datos dentro de un cluster, mientras que la separación evalúa la distancia entre clusters diferentes. Estas métricas ayudan a verificar si los clusters son significativos y bien definidos en los datos de texto, permitiendo identificar comunidades de manera clara.

Figura 4
Evaluación de Cohesión y Separación en Clustering.



Nota: Representación gráfica de los conceptos de cohesión y separación en clustering. Elaborado por el autor.

Silhouette Score. El Silhouette Score es una métrica de evaluación interna que mide la coherencia de los clusters formados en un análisis de clustering. Su valor oscila entre -1 y 1, donde un valor cercano a 1 indica que los puntos están bien agrupados dentro de su cluster y bien separados

de otros clusters. Un valor cercano a 0 sugiere que los puntos están en el límite entre dos clusters, lo que indica una agrupación ambigua. Un valor negativo indica que los puntos pueden estar agrupados incorrectamente (Soria-Olivas et al., 2023).

Para calcular el Silhouette Score se siguen estos pasos:

- **Cálculo de a_i**: Se calcula como la media de las distancias entre un punto y todos los demás puntos del mismo cluster.
- **Cálculo de b_i**: Se calcula la mínima distancia media entre un punto y los puntos en otros clusters.
- **Cálculo de S_i**: El índice de Silhouette para cada punto se calcula $como\ (b_i - a_i)\ /\ max(a_i, b_i)$.
- **Cálculo de C**: El Silhouette Score general se calcula como la media de los S_i para todos los puntos en los clusters.

  Cuanto mayor es el Silhouette Score, mejor es la calidad de la agrupación, ya que se minimiza la distancia intracluster y se maximiza la distancia intercluster.

Índice de Davies-Bouldin. El Índice de Davies-Bouldin evalúa la relación entre la cohesión de los clusters (distancia dentro del cluster) y la separación entre clusters (distancia entre centroides). Su valor varía entre 0 y 1, donde valores más bajos indican clusters más ajustados y mejor separados (Ros et al., 2023).

Para calcularlo, se sigue esta fórmula:

- Se calcula la varianza dentro de cada cluster (σ_x) y la distancia media entre los puntos y el centroide del cluster.
- Se mide la distancia entre los centroides de dos clusters $(d(ci, cj))$.
- El índice final es la media de las relaciones entre las distancias de todos los clusters.

  Cuanto más cerca esté de 0, mejor será el resultado del clustering.

Métricas de Evaluación Externa

Las métricas de evaluación externa comparan los clusters generados con datos etiquetados o clasificaciones reales. Una de las métricas más utilizadas es el Índice de Rand Ajustado (ARI), que mide la similitud entre las agrupaciones obtenidas y las etiquetas reales. Un ARI alto indica una fuerte concordancia entre los clusters generados y las categorías de referencia. Sin embargo, estas métricas requieren que los datos estén etiquetados, lo cual no siempre es posible. En tales casos, se deben emplear métricas internas o técnicas de aprendizaje semisupervisado (Sundqvist et al., 2023).

Métricas de Evaluación Visual

La evaluación visual permite una inspección más intuitiva de los resultados del clustering, mediante herramientas como diagramas de dispersión, mapas de calor o dendrogramas. Estos métodos permiten a los analistas identificar patrones, outliers o inconsistencias que podrían pasar desapercibidos al utilizar únicamente métricas numéricas. La visualización proporciona un complemento importante a las métricas internas y externas, ofreciendo una perspectiva más completa de la calidad del agrupamiento.

Subjetividad en la evaluación de algoritmos

Aunque estas métricas cuantifican la calidad de los clusters, los resultados pueden ser subjetivos y depender de los objetivos específicos del análisis y de la naturaleza de los datos. Por ejemplo, en un análisis de opinión pública, puede ser más importante capturar la variedad de temas que crear clusters compactos. Es crucial ajustar las métricas y su interpretación al propósito específico de la investigación, ya que los patrones y clusters pueden variar en función de los intereses de estudio (Soria-Olivas et al., 2023).

A continuación, se presenta un resumen de las métricas aplicables a cada algoritmo:

Tabla 2
Resumen de Métricas de Evaluación por Algoritmo de Clustering

Algoritmo de Clustering	Métrica	Descripción
K-means	Errores cuadráticos medios	Mide la compacidad de los clusters evaluando la dispersión de los puntos alrededor de su centroide.
	Silhouette Score	Evalúa la coherencia de los clusters, indicando qué tan cerca están los puntos dentro del mismo cluster en comparación con otros clusters.
	Índice de Dunn	Calcula la relación entre la distancia mínima inter-cluster y la distancia máxima intra-cluster, útil para medir la separación de los clusters.
	Índice de Davies-Bouldin	Mide la relación entre cohesión y separación de los clusters, donde valores más bajos reflejan clusters más definidos.
DBSCAN	Silhouette Score	Útil para medir la calidad de los clusters formados en función de su densidad.
	Número de Clusters	Evalúa el número de clusters generados, importante dado que DBSCAN puede detectar clusters de tamaño variable.
	ARI (Adjusted Rand Index)	Si se dispone de etiquetas de referencia, ARI mide la concordancia entre el agrupamiento obtenido y el agrupamiento esperado.
	Distancias Intra e Inter-cluster	Mide la compacidad dentro de cada cluster y la distancia entre clusters, ayudando a evaluar la cohesión en clusters densos.
Clustering Jerárquico (Agglomerative Clustering)	Silhouette Score	Permite evaluar la coherencia y separación entre los clusters en diferentes niveles de la jerarquía.
	Índice de Davies-Bouldin	Valora la relación entre cohesión y separación de los clusters en cada nivel jerárquico.
	Índice de Dunn	Similar a K-means, mide la compacidad y separación de los clusters, evaluando la calidad del agrupamiento en la estructura jerárquica.
	ARI (si hay etiquetas)	Para datos etiquetados, ARI evalúa la precisión del agrupamiento en comparación con las categorías de referencia.

Nota: La tabla muestra una comparación de las métricas de evaluación que se pueden aplicar a diferentes algoritmos de clustering, cada una de las cuales aporta información clave sobre la cohesión, separación y precisión de los clusters formados.

Identificación de Patrones de Opinión y Comunidades

En el análisis de redes sociales, identificar patrones de opinión y comunidades permite

comprender mejor las interacciones y temas de interés entre los usuarios. Las siguientes secciones explican cómo el clustering facilita la identificación de estos patrones, con énfasis en la detección de trending topics.

Identificación de trending topics

El clustering es especialmente útil para detectar trending topics o temas emergentes de conversación en redes sociales. Estos trending topics reflejan discusiones actuales sobre temas relevantes, como educación, economía o violencia de género, y su identificación temprana permite a los investigadores y analistas en marketing político responder rápidamente a temas que capturan el interés público.

La identificación de trending topics mediante clustering no solo revela qué temas son de interés inmediato, sino que también permite anticipar posibles cambios en la opinión pública. Esto resulta clave para la toma de decisiones informadas y la planificación estratégica, ya que permite a las organizaciones y analistas responder de manera proactiva a las inquietudes y demandas emergentes de la sociedad, manteniendo sus mensajes en sintonía con el pulso de la conversación pública.

Comunidades políticas en redes sociales

Una comunidad en redes sociales se define como un grupo de usuarios conectados por intereses o temas comunes. En el caso de temas políticos, los algoritmos de clustering pueden identificar grupos de usuarios que comparten opiniones similares sobre temas específicos, como partidos políticos o figuras públicas. La detección de estas comunidades es valiosa para estudios de opinión, ya que permite observar cómo se agrupan las opiniones y cómo se influencian mutuamente los usuarios dentro de la plataforma.

Relevancia de los patrones de opinión

Identificar patrones de opinión en redes sociales es fundamental en estudios políticos y de opinión pública, ya que permite comprender el sentimiento colectivo y las preferencias de los usuarios. Estos patrones ofrecen información sobre temas de interés y revelan cómo evolucionan las opiniones en función de eventos o campañas específicas. Comprender estos patrones resulta crucial para estrategias de marketing, estudios sociopolíticos y toma de decisiones en políticas públicas.

Metodología

Para llevar a cabo esta investigación, se adoptó un enfoque de análisis de datos que combina técnicas de procesamiento de lenguaje natural (PLN) y algoritmos de clustering, con el objetivo de identificar comunidades de opinión en un conjunto de tweets en español sobre temas políticos. A continuación, se describen los pasos clave de la metodología:

Obtención del Conjunto de Datos

Se utilizó un dataset descargado que contiene tweets en español sobre temas políticos. Este conjunto de datos fue elegido por su relevancia y representatividad en el contexto de conversaciones políticas en español.

Preprocesamiento de los Datos

El conjunto de datos fue sometido a una serie de pasos de preprocesamiento para eliminar el "ruido" y preparar los tweets para el análisis de clustering. Este proceso incluyó la normalización de texto, eliminación de URLs, menciones y hashtags irrelevantes, conversión a minúsculas y manejo de emojis y abreviaturas. Se utilizó el Tweet Tokenizer, una herramienta especializada en tokenizar texto en redes sociales, para segmentar cada tweet en unidades significativas.

Aplicación de Algoritmos de Clustering

Se aplicaron y compararon tres algoritmos de clustering ampliamente utilizados: K-means, DBSCAN y Agglomerative Clustering. Cada algoritmo fue configurado según los parámetros específicos que maximizan su efectividad en la identificación de patrones en datos textuales. En el caso de K-means, se utilizó el método del codo para identificar el mejor número de clusters. Para DBSCAN, se experimentó con diferentes valores de densidad y distancia para identificar comunidades de densidad variable. Para el clustering jerárquico, se utilizó un enfoque aglomerativo para construir dendrogramas que permitan analizar la estructura jerárquica de las relaciones entre usuarios.

Evaluación de los Resultados

Para evaluar la calidad de las agrupaciones, se utilizaron métricas de evaluación interna, como el Silhouette Score, el Índice de Dunn y el Índice de Davies-Bouldin, las cuales miden la coherencia y separación de los clusters sin necesidad de etiquetas predefinidas. Además, se utilizaron métricas de evaluación externa, como el Índice de Rand Ajustado (ARI), cuando las etiquetas de referencia estuvieron disponibles en el conjunto de datos.

Interpretación de Comunidades y Patrones

Finalmente, se interpretaron los clusters resultantes para identificar comunidades de opinión y analizar los temas comunes en las conversaciones. Los patrones de opinión y los trending topics detectados permitieron entender la estructura de las interacciones y la alineación de usuarios en torno a temas específicos, contribuyendo a una comprensión más profunda de las dinámicas políticas en redes sociales.

Capítulo 3. Implementación y Análisis de Resultados

Introducción

El objetivo de este capítulo es detallar los métodos, herramientas y técnicas empleados para llevar a cabo el análisis comparativo de algoritmos de agrupamiento aplicados a un conjunto de tweets en español relacionados con temas políticos.

Con base en el marco teórico expuesto en el capítulo 2, la metodología desarrollada tiene como propósito evaluar la efectividad de diversos algoritmos en la identificación de comunidades de opinión y patrones semánticos presentes en las conversaciones políticas analizadas.

El proceso metodológico se estructura en varias etapas:

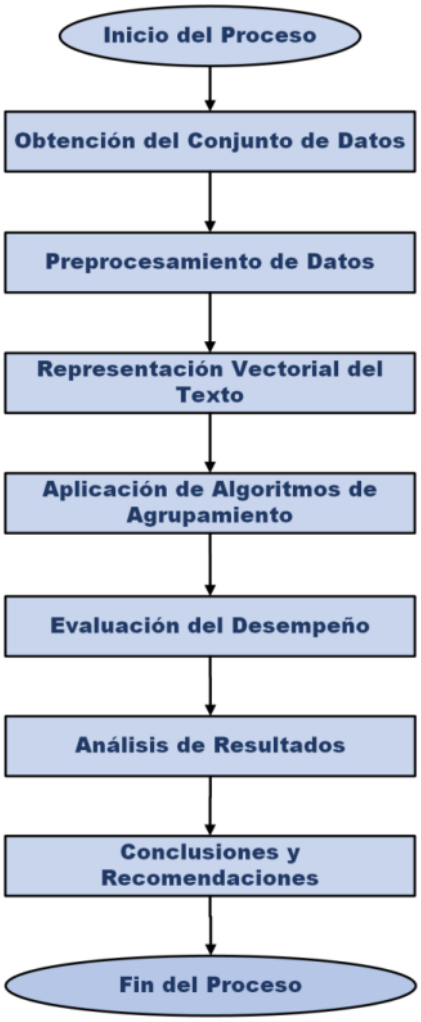
- Obtención y preprocesamiento de datos: Asegurando la limpieza y preparación adecuada del corpus de tweets.
- Representación del texto: Empleando técnicas como la vectorización (TF-IDF) y modelos semánticos.
- Aplicación de algoritmos de agrupamiento: Implementando métodos como K-Means, DBSCAN y Agglomerative Clustering.
- Evaluación y análisis de resultados: Comparando el desempeño de los algoritmos a través de métricas específicas y análisis de coherencia temática.

Los resultados obtenidos se presentan y analizan al final del capítulo, destacando las fortalezas y limitaciones de cada algoritmo en el contexto del estudio. Este análisis permite contrastarlos y determinar su efectividad para la identificación de patrones de opinión en redes sociales.

Una representación gráfica del procedimiento completo puede visualizarse en la **Figura *5***, la cual ilustra de manera secuencial cada una de las etapas metodológicas.

Figura 5

Flujograma del Proceso Metodológico para el Análisis Comparativo de Algoritmos de Agrupamiento



Nota: Este diagrama ilustra las principales fases involucradas en la comparación de distintos algoritmos de agrupamiento

Obtención del Conjunto de Datos

Fuente del Conjunto de Datos

El conjunto de datos utilizado en esta investigación se obtuvo de la plataforma Kaggle, específicamente del proyecto titulado *Tweets Política España* creado por Ricardo Moya (2021). Este corpus contiene tweets escritos en español sobre temas relacionados a política española pertenecientes a militantes de los cinco principales partidos políticos de España: PSOE, PP, VOX,

Unidas Podemos y Ciudadanos. La información fue recopilada por el autor a través de scraping utilizando la API de Twitter, siguiendo los términos y condiciones de la plataforma.

El conjunto de datos fue seleccionado por su capacidad para analizar dinámicas de opinión política en español. En particular:

- Relevancia temática: Refleja la actividad política en Twitter, una plataforma clave para la comunicación de partidos y representantes con el electorado.
- Representatividad: Incluye información de cinco partidos políticos principales en España, lo que permite estudiar diferencias discursivas y estructuras de comunidad.

Resumen de los Datos

El conjunto de datos consta de 245,789 tweets originalmente, pero después de realizar una eliminación de duplicados trabajamos con 245,116 tweets únicos.

El conjunto de datos incluye los siguientes campos principales:

- cuenta: Identificación hash de la cuenta que publica el tweet.
- partido: Afiliación política del autor del tweet.
- timestamp: Fecha y hora de publicación del tweet.
- tweet: Texto original del tweet.

Selección de los Datos

Para este análisis, se trabajó específicamente con la columna tweet, eliminando las demás columnas, ya que el objetivo principal de esta investigación es evaluar algoritmos de agrupamiento no supervisados. Esto permitió trabajar sin etiquetas predefinidas, adecuando el dataset a los requisitos de las técnicas de clustering. La columna seleccionada contiene 245,116 valores únicos, lo que asegura un corpus lo suficientemente amplio y variado para llevar a cabo un análisis significativo de patrones en las conversaciones políticas.

Preprocesamiento de Datos

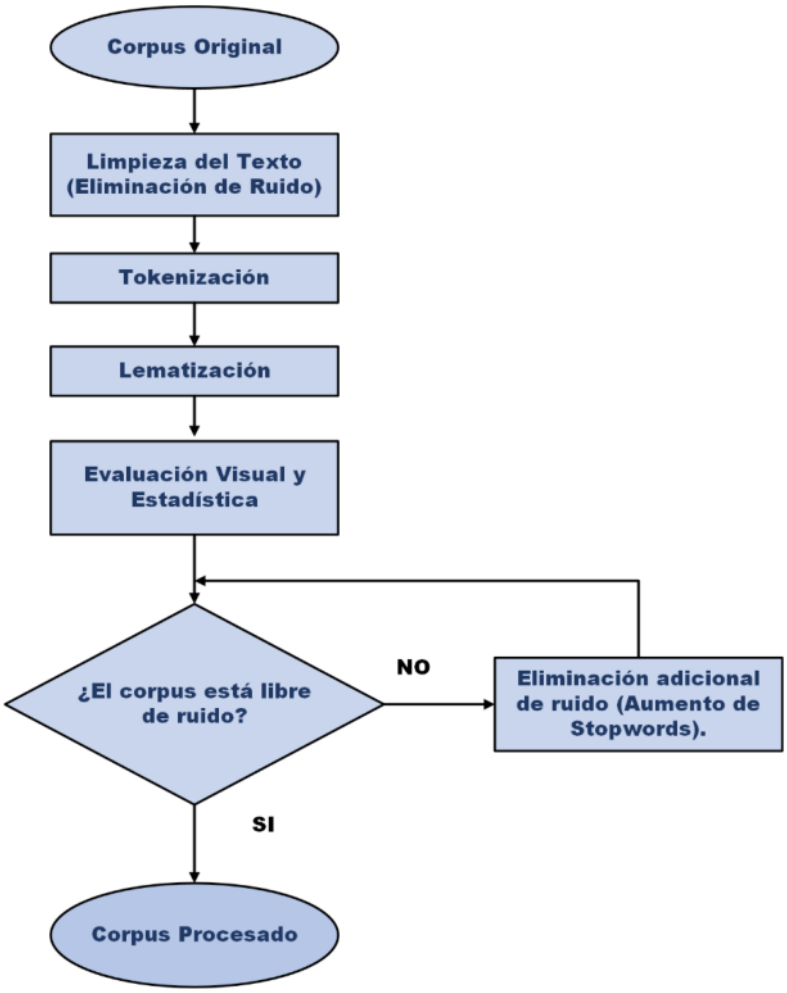
Metodología del Preprocesamiento

El preprocesamiento de datos se llevó a cabo siguiendo un flujo de trabajo sistemático representado en la **Figura *6***. Este proceso permitió transformar los datos crudos en una representación textual limpia y estructurada, adecuada para los análisis posteriores. Las etapas se desarrollaron de manera secuencial e iterativa, como se describe a continuación:

- Limpieza del Texto: Se eliminaron elementos irrelevantes como menciones, URLs, hashtags, caracteres especiales y palabras vacías (stopwords) para reducir el ruido en los datos.
- Tokenización: El texto fue segmentado en unidades individuales (tokens) para facilitar su procesamiento computacional.
- Lematización: Las palabras fueron transformadas a su forma base o raíz, permitiendo una normalización semántica del texto.
- Análisis Semántico y Estadístico: Se realizó una evaluación del corpus mediante análisis visual y estadístico para determinar la calidad y representatividad de los datos.

En caso de detectar ruido residual durante la evaluación, se aplicó un paso adicional de eliminación de ruido mediante un aumento en la lista de stopwords, garantizando así un corpus final optimizado.

Figura 6
Flujograma del Proceso de Preprocesamiento de Datos



Nota: Se muestra una representación gráfica de las etapas involucradas en la preparación de datos.

Resultados

El preprocesamiento permitió una transformación significativa en los tweets, como se observa en la Tabla 3. Cada etapa redujo la complejidad del texto, eliminó elementos irrelevantes y mejoró la representación semántica.

Tabla 3
Evolución de un Tweet en el Proceso de Preprocesamiento

Etapa	Resultado del Tweet
Original (Crudo)	@user ¡Qué increíble debate en #política hoy! https://abc
Limpieza del Texto	que increible debate politica hoy
Tokenización	['que', 'increible', 'debate', 'politica', 'hoy']
Lematización	['increíble', 'debate', 'política', 'hoy']

Nota: La tabla muestra las etapas de preprocesamiento de un tweet.

Análisis Estadístico

Tweets Únicos. Luego de aplicar todas las etapas del preprocesamiento, se obtuvo un total de 227,853 tweets únicos, lo que indica una reducción significativa de redundancias en el corpus.

Longitud Promedio de los Tweets. La longitud promedio de los tweets antes y después del preprocesamiento muestra una reducción del 34.0 %, lo cual evidencia la eliminación de ruido textual.

- Antes del preprocesamiento: 210.91 caracteres.
- Después del preprocesamiento: 139.21 caracteres.

Frecuencia de Palabras. Después de dos iteraciones de limpieza, las palabras más frecuentes en el corpus son:

- gobierno (31,788 apariciones),
- espana (17,441 apariciones),
- sanchez (15,426 apariciones),
- seguir (13,445 apariciones),
- espanol (11,792 apariciones),
- politico (11,419 apariciones),
- derecho (10,522 apariciones),
- ley (10,281 apariciones),
- madrid (10,045 apariciones),
- gracias (9,279 apariciones).

La nube de palabras presentada en la **Figura *7*** proporciona una representación visual de los términos con mayor frecuencia. Esta visualización confirma la relevancia temática del corpus, resaltando conceptos clave como “gobierno,” “espana,” “sanchez” y “politico”.

Figura 7

Nube de palabras de los términos más frecuentes

Nota: La nube de palabras fue generada a partir del análisis de frecuencia de los términos más comunes en los tweets recopilados.

La predominancia de términos como “gobierno,” “sanchez,” “espana” y “derecho” refleja la centralidad del corpus en temas políticos y legislativos, destacando discusiones relacionadas con actores políticos y asuntos de interés público. La nube de palabras confirma visualmente esta relevancia temática, al resaltar conceptos clave que evidencian el enfoque predominante del análisis en debates políticos y sociales.

Representación vectorial del Texto

La representación vectorial del corpus de tweets se realizó utilizando dos enfoques complementarios: el modelo TF-IDF (Term Frequency-Inverse Document Frequency) y Word2Vec, con el objetivo de transformar el texto preprocesado en estructuras numéricas aptas para análisis posteriores, como el agrupamiento y la extracción de patrones. A continuación, se describen los procedimientos, los hiperparámetros empleados y los resultados obtenidos.

TF-IDF

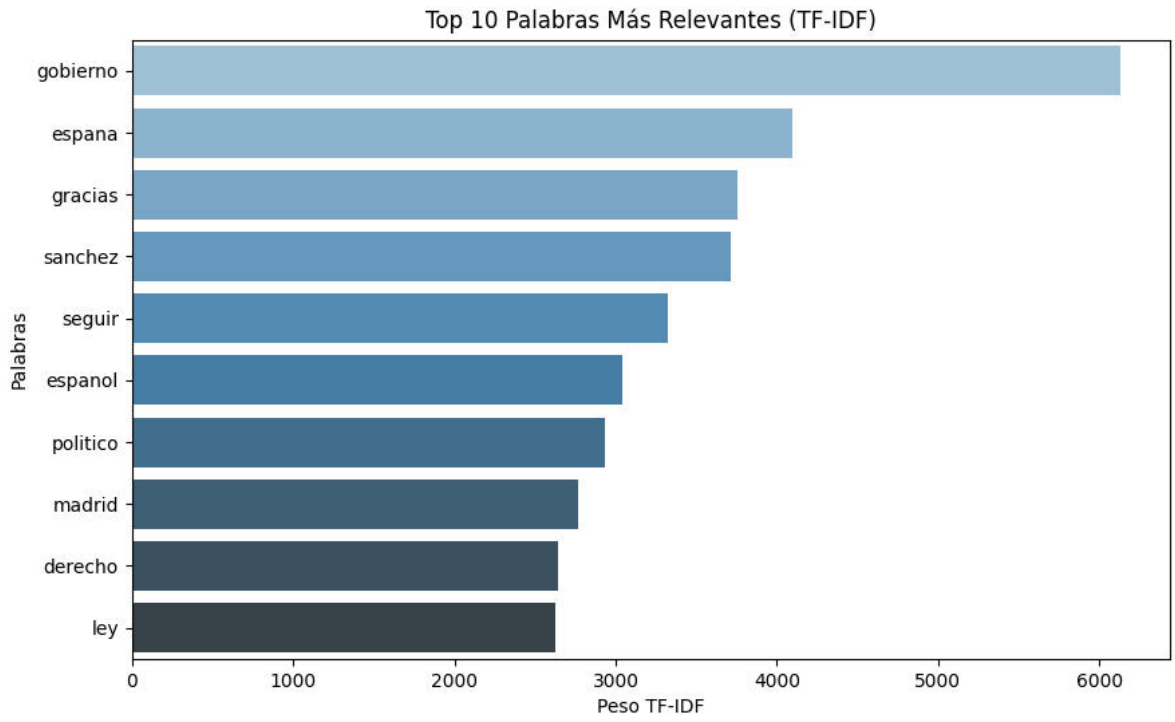
El modelo TF-IDF fue empleado para ponderar la relevancia de las palabras dentro del corpus, basado en su frecuencia local en cada documento (tweets) y su frecuencia global en el conjunto completo. Este enfoque permitió identificar los términos más representativos para el análisis temático.

Se configuraron los siguientes hiperparámetros para optimizar el modelo:

- max_features=1000: El modelo fue limitado a las 1000 palabras o combinaciones de palabras (n-gramas) más relevantes, lo que permite reducir la dimensionalidad del corpus sin perder información clave.
- min_df=10: Se descartaron palabras que aparecen en menos de 10 tweets, eliminando términos con baja frecuencia que podrían introducir ruido.
- max_df=0.85: Se excluyeron términos excesivamente frecuentes (presentes en más del 85% de los documentos), como conectores, que no aportan valor informativo.
- ngram_range=(1, 2): Se incluyeron unigramas y bigramas, capturando así tanto palabras individuales como relaciones contextuales simples.

El análisis del modelo reveló que las palabras más importantes del corpus, según su peso TF-IDF, fueron "gobierno," "españa," "gracias," y "sánchez." La visualización de estas palabras en la **Figura *8*** muestra su predominancia, destacando términos centrales en las discusiones políticas. Este enfoque permitió priorizar palabras que no solo son frecuentes, sino que además son informativas en relación con el tema general del corpus.

Figura 8
Distribución de los Pesos TF-IDF de las 10 Palabras más Relevantes del Corpus



Nota: El gráfico muestra la distribución de los pesos TF-IDF de las palabras más relevantes.

Word2Vec

El modelo Word2Vec se utilizó para obtener una representación distribuida de las palabras en un espacio vectorial de alta dimensionalidad, capturando relaciones semánticas entre términos. A diferencia de TF-IDF, que pondera la importancia basada en frecuencia, Word2Vec se enfoca en representar las palabras en función de su contexto dentro del corpus.

Los hiperparámetros configurados fueron:

- vector_size=200: Cada palabra fue representada en un vector de 200 dimensiones, lo que permite capturar información semántica detallada.
- window=10: Este parámetro amplió el contexto analizado a 10 palabras, lo que incrementa la capacidad del modelo para capturar relaciones a nivel de frase o ideas amplias.
- min_count=5: Se descartaron palabras con menos de 5 apariciones, eliminando términos poco frecuentes y potencialmente irrelevantes.
- workers=4: Se utilizó paralelización para acelerar el entrenamiento.

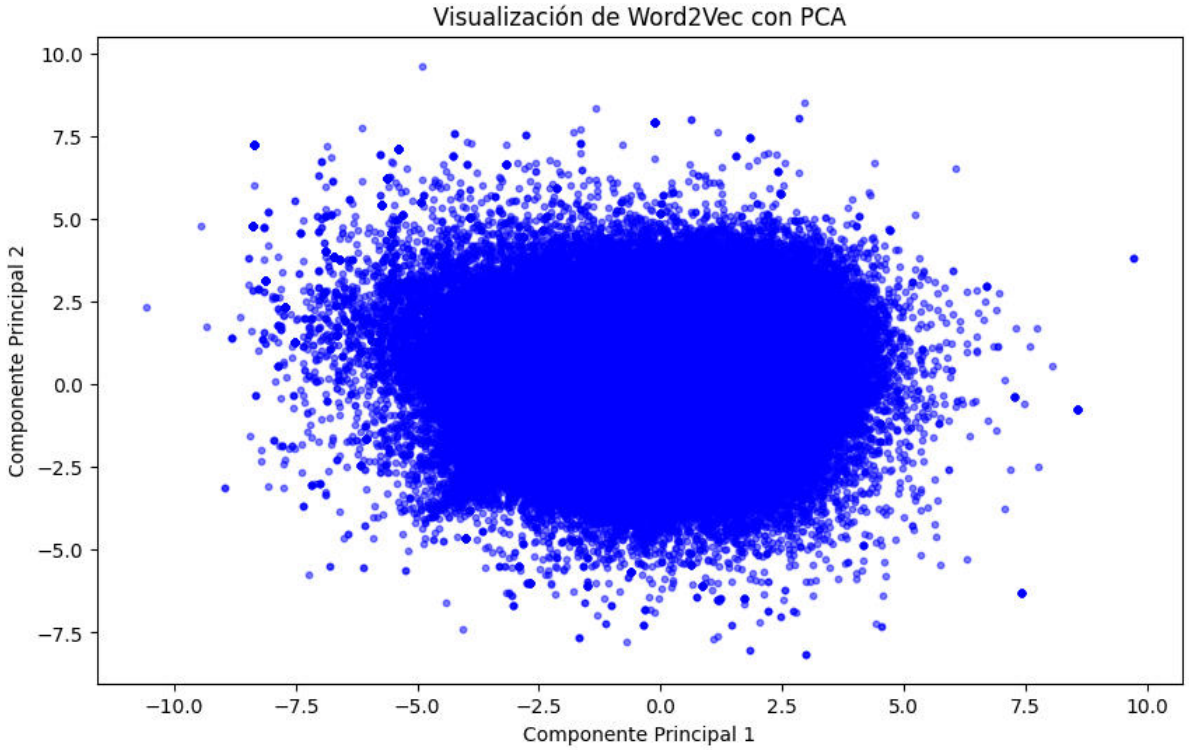
El modelo produjo una matriz de vectores con dimensiones (237.824,200), donde cada fila representa un tweet como el promedio de los vectores de las palabras que lo componen. Esto permitió generar una representación semántica condensada de cada documento.

Análisis de Resultados

La visualización y exploración de los resultados de Word2Vec se llevaron a cabo utilizando técnicas de reducción de dimensionalidad. En primer lugar, la proyección mediante PCA que se presenta en la

Figura *9*, mostró una dispersión controlada de los tweets, indicando que el modelo capturó patrones semánticos relevantes en el corpus. Esta dispersión sugiere que ciertos tweets comparten temáticas similares, mientras que otros presentan una mayor diversidad.

Figura 9
Proyección de los Vectores Word2Vec mediante PCA

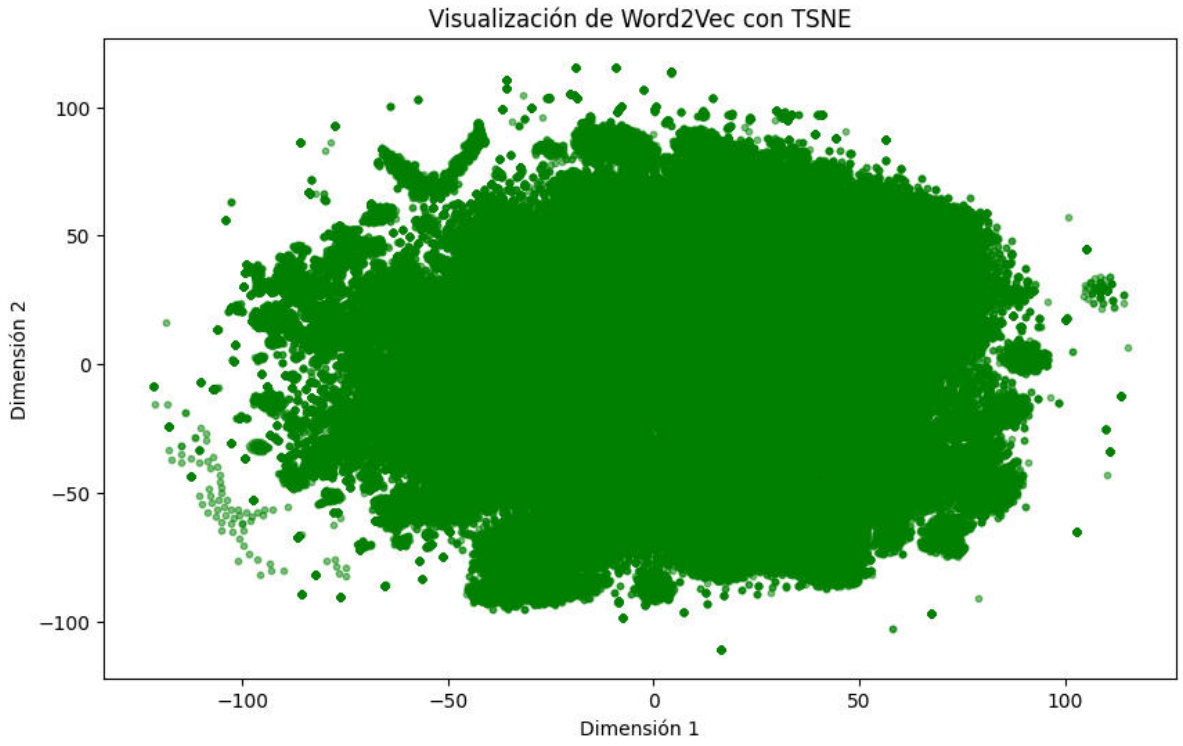


Nota: El gráfico muestra la proyección de los vectores de palabras en un espacio de dos dimensiones mediante el Análisis de Componentes Principales (PCA). Los vectores corresponden a un conjunto de

palabras relacionadas con el tema de política española. La proximidad entre las palabras refleja su similitud semántica en el espacio vectorial. Elaborado por el autor.

En segundo lugar, la visualización con t-SNE que se puede visualizar en la **Figura *10***, permitió una interpretación más detallada de las agrupaciones. Aunque este método es más sensible al ruido, reveló densidades claras en ciertas regiones del espacio, lo que sugiere comunidades de tweets con alto grado de similitud semántica.

Figura 10
Visualización de los Vectores Word2Vec utilizando t-SNE

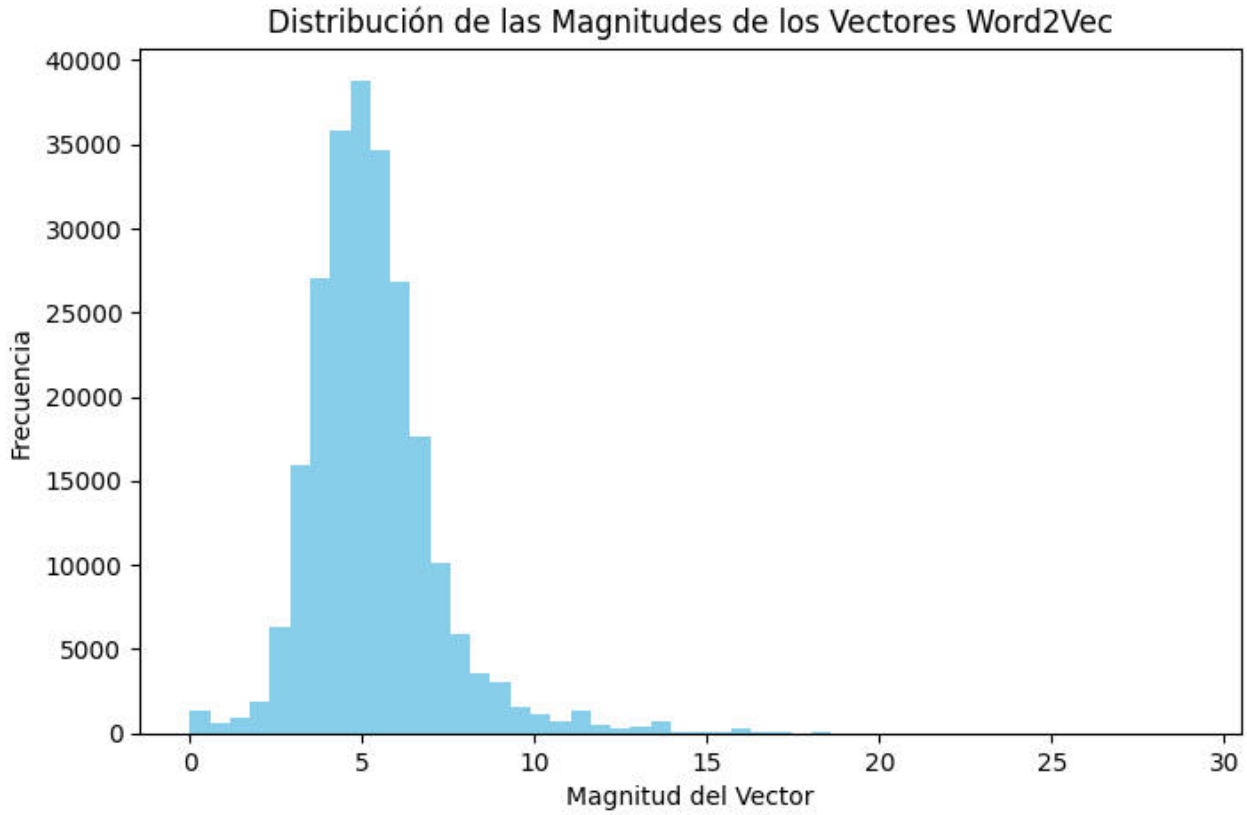


Nota: El gráfico muestra los vectores de palabras proyectados mediante t-SNE, basados en un modelo Word2Vec entrenado sobre un corpus de tweets políticos. Elaborado por el autor.

Por último, se evaluó la distribución de las magnitudes de los vectores (***Figura 11***), mostrando que la mayoría de los tweets se encuentran dentro de un rango medio de magnitudes, lo cual es consistente con un corpus bien balanceado en términos de información semántica.

Figura 11

Distribución de las Magnitudes de los Vectores Generados por Word2Vec



Nota: El gráfico muestra la distribución de las magnitudes de los vectores generados por el modelo Word2Vec, calculadas a partir de un corpus de tweets políticos. Elaborado por el autor.

La combinación de TF-IDF y Word2Vec proporcionó representaciones complementarias del corpus. Mientras que TF-IDF destacó palabras relevantes basadas en frecuencia y rareza, Word2Vec permitió capturar relaciones contextuales complejas. Las visualizaciones confirmaron que las representaciones obtenidas son adecuadas para tareas posteriores de análisis, como el agrupamiento de tweets o la identificación de comunidades. Estos resultados respaldan la robustez del preprocesamiento y la configuración de los hiperparámetros seleccionados.

Aplicación de algoritmos de agrupamiento

La implementación de algoritmos de agrupamiento no supervisados se enfocó en tres

métodos ampliamente utilizados: K-Means, DBSCAN y Agglomerative Clustering. La selección de estos algoritmos responde a su capacidad para identificar patrones en datos textuales vectorizados, aprovechando tanto las representaciones basadas en TF-IDF como en Word2Vec. En este estudio, cada algoritmo se implementó de manera rigurosa, considerando métricas de evaluación internas y la visualización de resultados en espacios reducidos dimensionalmente mediante PCA y UMAP.

Selección del Número Óptimo de Clústeres para K-Means

El algoritmo de K-Means requiere la especificación previa del número de clústeres (k). La determinación de un valor óptimo para k es un proceso muy relevante, ya que afecta directamente la calidad del agrupamiento. Para este estudio, se utilizaron dos métodos ampliamente reconocidos: el método del codo y el método de la silueta.

Método de la Silueta

El método de la silueta evalúa la calidad del agrupamiento mediante dos criterios principales: cohesión interna y separación externa. Este índice oscila entre -1 y 1, donde un valor cercano a 1 indica clústeres bien definidos, mientras que valores cercanos a 0 sugieren puntos ubicados en las fronteras entre clústeres, y valores negativos indican asignaciones incorrectas.

En este análisis, el Silhouette Score se calculó para diferentes valores de k (de 2 a 10). Como se observa en la

Figura *12*, el valor máximo del Silhouette Score se alcanzó en k=5, lo que sugiere que este número de clústeres proporciona un equilibrio óptimo entre cohesión y separación. A partir de k=6, el puntaje disminuye significativamente, indicando una pérdida en la calidad del agrupamiento.

Figura 12
Distribución del Silhouette Score para diferentes valores de k

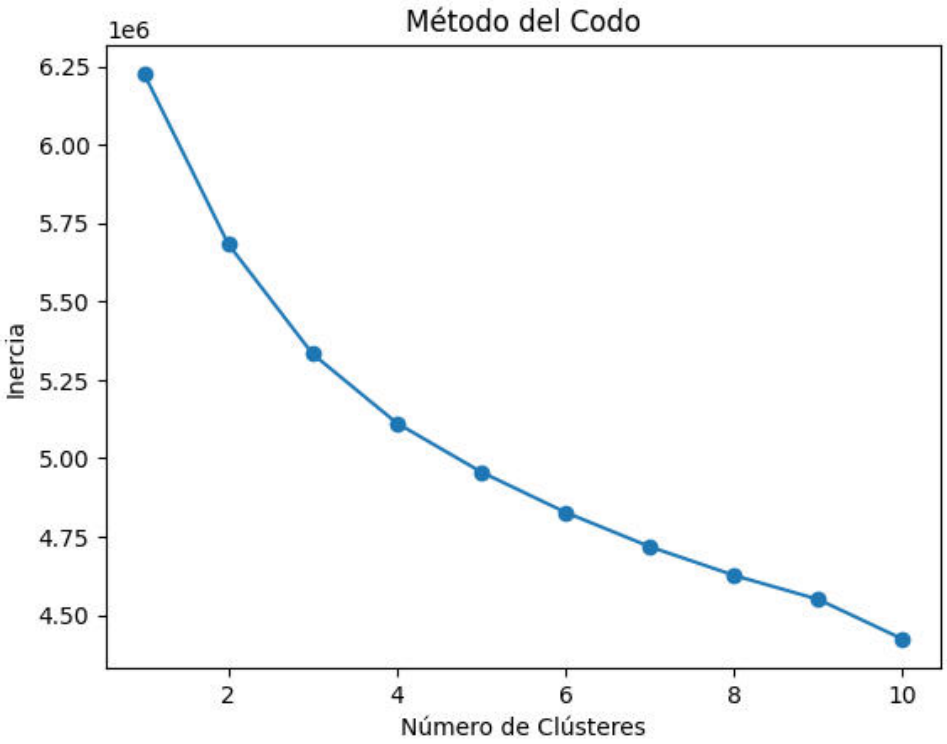
Nota: El gráfico muestra la variación del Silhouette Score obtenido para diferentes valores de k en un análisis de agrupamiento. El Silhouette Score mide la calidad de los clusters generados, donde valores cercanos a 1 indican una buena agrupación.

Método del Codo

El método del codo analiza la inercia, definida como la suma de las distancias cuadradas entre los puntos y sus respectivos centroides. Este método busca identificar el punto de inflexión en la curva de inercia, donde agregar más clústeres no mejora significativamente la agrupación.

En la Figura 2, se observa que el "codo" ocurre alrededor de k=5, lo que confirma la elección del número óptimo de clústeres. Después de este punto, la disminución en la inercia es menos pronunciada, indicando que aumentar el número de clústeres no aporta beneficios significativos al modelo.

Figura 13
Distribución del método del codo para diferentes valores de k



Nota: El gráfico muestra la relación entre el número de clusters (k) y la suma de los errores cuadráticos dentro de los clusters (WSS) en un análisis de agrupamiento utilizando el método del codo. El valor de k que minimiza la WSS antes de alcanzar una disminución marginal se considera el número óptimo de clusters.

Implementación de K-Means

Una vez determinado que k=5 es el número óptimo de clústeres, se procedió a implementar el algoritmo de K-Means con los siguientes parámetros:

- Inicialización: Se utilizó el método "k-means++", que mejora la selección inicial de los centroides al minimizar la probabilidad de que estos se encuentren demasiado cerca entre sí.
- Semilla aleatoria: El parámetro random_state=42 se incluyó para garantizar la reproducibilidad de los resultados.

- La implementación de K-Means se evaluó utilizando las métricas de Silhouette Score, Calinski-Harabasz Index y Davies-Bouldin Index, cuyos valores indicaron un agrupamiento moderado pero consistente con los datos vectorizados, tanto en TF-IDF como en Word2Vec.

Implementación DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

El algoritmo DBSCAN no requiere especificar un número fijo de clústeres, lo cual lo hace especialmente útil para conjuntos de datos en los que el número de agrupaciones no es evidente. En su lugar, utiliza dos parámetros clave:

1. **Radio de vecindad (eps)**. Se definió como eps=0.5, basado en un análisis de la densidad promedio del corpus. Este valor busca equilibrar entre identificar suficientes puntos vecinos para formar un clúster y evitar que se generen agrupaciones excesivamente densas.
2. **Muestras mínimas (min_samples).** Se estableció en min_samples=5, lo que garantiza que un grupo de puntos pueda formar un clúster siempre que al menos 5 puntos estén dentro del radio de vecindad.

La naturaleza basada en densidad del DBSCAN permite identificar clústeres de formas arbitrarias y manejar eficazmente el ruido en los datos. Sin embargo, esta metodología también implica que el algoritmo es altamente sensible a los valores de eps y min_samples, lo que puede requerir calibración para diferentes corpus.

La evaluación del rendimiento del algoritmo incluyó métricas como el Silhouette Score, el Calinski-Harabasz Index y el Davies-Bouldin Index.

Implementación Agglomerative Clustering

El algoritmo Agglomerative Clustering organiza los datos en una estructura de árbol o dendrograma, permitiendo observar cómo los clústeres se forman progresivamente. En este estudio,

se emplearon los siguientes parámetros clave:

- Enlace jerárquico (linkage): Se utilizó el método "ward", que minimiza la varianza dentro de los clústeres al fusionarlos, optimizando así la cohesión interna.
- Número de clústeres (n_clusters): Con base en los resultados obtenidos del análisis con K-Means, se definió el número de clústeres en k=5, asegurando consistencia y comparabilidad entre los algoritmos.

Agglomerative Clustering es especialmente valioso para explorar relaciones jerárquicas entre los datos, permitiendo una visualización intuitiva de cómo los puntos se agrupan. Su implementación se evaluó mediante las métricas de Silhouette Score, Calinski-Harabasz Index y Davies-Bouldin Index, mostrando resultados consistentes en ambas representaciones vectoriales.

Los datos TF-IDF ofrecieron clústeres bien definidos con una fuerte cohesión interna, mientras que Word2Vec reflejó relaciones semánticas más profundas, a pesar de mostrar una ligera superposición entre clústeres. La representación visual de los clústeres en PCA y UMAP reveló una buena diferenciación en el espacio bidimensional, respaldando la utilidad del algoritmo para datos textuales complejos.

Evaluación de algoritmos de agrupamiento

Evaluación de Métricas para K-means

El algoritmo K-Means fue evaluado utilizando dos representaciones vectoriales: TF-IDF y Word2Vec. Las métricas de evaluación empleadas, como el Silhouette Score, el Calinski-Harabasz Index y el Davies-Bouldin Index, permitieron analizar la calidad del agrupamiento en términos de cohesión interna y separación entre clústeres.

En la representación TF-IDF, los valores obtenidos fueron: Silhouette Score (0.0161), Calinski-Harabasz Index (102.3372) y Davies-Bouldin Index (7.3943). Estos valores indican una

calidad moderada en la separación de los clústeres, sugiriendo un nivel bajo de cohesión interna y una limitada separación entre grupos.

Por otro lado, Word2Vec demostró un desempeño superior, con métricas de Silhouette Score (0.0711), Calinski-Harabasz Index (1501.8709) y Davies-Bouldin Index (2.7189). Estos resultados reflejan clústeres más definidos, con mejor cohesión interna y una separación más clara, lo que sugiere que Word2Vec es más efectivo para capturar relaciones semánticas complejas en el conjunto de datos.

Visualización de Resultados para el algoritmo K-means

En la **Figura *14*** se presenta la distribución de los datos vectorizados con TF-IDF, proyectados en dos dimensiones mediante PCA, mostrando la asignación de clústeres realizada por K-Means. La **Figura *15*** ilustra la misma metodología utilizando Word2Vec como representación vectorial. En las **Figura *16*** y

Figura *17* se incluyen visualizaciones utilizando UMAP como técnica de reducción dimensional aplicada a las representaciones TF-IDF y Word2Vec, respectivamente.

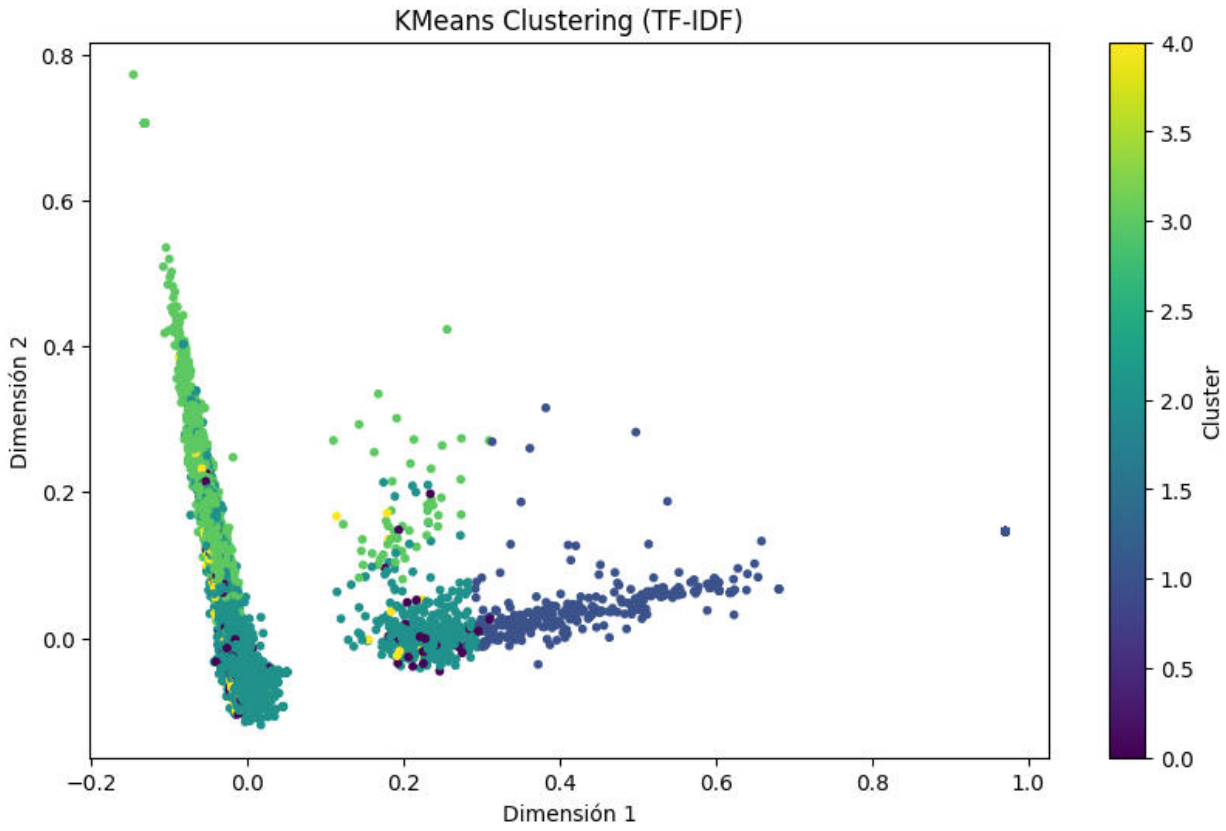
La comparación entre PCA y UMAP revela diferencias significativas en la capacidad para capturar relaciones entre los datos. PCA, al ser una técnica lineal, simplifica las dimensiones, pero puede no reflejar completamente las relaciones no lineales entre las palabras. Esto es evidente en las visualizaciones con TF-IDF (**Figura *14***), donde los clústeres muestran un alto grado de solapamiento, reflejando las métricas bajas de cohesión y separación. Por otro lado, Word2Vec (**Figura *15***) mejora la cohesión y separación de los clústeres, aunque algunos puntos aún están dispersos.

UMAP, al capturar relaciones no lineales, mejora la definición de los clústeres en ambas representaciones vectoriales. En la **Figura *16*** (UMAP con TF-IDF), algunos puntos están más agrupados en comparación con PCA, aunque el solapamiento de clústeres sigue siendo notable. La

Figura *17* (UMAP con Word2Vec) muestra clústeres más compactos y diferenciados, con menos puntos dispersos, reflejando las relaciones semánticas complejas que Word2Vec puede capturar. Esto refuerza la eficacia de UMAP combinado con Word2Vec para este conjunto de datos.

Figura 14

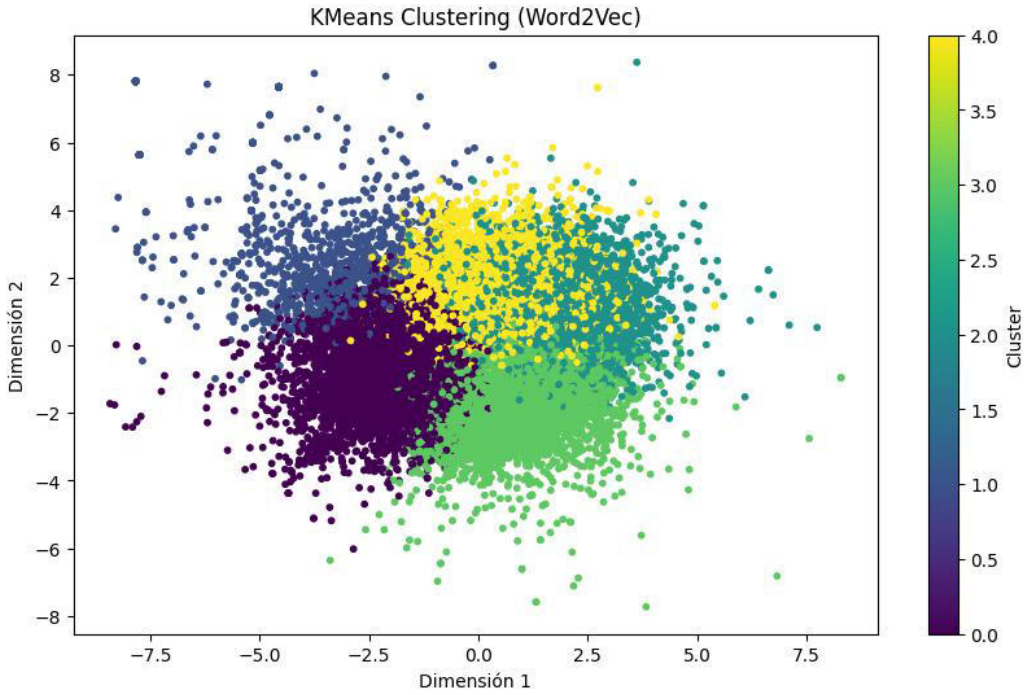
Visualización de los clústeres generados por K-Means con representación TF-IDF utilizando PCA



Nota: El gráfico muestra los clústeres generados por el algoritmo K-Means aplicado a una representación TF-IDF de los datos, proyectados en un espacio de dos dimensiones mediante PCA (Análisis de Componentes Principales). Los clústeres corresponden a un conjunto de tweets sobre política española. Los puntos del gráfico representan los documentos (tweets), y los colores indican a qué clúster pertenece cada uno. La proyección en 2D facilita la visualización de la agrupación semántica de los tweets. Elaborado por el autor.

Figura 15

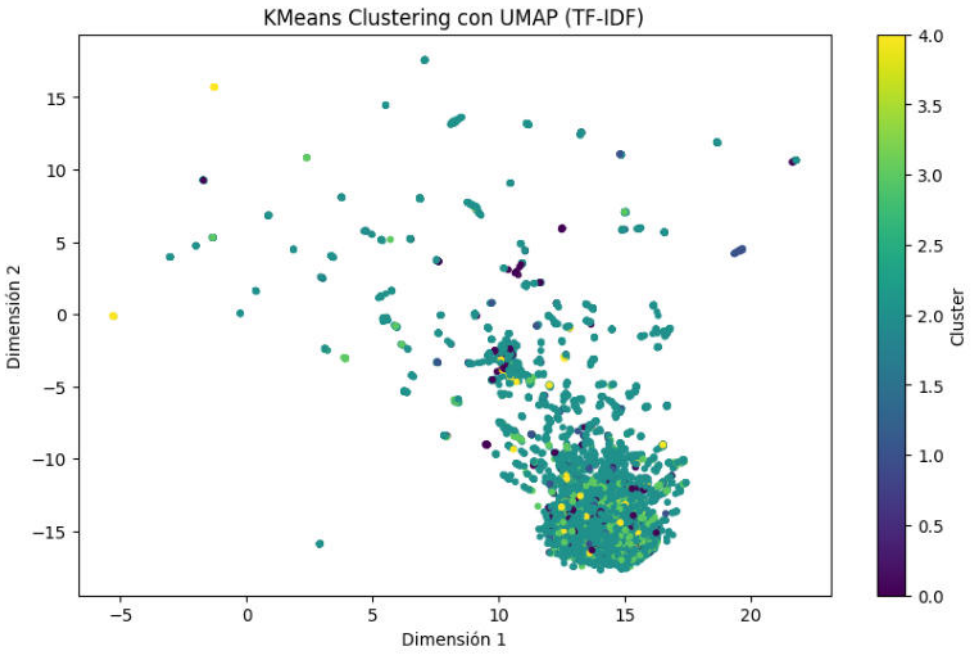
Visualización de los clústeres generados por K-Means con representación Word2Vec utilizando PCA



Nota: El gráfico muestra los clústeres generados por el algoritmo K-Means aplicado a la representación vectorial de las palabras mediante Word2Vec, proyectados en un espacio de dos dimensiones utilizando PCA (Análisis de Componentes Principales). Los clústeres corresponden a un conjunto de tweets sobre política española. Los puntos del gráfico representan los tweets, y los colores indican a qué clúster pertenece cada uno, destacando la agrupación semántica de los datos en función de las representaciones generadas por Word2Vec. Elaborado por el autor.

Figura 16

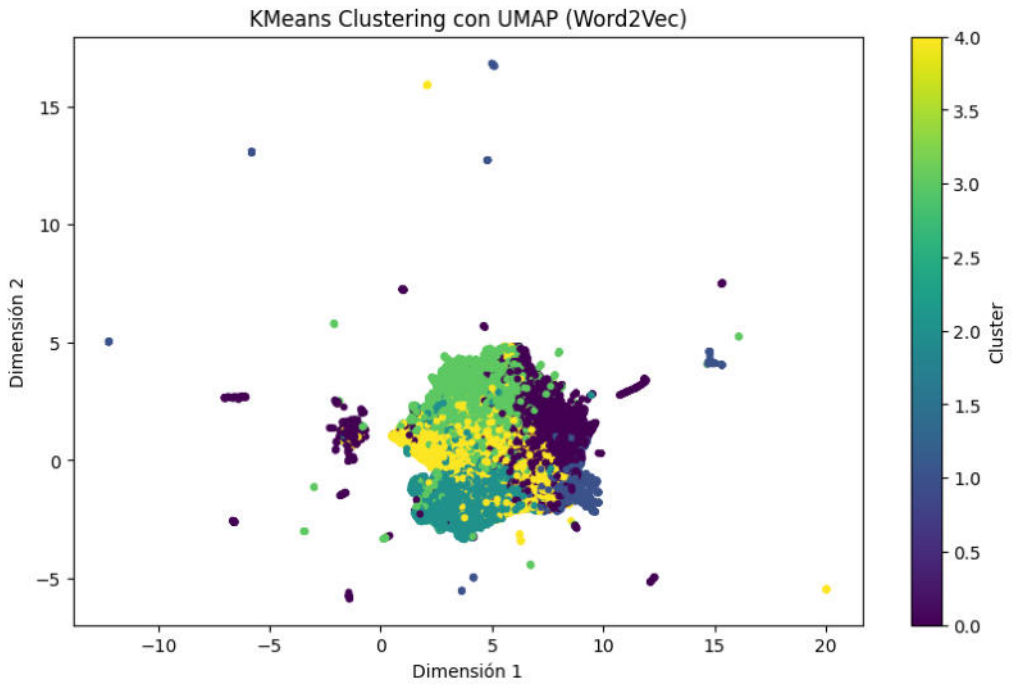
Visualización de los clústeres generados por K-Means con representación TF-IDF utilizando UMAP



Nota: El gráfico muestra los clústeres generados por el algoritmo K-Means aplicado a una representación TF-IDF de los datos, proyectados en un espacio de dos dimensiones utilizando UMAP (Uniform Manifold Approximation and Projection). Los clústeres corresponden a un conjunto de tweets sobre política española. Los puntos del gráfico representan los tweets, y los colores indican a qué clúster pertenece cada uno, destacando la agrupación semántica de los tweets en función de sus representaciones TF-IDF. Elaborado por el autor.

Figura 17

Visualización de los clústeres generados por K-Means con representación Word2Vec utilizando UMAP



Nota: El gráfico muestra los clústeres generados por el algoritmo K-Means aplicado a la representación vectorial de las palabras mediante Word2Vec, proyectados en un espacio de dos dimensiones utilizando UMAP (Uniform Manifold Approximation and Projection). Los clústeres corresponden a un conjunto de tweets sobre política española. Los puntos del gráfico representan los tweets, y los colores indican a qué clúster pertenece cada uno, revelando las relaciones semánticas entre los datos según sus representaciones generadas por Word2Vec. Elaborado por el autor.

Análisis Léxico por Clúster K-means

TF-IDF. La Tabla 4 resume las palabras clave representativas de cada clúster identificadas mediante TF-IDF. Estas palabras reflejan temáticas políticas, sociales y económicas que permiten interpretar los datos textuales de manera eficiente.

Tabla 4
Palabras clave representativas por clúster K-means (TF-IDF)

Clúster	Palabras Clave

0	poder, pasar, momento, seguir, madrid, espana, politico, entrevista, sanchez
1	gracias, querido, trabajo, amigo, mil, venir, apoyo, seguir, honor, companero
2	espana, sanchez, politico, seguir, madrid, espanol, via, derecho, ley, partido
3	gobierno, sanchez, espana, espanol, ley, espán, medida, gobierno espana, politico, seguir
4	social, agenda, dejar, red, justicia social, justicia, politica, gobierno, economico, seguir

Nota: La tabla muestra las palabras clave más representativas para cada clúster generado mediante el algoritmo K-Means, utilizando la representación TF-IDF de un conjunto de tweets sobre política española. Las palabras clave corresponden a los términos con mayor peso en cada clúster, reflejando las características semánticas y temáticas de los agrupamientos. Los valores de TF-IDF se utilizan para identificar las palabras más importantes dentro de cada grupo, lo que permite comprender mejor las temáticas predominantes en los clústeres. Elaborado por el autor.

Word2Vec. El análisis léxico con Word2Vec, presentado en la Tabla 5, muestra términos semánticamente ricos asociados a cada clúster. Esta representación enfatiza relaciones conceptuales más complejas en comparación con TF-IDF.

Tabla 5
Palabras clave representativas por clúster K-means (Word2Vec)

Clúster	Palabras Clave
0	maravillar, sintonizano, caseto, esperamo, armilla, klimt, politicasiguelo, acompanarar, llenazo, preparativo
1	felicidad, teson, gracia, dedicacion, agradecido, gracios, pasion, suerte, felicitacion
2	aumentar, comprometer, funcionado, certidumbre, reducir, incrementar, adecuado, agil, prioridad, beneficioso
3	aceptar, sistematicamente, unicamente, tolera, sentar, abandonar, chantaje, mercadear, apoye, instituciones
4	trabajaremos, garantiar, pai, fortalecer, necesitamos, plenamente, determinacion, paislo, compromisir, lograr

Nota: La tabla muestra las palabras clave más representativas de cada clúster generado mediante el algoritmo K-Means, utilizando la representación vectorial de palabras generada por Word2Vec. Los datos corresponden a un conjunto de tweets sobre política española. Las palabras clave reflejan los términos semánticamente más cercanos dentro de cada clúster, lo que permite identificar las temáticas predominantes según las representaciones vectoriales de las palabras. Este análisis muestra cómo las relaciones semánticas entre las palabras contribuyen a la formación de los clústeres. Elaborado por el autor.

La comparación entre TF-IDF y Word2Vec evidencia que la representación Word2Vec genera clústeres más cohesionados y con mayor riqueza semántica. Esto se refleja tanto en las métricas de evaluación como en las visualizaciones generadas. Word2Vec logra capturar relaciones conceptuales más complejas, lo que se traduce en clústeres más definidos y con menor solapamiento.

Las visualizaciones y métricas obtenidas indican que Word2Vec es más adecuado para tareas de análisis de texto donde la semántica juega un rol importante, mientras que TF-IDF puede ser útil en escenarios donde la frecuencia de términos sea prioritaria.

Evaluación de Métricas para DBSCAN

El algoritmo DBSCAN fue evaluado con las métricas internas estándar: Silhouette Score, Calinski-Harabasz Index y Davies-Bouldin Index. Los resultados obtenidos para ambas representaciones vectoriales muestran diferencias significativas en la calidad del agrupamiento.

DBSCAN (TF-IDF):

- Silhouette Score: -0.1879
- Calinski-Harabasz Index: 12.9216
- Davies-Bouldin Index: 1.2266

Estos valores indican que el agrupamiento no es ideal, ya que el Silhouette Score negativo sugiere que muchos puntos están más cerca de clústeres vecinos que del propio. El bajo Calinski-

Harabasz Index refleja que la separación entre los clústeres es limitada. Sin embargo, el Davies-Bouldin Index muestra cierta coherencia en la forma de los clústeres.

DBSCAN (Word2Vec):

- Silhouette Score: -0.2283
- Calinski-Harabasz Index: 53.3354
- Davies-Bouldin Index: 0.6386

En esta representación, el Silhouette Score es también negativo, reflejando una estructura de clústeres débil. Sin embargo, el Calinski-Harabasz Index es notablemente mayor que en TF-IDF, lo que sugiere una mejor definición de los clústeres. El Davies-Bouldin Index más bajo indica que los clústeres son más compactos y están mejor separados.

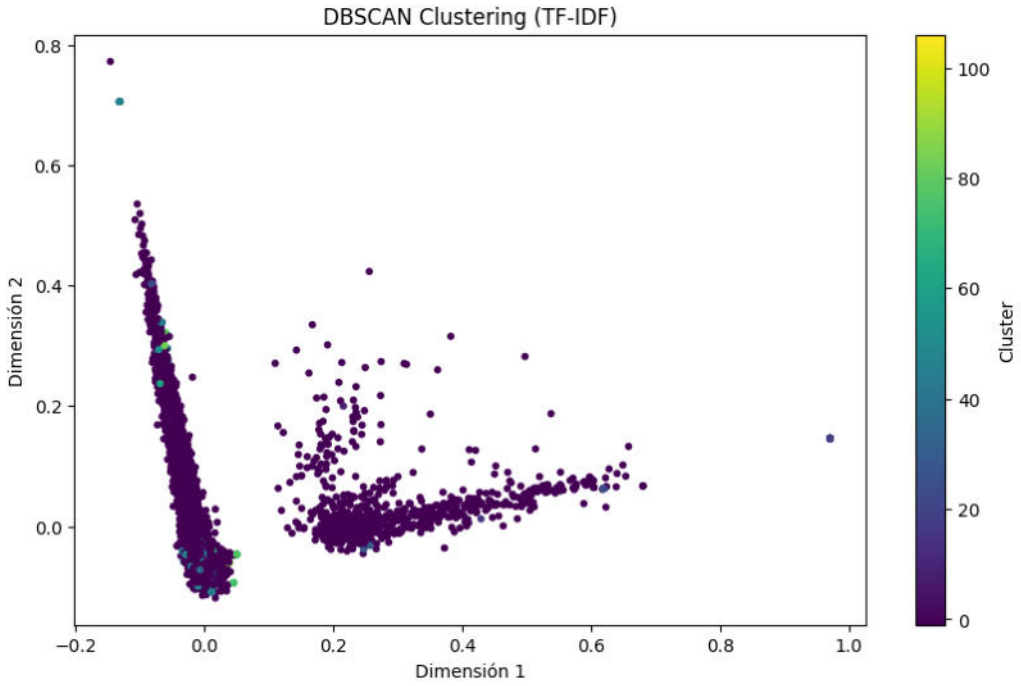
Visualización de Resultados para el Algoritmo DBSCAN

La proyección visual de los clústeres generados por DBSCAN se realizó utilizando dos métodos de reducción dimensional: PCA y UMAP. Estas visualizaciones permiten destacar las diferencias en la estructura y organización de los datos en función de las representaciones vectoriales utilizadas (TF-IDF y Word2Vec).

La representación en la

Figura 18 proyecta los datos vectorizados con TF-IDF a dos dimensiones mediante PCA. Los resultados muestran múltiples clústeres junto con una cantidad significativa de puntos clasificados como ruido (etiquetados como -1). La alta densidad en ciertas regiones refleja la capacidad de TF-IDF para identificar áreas densas de datos, aunque su naturaleza esparsa y de alta dimensionalidad limita la separación clara entre clústeres. Esto se traduce en una agrupación débil y una cantidad considerable de ruido.

Figura 18
Representación de los clústeres DBSCAN generados con TF-IDF y proyectados con PCA

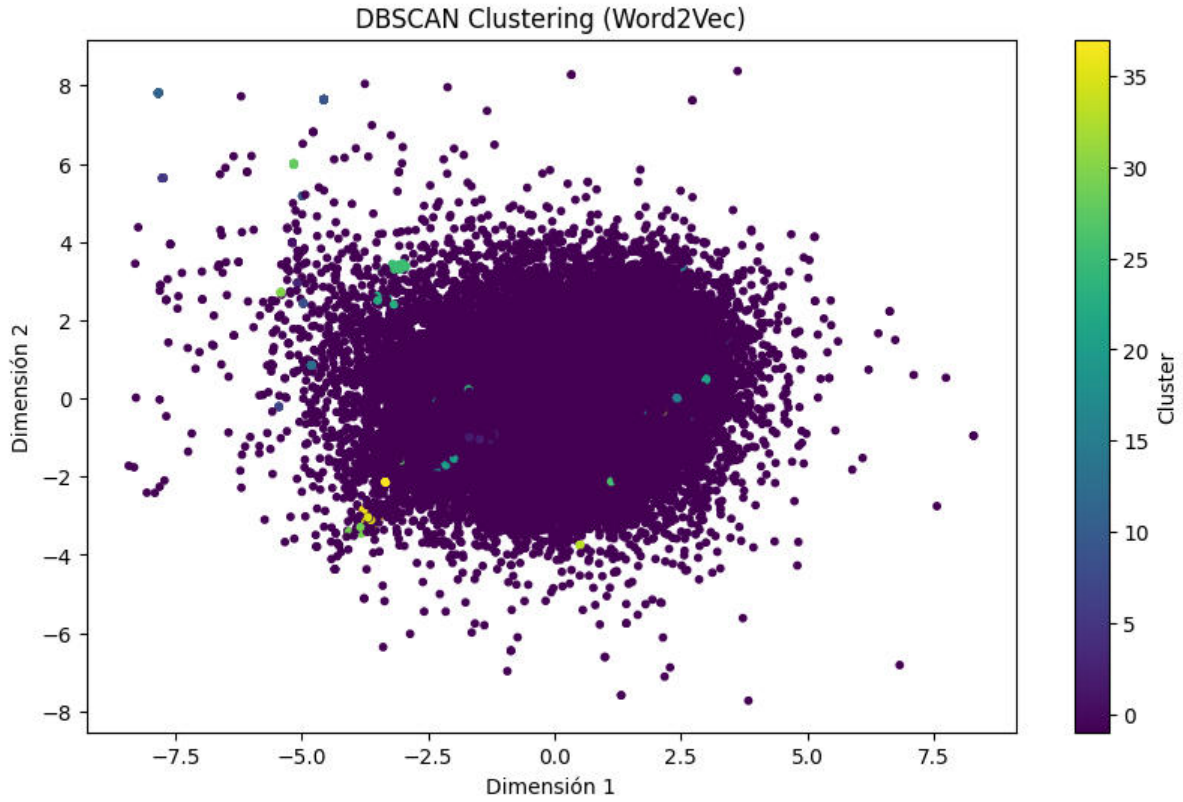


Nota: El gráfico muestra los clústeres generados por el algoritmo DBSCAN (Density-Based Spatial Clustering of Applications with Noise), aplicado a la representación TF-IDF de un conjunto de tweets sobre política española. Los clústeres se proyectaron en un espacio de dos dimensiones utilizando PCA (Análisis de Componentes Principales). Los puntos del gráfico representan los tweets, y los colores indican a qué clúster pertenece cada uno. DBSCAN permite identificar agrupamientos de alta densidad, así como puntos considerados ruido (que no pertenecen a ningún clúster). Esta representación facilita la visualización de la distribución de los datos y las relaciones semánticas entre los tweets, según sus representaciones TF-IDF. Elaborado por el autor.

En la **Figura *19***, los datos proyectados con PCA y representados mediante Word2Vec muestran clústeres más compactos y mejor definidos en comparación con TF-IDF. Aunque persisten puntos aislados clasificados como ruido, la capacidad de Word2Vec para capturar relaciones semánticas entre palabras mejora significativamente la cohesión dentro de los clústeres. No obstante, DBSCAN enfrenta desafíos para separar datos en áreas con baja densidad, lo que indica que los parámetros del algoritmo podrían optimizarse aún más.

Figura 19

Visualización clúster DBSCAN con Word2Vec y PCA



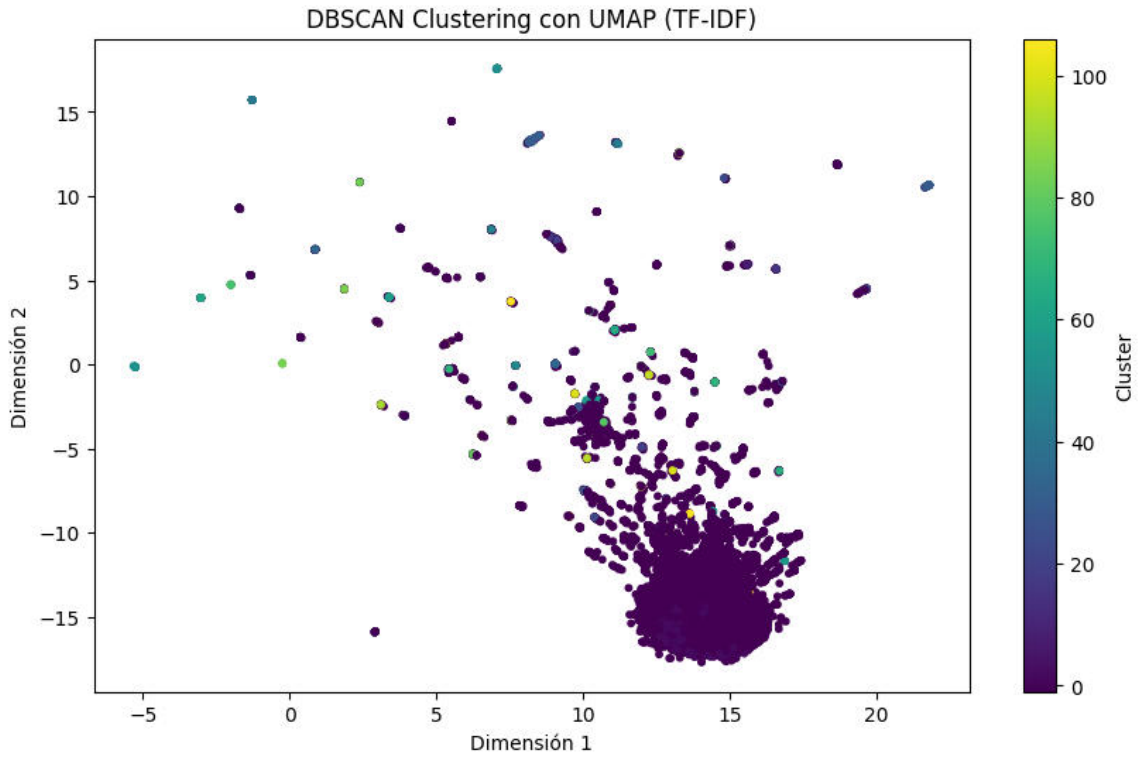
Nota: El gráfico muestra los clústeres generados por el algoritmo DBSCAN (Density-Based Spatial Clustering of Applications with Noise), aplicado a la representación vectorial de las palabras generadas por Word2Vec, proyectados en un espacio de dos dimensiones utilizando PCA (Análisis de Componentes Principales). Los puntos del gráfico representan los tweets sobre política española y los colores indican a qué clúster pertenece cada uno. DBSCAN permite identificar agrupamientos de alta densidad y detectar puntos ruidosos que no pertenecen a ningún clúster. Esta visualización resalta la distribución semántica de los datos y las relaciones entre los tweets según sus representaciones semánticas generadas por Word2Vec. Elaborado por el autor.

La **Figura *20*** presenta la proyección de los datos con UMAP, utilizando la representación TF-IDF. En esta visualización, los clústeres son más dispersos y menos definidos, con una proporción significativa de puntos etiquetados como ruido. Esto pone de manifiesto las limitaciones

de TF-IDF para agrupar datos textuales de manera efectiva en estructuras complejas, incluso al usar una técnica de reducción dimensional avanzada como UMAP**.**

Figura 20

Proyección clúster DBSCAN con UMAP y TF-IDF



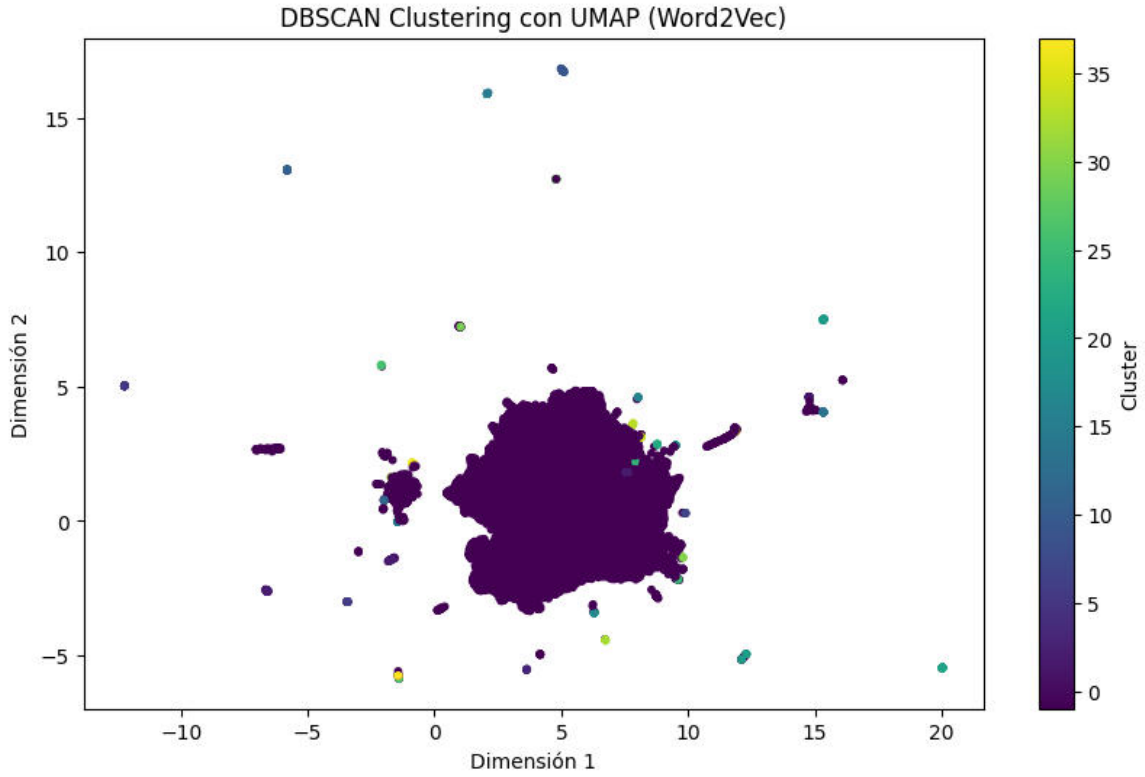
Nota: El gráfico muestra los clústeres generados por el algoritmo DBSCAN (Density-Based Spatial Clustering of Applications with Noise) aplicado a la representación TF-IDF de un conjunto de tweets sobre política española. Los clústeres fueron proyectados en un espacio de dos dimensiones utilizando UMAP (Uniform Manifold Approximation and Projection). Los puntos del gráfico representan los tweets, y los colores indican a qué clúster pertenece cada uno. DBSCAN identifica agrupamientos de alta densidad y detecta puntos ruidosos, mientras que UMAP permite una representación visual más eficiente en 2D de las relaciones semánticas entre los tweets basados en sus representaciones TF-IDF. Elaborado por el autor.

La **Figura *21*** muestra los resultados de DBSCAN aplicados a datos vectorizados con Word2Vec y proyectados mediante UMAP. Esta configuración genera clústeres notablemente más

cohesionados y definidos, con una reducción considerable de puntos clasificados como ruido. Word2Vec demuestra su eficacia al capturar relaciones semánticas complejas, mientras que UMAP conserva la estructura no lineal de los datos, maximizando la cohesión dentro de los clústeres. Esta combinación resalta subgrupos específicos y proporciona una representación clara y precisa de la organización de los datos.

Figura 21

Proyección clúster DBSCAN con UMAP y Word2Vec



Nota: El gráfico muestra los clústeres generados por el algoritmo DBSCAN (Density-Based Spatial Clustering of Applications with Noise), aplicado a la representación vectorial de palabras generada por Word2Vec, proyectados en un espacio de dos dimensiones utilizando UMAP (Uniform Manifold Approximation and Projection). Los puntos del gráfico representan los tweets sobre política española y los colores indican a qué clúster pertenece cada uno. DBSCAN identifica agrupamientos de alta densidad y detecta puntos ruidosos (outliers), mientras que UMAP facilita la visualización de las relaciones semánticas entre los tweets al

reducir la dimensionalidad de sus representaciones vectoriales obtenidas mediante Word2Vec. Elaborado por el autor.

En general, Word2Vec combinado con UMAP ofrece la representación más efectiva para DBSCAN, logrando clústeres compactos y bien definidos con menor ruido en comparación con otras configuraciones.

Análisis Léxico por Clúster DBSCAN

El análisis léxico de los clústeres generados por DBSCAN revela patrones clave en los datos textuales procesados. En total, el algoritmo identificó 106 clústeres, junto con un grupo adicional etiquetado como -1, correspondiente al ruido o datos atípicos. Este resultado ofrece información relevante sobre la organización de los datos y la eficacia del modelo en la identificación de patrones.

Cada clúster representa un conjunto de puntos densamente conectados que comparten características semánticas o contextuales similares. Los clústeres principales (0 al 106) permiten identificar temas o conceptos clave en los datos textuales. La cohesión interna y la separación entre estos grupos varía según la representación vectorial utilizada:

- Con TF-IDF
    - Los clústeres suelen contener términos relevantes, pero presentan mayor dispersión y menor cohesión semántica, como reflejan las métricas de evaluación.
    - Los clústeres formados tienden a ser más pequeños y menos definidos debido a la alta dimensionalidad y esparsidad de la representación TF-IDF.

Tabla 6
Palabras clave representativas por clúster DBSCAN (TF-IDF)

Clúster	Palabras Clave
0	pasar, zona, esfuerzo, esencial, escuela, escuchar, error, ERC
1	libro, pandemia, entrevista, entregar, equipo, zona, esfuerzo
2	zona, entrar, espacio, esfuerzo, esencial, error, escuchar

-1	gobierno, España, Sánchez, seguir, político, derecho, ley, gracias

Nota: La tabla presenta las palabras clave más representativas de cada clúster generado mediante el algoritmo DBSCAN (Density-Based Spatial Clustering of Applications with Noise), utilizando la representación TF-IDF de un conjunto de tweets sobre política española. Las palabras clave fueron seleccionadas según su valor TF-IDF, que mide la importancia de un término dentro de un clúster específico en relación con el corpus general.

El ruido incluye palabras comunes y genéricas, como "gobierno" o "Sánchez", que son relevantes, pero no forman parte de patrones densos o claramente delimitados.

- Con Word2Vec:
    - Los clústeres muestran una mayor riqueza semántica y cohesión interna, lo que sugiere que Word2Vec captura relaciones más profundas entre palabras.
    - La densidad de los puntos dentro de los clústeres es más alta, lo que facilita la identificación de temas concretos.

Tabla 7
Palabras clave representativas por clúster DBSCAN (TF-IDF)

Clúster	Palabras Clave
0	nortar, investir, coincidencia, negociar, preferente, diálogo
1	felicidad, desear, navidad, pascua, fiesta, bendiga, inmaculada
2	ánimo, cariño, admiración, afecto, sincero, querido, deseo
3	gobierno, España, Sánchez, seguir, político, derecho, ley, gracias
5	felicidades, enhorabuena, deportista, tesón, éxito, felicitación
-1	términos dispersos y genéricos, como "plan", "unión", "global"

Nota: La tabla muestra las palabras clave más representativas de cada clúster generado mediante el algoritmo DBSCAN (Density-Based Spatial Clustering of Applications with Noise), *utilizando la representación TF-IDF de un conjunto de tweets sobre política española.*

Las palabras clave corresponden a los términos con mayor peso dentro de cada clúster, identificados mediante el cálculo de la frecuencia de término y la frecuencia inversa de documento (TF-IDF). Este análisis permite comprender las temáticas predominantes en los clústeres y proporciona una visión sobre los temas tratados en los tweets, según su relevancia dentro de cada agrupamiento semántico.

El análisis léxico para Word2Vec revela que los clústeres capturan contextos específicos, como emociones ("felicidad", "admiración"), eventos ("navidad", "pascua") o temas de interacción política ("negociar", "diálogo"). Esto demuestra una mayor riqueza semántica y cohesión en comparación con TF-IDF.

Evaluación de Métricas para Agglomerative Clustering

El rendimiento del algoritmo Agglomerative Clustering fue evaluado utilizando las métricas Silhouette Score, Calinski-Harabasz Index y Davies-Bouldin Index. Los resultados obtenidos para las dos representaciones vectoriales empleadas son los siguientes:

Agglomerative Clustering (TF-IDF):

- Silhouette Score: 0.0125
- Calinski-Harabasz Index: 83.0367
- Davies-Bouldin Index: 2.2840

Estos valores indican una baja cohesión interna y una separación moderada entre los clústeres. La representación TF-IDF presenta limitaciones al generar una estructura clara en los agrupamientos debido a la alta dimensionalidad y al carácter esparso de los datos.

Agglomerative Clustering (Word2Vec):

- Silhouette Score: 0.0435
- Calinski-Harabasz Index: 1057.1847
- Davies-Bouldin Index: 3.2281

En comparación, Word2Vec ofrece una mayor cohesión interna, como lo refleja el Silhouette Score más alto. Sin embargo, el Davies-Bouldin Index más elevado sugiere que los clústeres son más dispersos y presentan una menor separación.

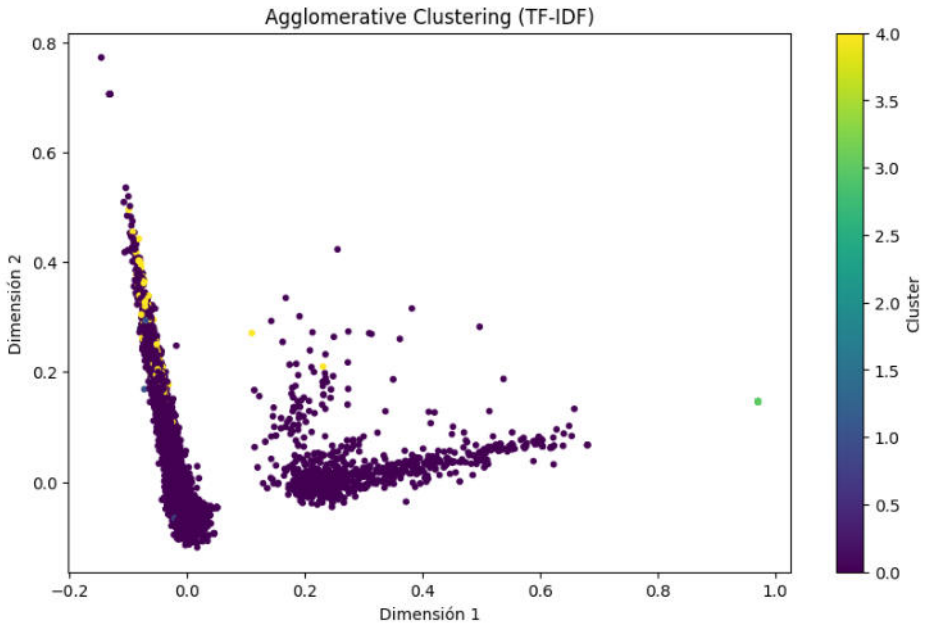
Visualización de Resultados para el Algoritmo Agglomerative Clustering

La proyección visual de los resultados del algoritmo Agglomerative Clustering fue realizada mediante PCA y UMAP, permitiendo analizar las configuraciones estructurales generadas por las representaciones TF-IDF y Word2Vec.

La **Figura *22*** muestra clústeres dispersos y una significativa proximidad entre los puntos, lo que complica la demarcación clara de los grupos. Este comportamiento puede atribuirse a la naturaleza esparsa de TF-IDF, que limita la capacidad del algoritmo para definir fronteras robustas. La cantidad de ruido en la proyección también sugiere que esta representación no es óptima para identificar estructuras cohesivas con este método.

Figura 22

Proyección de Agglomerative Clustering con TF-IDF y PCA

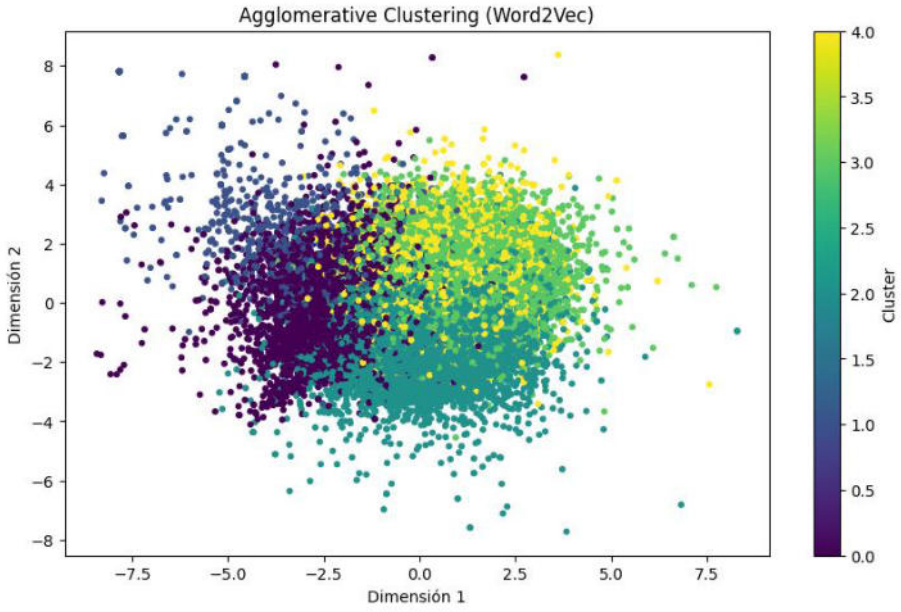


Nota: El gráfico muestra los clústeres generados por el algoritmo Agglomerative Clustering, utilizando la representación TF-IDF de un conjunto de tweets sobre política española. Los clústeres fueron proyectados en un espacio de dos dimensiones utilizando PCA (Análisis de Componentes Principales). Los puntos del gráfico representan los tweets, y los colores indican a qué clúster pertenece cada uno. Agglomerative Clustering es un método jerárquico que construye los clústeres a partir de las relaciones de proximidad entre los datos, mientras que PCA se usa para reducir la dimensionalidad y visualizar los clústeres en un espacio bidimensional. Elaborado por el autor.

La representación Word2Vec ofrece una estructura ligeramente más compacta, con clústeres más definidos en comparación con TF-IDF. Sin embargo, aún persiste el problema de solapamiento entre los puntos de diferentes clústeres, lo que indica que Agglomerative Clustering tiene dificultades para separar adecuadamente grupos en contextos semánticamente complejos.

Figura 23

Proyección de Agglomerative Clustering con Word2Vec y PCA



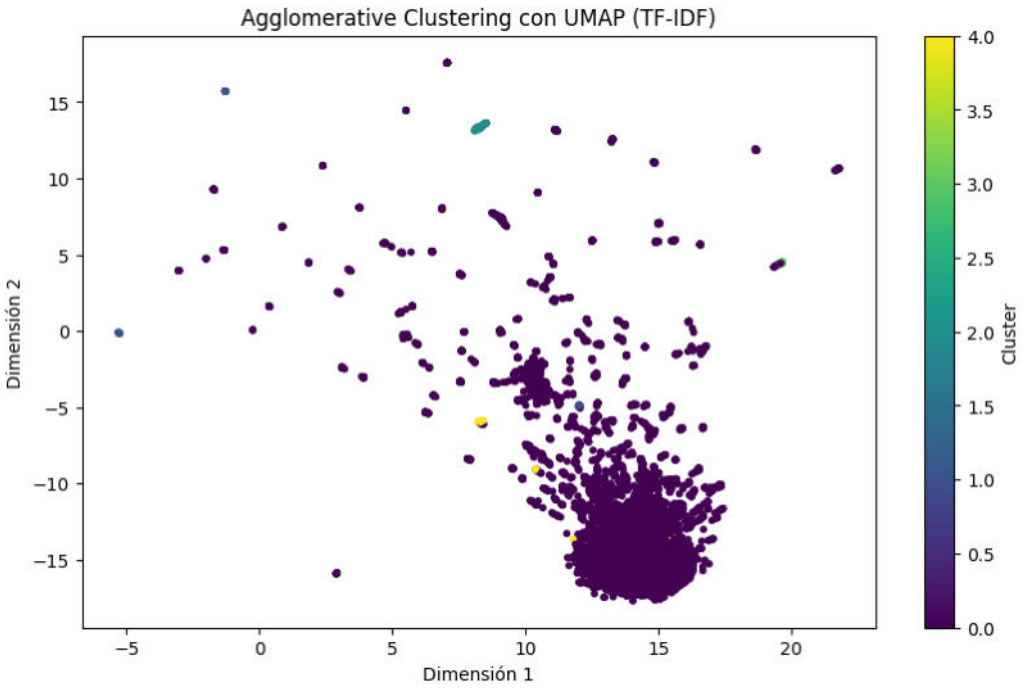
Nota: El gráfico muestra los clústeres generados por el algoritmo Agglomerative Clustering, utilizando la representación vectorial de las palabras generada por Word2Vec, proyectados en

un espacio de dos dimensiones utilizando PCA (Análisis de Componentes Principales). Los puntos en el gráfico representan los tweets sobre política española y los colores indican a qué clúster pertenece cada uno. Agglomerative Clustering es un algoritmo jerárquico que agrupa los datos en función de su similitud semántica, y PCA se utiliza para reducir la dimensionalidad de los vectores de palabras obtenidos con Word2Vec y facilitar su visualización en un espacio bidimensional. Elaborado por el autor.

La reducción dimensional con UMAP que se muestra en la **Figura *24*** evidencia una dispersión aún mayor en los clústeres generados por TF-IDF. Esto resalta la limitada capacidad de Agglomerative Clustering para manejar datos esparsos. A pesar de que UMAP preserva relaciones no lineales, los clústeres formados son débiles y muestran una considerable cantidad de ruido.

Figura 24

Proyección de Agglomerative Clustering con TF-IDF y UMA



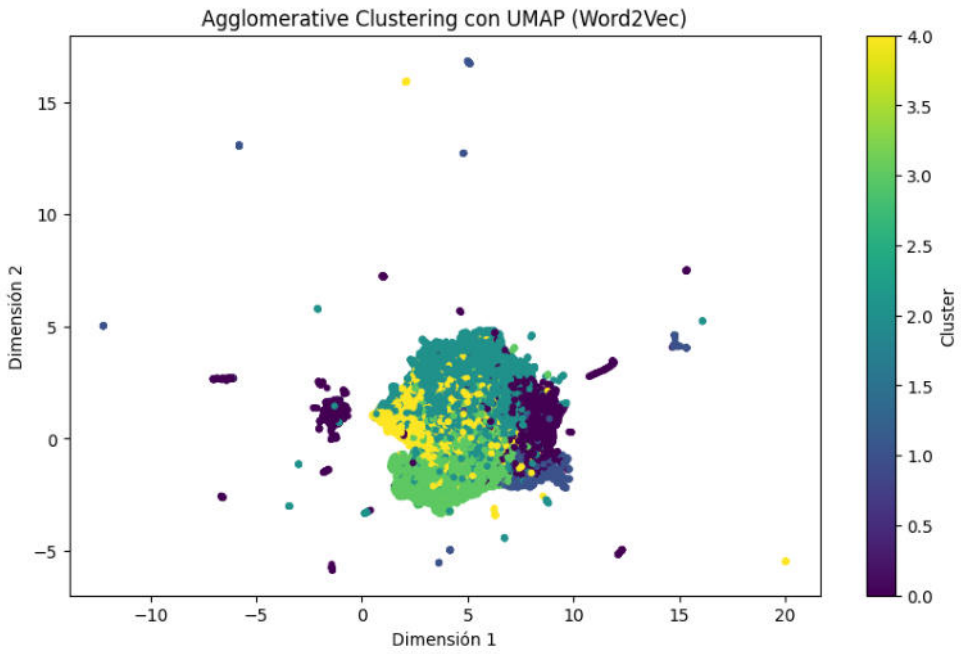
Nota: El gráfico muestra los clústeres generados por el algoritmo Agglomerative Clustering, utilizando la representación TF-IDF de un conjunto de tweets sobre política española. Los clústeres fueron proyectados en un espacio de dos dimensiones utilizando UMAP (Uniform

Manifold Approximation and Projection). Los puntos del gráfico representan los tweets, y los colores indican a qué clúster pertenece cada uno. Agglomerative Clustering es un algoritmo jerárquico que agrupa los datos en función de la similitud semántica, y UMAP se utiliza para reducir la dimensionalidad de los vectores TF-IDF y facilitar su visualización en un espacio bidimensional, manteniendo las relaciones estructurales y semánticas de los datos. Elaborado por el autor.

La combinación de Word2Vec y UMAP refuerza la cohesión interna de los clústeres. Los datos dentro de cada grupo están más densamente agrupados, y el solapamiento es menor en comparación con las demás configuraciones. Esto es evidenciado en la **Figura *25*** la capacidad de Word2Vec para capturar relaciones semánticas y la eficacia de UMAP para mejorar la visualización de las estructuras latentes.

Figura 25

Proyección de Agglomerative Clustering con Word2Vec y UMAP



Nota: El gráfico muestra los clústeres generados por el algoritmo Agglomerative Clustering, utilizando la representación TF-IDF de un conjunto de tweets sobre política española. Los clústeres fueron proyectados en un espacio de dos dimensiones utilizando UMAP (Uniform

Manifold Approximation and Projection). Los puntos del gráfico representan los tweets, y los colores indican a qué clúster pertenece cada uno. Agglomerative Clustering es un algoritmo jerárquico que agrupa los datos en función de la similitud semántica, y UMAP se utiliza para reducir la dimensionalidad de los vectores TF-IDF y facilitar su visualización en un espacio bidimensional, manteniendo las relaciones estructurales y semánticas de los datos. Elaborado por el autor.

En la **Figura *22*,** los clústeres generados con TF-IDF y PCA son dispersos, dificultando su separación. La **Figura *23*,** con Word2Vec y PCA, muestra una mejora en la compactación de los clústeres, aunque persiste el solapamiento. La **Figura *24***, que combina TF-IDF y UMAP, evidencia una mayor dispersión, reflejando limitaciones en la cohesión de los clústeres. Por último, la **Figura *25***, con Word2Vec y UMAP, presenta clústeres más definidos y cohesionados, resaltando la efectividad de esta configuración.

En general, Word2Vec y UMAP ofrecen mejores resultados en términos de definición y cohesión de clústeres, destacando la importancia de elegir representaciones y técnicas adecuadas para optimizar el clustering.

Análisis Léxico por Clúster para Agglomerative Clustering

En este análisis se evaluaron los clústeres generados mediante Agglomerative Clustering con representaciones vectoriales TF-IDF y Word2Vec, explorando los términos más representativos para identificar patrones temáticos clave. Las tablas a continuación presentan las palabras clave asociadas a cada clúster, destacando aquellos que presentan temáticas definidas y aquellos con menor cohesión semántica. Este análisis permite identificar los clústeres más relevantes y sugiere áreas de mejora para optimizar los resultados.

Tabla 8
Palabras clave representativas por clúster agglomerative clustering (TF-IDF)

Clúster	Palabras Clave
0	gobierno, españa, seguir, sanchez, español, político, madrid, ley, derecho, gracias
1	agenda, dejar, sanchez, ir, ayuda, español, libertad, gente, precio, gobierno
2	the, to, and, in, of, global, president, plan, union, debate
3	gracias, zona, entender, espacio, esfuerzo, esencial, escuela, escuchar, escribir, error
4	pedro, pedro sanchez, sanchez, gobierno, español, congreso, presidente, españa, indulto, socio

Nota: La tabla muestra las palabras clave más representativas de cada clúster generado mediante el algoritmo Agglomerative Clustering, utilizando la representación TF-IDF de un conjunto de tweets sobre política española. Las palabras clave fueron seleccionadas según su valor TF-IDF.

Entre los hallazgos más relevantes, el Clúster 0 destaca por su consistencia en temas políticos y legales, indicando una fuerte cohesión en torno a términos como gobierno, sanchez y ley. De manera similar, el Clúster 4 presenta una temática bien definida alrededor de figuras públicas y eventos políticos clave, lo que refuerza su relevancia. Por el contrario, el Clúster 2 es menos cohesivo debido a la presencia de términos genéricos en inglés que generan ruido.

Tabla 9
Palabras clave representativas por clúster agglomerative clustering (Word2Vec)

Clúster	Palabras Clave
0	arranca, arrancado, guadairar, contaro, preciosa, conecto, esperamo, armilla, xv, maravillar
1	gracia, gracias, dedicación, tesón, agradecido, felicidad, gracios, graciar, suerte, corazón
2	tolera, confrontar, concernido, sentar, mercadear, acaben, sistemáticamente, apoye, únicamente, reconsiderar
3	aumentar, funcionado, comprometer, certidumbre, ágil, adecuado, beneficioso, reducir, prioridad, traducir
4	trabajaremos, garantiar, necesitamos, país, plenamente, firmemente, paislo, determinación, obligación, fortalecer

Nota: La tabla muestra las palabras clave más representativas de cada clúster generado mediante el algoritmo Agglomerative Clustering, utilizando la representación Word2Vec de un conjunto de tweets sobre política española. Las palabras clave fueron seleccionadas en función de su cercanía semántica dentro de cada clúster, medida a través de los vectores de palabras generados por Word2Vec, que captura las relaciones semánticas entre los términos.

En esta representación, el Clúster 1 es particularmente fuerte, con un enfoque en mensajes de agradecimiento y reconocimiento que destacan por su alta cohesión semántica. Asimismo, el Clúster 4 refleja un mensaje motivacional y colaborativo, capturando objetivos colectivos claros. Sin embargo, el Clúster 2 presenta términos relacionados con dilemas y conflictos, pero su cohesión es limitada en comparación con otros clústeres.

El análisis léxico revela que Word2Vec genera clústeres más ricos semánticamente en comparación con TF-IDF, destacándose los Clústeres 0 y 4 en TF-IDF por su relevancia política y legal, y los Clústeres 1 y 4 en Word2Vec por su temática emocional y motivacional. Estos hallazgos subrayan la importancia de elegir la representación vectorial adecuada para maximizar la calidad de los clústeres, especialmente en contextos donde la semántica y la cohesión son fundamentales para el análisis de datos textuales.

Análisis General de los Resultados

Análisis Comparativo del Rendimiento de los Algoritmos

Tabla 10

Rendimiento de los Algoritmos de Clustering con Diferentes Representaciones Vectoriales

Algoritmo	Representación	Silhouette Score	Calinski-Harabasz Index	Davies-Bouldin Index
K-means	TF-IDF	0.0161	102.3372	7.3943
K-means	Word2Vec	0.0711	1501.8709	2.7189

DBSCAN	TF-IDF	-0.1879	12.9216	1.2266
DBSCAN	Word2Vec	-0.2283	53.3354	0.6386
Agglomerative	TF-IDF	0.0125	83.0367	2.2840
Agglomerative	Word2Vec	0.0435	1057.1847	3.2281

Nota: La tabla muestra el rendimiento comparativo de varios algoritmos de clustering (K-Means, DBSCAN, Agglomerative Clustering) al aplicar diferentes representaciones vectoriales de texto: TF-IDF, Word2Vec..Los resultados se evaluaron utilizando métricas de Silhouette Score, Davies-Bouldin Index y Calinski-Harabasz Index, las cuales proporcionan una medida de la calidad y cohesión de los clústeres generados.

Comparación entre Representaciones Vectoriales

TF-IDF. La representación TF-IDF demostró ser menos efectiva debido a su naturaleza esparsa y su limitada capacidad para capturar relaciones semánticas complejas, lo que resultó en un bajo rendimiento en todos los algoritmos.

Word2Vec. Word2Vec sobresale en la captura de relaciones semánticas y contextuales, mejorando significativamente la cohesión interna y la separación entre clústeres, especialmente cuando se combina con KMeans.

Discusión final

El análisis general señala que KMeans con Word2Vec es la solución más robusta para identificar trending topics relevantes en datos textuales. Esta combinación permite formar clústeres bien definidos y minimiza el solapamiento entre grupos, lo cual es crucial para estrategias de marketing político dirigidas.

DBSCAN mostró potencial para manejar datos con densidades heterogéneas, pero su desempeño en cohesión interna fue limitado. Por su parte, Agglomerative Clustering logró resultados

razonables, aunque no alcanzó el nivel de claridad y cohesión demostrado por KMeans.

Para la identificación de tendencias temáticas y la optimización de campañas de marketing político, KMeans con Word2Vec se posiciona como el algoritmo más efectivo. Este enfoque combina cohesión interna, separación entre clústeres y una representación semántica sólida. En investigaciones futuras, sería valioso explorar combinaciones más avanzadas, como Word2Vec con técnicas jerárquicas basadas en densidad, para abordar escenarios más complejos y heterogéneos.

Capítulo 4. Conclusiones

El análisis de clustering aplicado a datos textuales con KMeans, DBSCAN y Agglomerative Clustering permitió evaluar representaciones vectoriales como TF-IDF y Word2Vec para identificar tendencias temáticas. KMeans combinado con Word2Vec mostró un desempeño superior, logrando clústeres bien definidos con un Silhouette Score de 0.0711, un Calinski-Harabasz Index de 1501.8709 y un Davies-Bouldin Index de 2.7189. DBSCAN, aunque efectivo en identificar estructuras densas con Word2Vec, presentó problemas de cohesión interna reflejados en un Silhouette Score negativo (-0.2283). Por su parte, Agglomerative Clustering mostró resultados intermedios, destacándose levemente con Word2Vec.

El proceso enfrentó limitaciones, como el alto consumo de memoria RAM al manejar representaciones de alta dimensionalidad como TF-IDF, lo que restringió el procesamiento de corpus más grandes. Además, la naturaleza esparsa de TF-IDF afectó la cohesión interna de los clústeres en escenarios con relaciones semánticas complejas.

Para superar estas limitaciones, se proponen acciones como una limpieza más profunda del corpus mediante técnicas avanzadas de preprocesamiento y reducción de ruido, optimización de recursos computacionales con infraestructuras de mayor capacidad y el uso de algoritmos complementarios como HDBSCAN o Gaussian Mixture Models. También se sugiere explorar técnicas de aprendizaje profundo para capturar relaciones más detalladas y validar resultados mediante configuraciones iterativas.

En conclusión, aunque KMeans con Word2Vec fue la solución más robusta, la implementación de estrategias de mejora y la incorporación de técnicas avanzadas pueden ampliar significativamente el alcance del análisis, permitiendo abordar problemas más complejos y enriquecer futuras aplicaciones.

Referencias

Ahmad, I. (2024). *50 algoritmos que todo programador debe conocer* (2a ed.). Marcombo.

Almeida, F., & Xexéo, G. (2019). *Word Embeddings: A Survey*. http://arxiv.org/abs/1901.09069

Alsina, Á. (2024). Modelo de Transformación Docente a través de Redes Sociales: ejemplificación en X para la educación matemática en la era digital. *Revista Digital: Matemática, Educación e Internet*, *25*(1). https://doi.org/10.18845/meij.v25i1.7236

Amin Baybon. (2023, julio 22). *Tokenization using NLTK (TweetTokenizer)*. Medium. https://medium.com/@aminbaybon/tokenization-using-nltk-tweettokenizer-d1213c1412d9

Arcila-Calderón, C., Ortega-Mohedano, F., Jiménez-Amores, J., & Trullenque, S. (2017). Análisis supervisado de sentimientos políticos en español: clasificación en tiempo real de tweets basada en aprendizaje automático. *El Profesional de la Información*, *26*(5), 973. https://doi.org/10.3145/epi.2017.sep.18

Avila Perez-Grovas, P. (2023). *Análisis de Patrones y Predicción de Comportamiento de Usuarios a través de Técnicas de Agrupamiento* [Tesis de grado, EscuelaTécnica Superior de Ingeniería (ICAI)]. http://hdl.handle.net/11531/79716

Bimber, B. (2014). Digital Media in the Obama Campaigns of 2008 and 2012: Adaptation to the Personalized Political Communication Environment. *Journal of Information Technology & Politics*, *11*(2), 130–150. https://doi.org/10.1080/19331681.2014.895691

Campos-Domínguez, E. (2017). Twitter and political communication. *Profesional de la Informacion*, *26*(5), 785–793. https://doi.org/10.3145/epi.2017.sep.01

Campos-Domínguez, E., & Calvo, D. (2017). La campaña electoral en Internet: planificación, repercusión y viralización en Twitter durante las elecciones españolas de 2015. *Comunicación y sociedad*, *29*, 93–116.

Criado, J. I., & Villodre, J. (2018). Local public sector big data communication on social media. A sentiment analysis in Twitter. *Profesional de la Informacion*, *27*(3), 614–623. https://doi.org/10.3145/epi.2018.may.14

Deng, D. (2020). DBSCAN Clustering Algorithm Based on Density. *2020 7th International Forum on Electrical Engineering and Automation (IFEEA)*, 949–953. https://doi.org/10.1109/IFEEA51475.2020.00199

Esteban, A., Zafra, A., & Ventura, S. (2021). Estudio comparativo de medidas de disimilitud para Clustering Multi-Instancia. En E. Alba (Ed.), *Actas de la XIX Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA 20/21).* (pp. 673–679). CAEPIA. https://caepia20-21.uma.es/inicio.html

García Diéguez, A. (2014). *Técnicas de clustering aplicadas al análisis de trending topics en conjunto de tweets* [Tesis de grado, Universidad Carlos III de Madrid]. https://hdl.handle.net/10016/22233

Godoy Viera, Á. F. (2017). Técnicas de aprendizaje de máquina utilizadas para la minería de texto. *Investigación Bibliotecológica. Archivonomía, Bibliotecología e Información*, *31*(71), 103. https://doi.org/10.22201/iibi.0187358xp.2017.71.57812

Graham, T., Jackson, D., & Broersma, M. (2016). New platform, old habits? Candidates' use of Twitter during the 2010 British and Dutch general election campaigns. *New Media & Society*, *18*(5), 765–783. https://doi.org/10.1177/1461444814546728

Gupta, I., & Joshi, N. (2017, diciembre 18). Tweet Normalization: A Knowledge Based Approach. *2017 International Conference on Infocom Technologies and Unmanned Systems (ICTUS'2017)*.

Jia, Q., & Xu, S. (2022). An Overall Analysis of Twitter and Elon Musk M&A Deal. En Gerald Bartlett & Jie Zhang (Eds.), *Highlights in Business, Economics and Management EMFT* (Vol. 2022, pp. 436–441). Darcy & Roy Press. https://doi.org/10.54097/hbem.v2i.2189

Jungherr, A. (2014). The Logic of Political Coverage on Twitter: Temporal Dynamics and Content. *Journal of Communication*, *64*(2), 239–259. https://doi.org/10.1111/jcom.12087

Jürgens, P., & Jungherr, A. (2015). The Use of Twitter during the 2009 German National Election. *German Politics*, *24*(4), 469–490. https://doi.org/10.1080/09644008.2015.1116522

Mahdiraji, H. A., Kazimieras Zavadskas, E., Kazeminia, A., & Abbasi Kamardi, A. (2019). Marketing strategies evaluation based on big data analysis: a CLUSTERING-MCDM approach. *Economic Research-Ekonomska Istraživanja*, *32*(1), 2882–2898. https://doi.org/10.1080/1331677X.2019.1658534

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. http://arxiv.org/abs/1301.3781

Mishra, N., Schreiber, R., Stanton, I., & Tarjan, R. E. (2007). Clustering Social Networks. En Heidelberg: Springer Berlin Heidelberg. (Ed.), *International Workshop on Algorithms and Models for the Web-Graph* (pp. 56–67).

NLTK Project. (2024, agosto 19). *NLTK Documentation*. NLTK. https://www.nltk.org/api/nltk.tokenize.casual.html#module-nltk.tokenize.casual

OpenAI. (2024). *A visually engaging image illustrating social media clustering, specifically on Twitter.The image shows interconnected groups or clusters of users [Imagen generada por DALL-E].* https://openai.com/dall-e/

Peng, H. (2008). Bioimage informatics: a new area of engineering biology. *Bioinformatics*, *24*(17), 1827–1836. https://doi.org/10.1093/bioinformatics/btn346

Piñeiro-Otero, T. (2023). De Twitter a X. El riesgo de la dependencia de fuentes privadas para la investigación. *Anuario ThinkEPI*, *17*. https://doi.org/10.3145/thinkepi.2023.e17a46

Quintana, F. (2003). Cluster analysis of human autoantibody reactivities in health and in type 1 diabetes mellitus: a bio-informatic approach to immune complexity. *Journal of Autoimmunity*, *21*(1), 65–75. https://doi.org/10.1016/S0896-8411(03)00064-7

Quintana, F., Getz, G., Hed, G., Domany, E., & Cohen, I. (2003). Cluster analysis of human autoantibody reactivities in health and in type 1 diabetes mellitus: a bio-informatic approach to immune complexity. *Journal of Autoimmunity*, *21*(1), 65–75. https://doi.org/10.1016/S0896-8411(03)00064-7

Rehman, S. U., Asghar, S., Fong, S., & Sarasvady, S. (2014). DBSCAN: Past, present and future. *The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)*, 232–238. https://doi.org/10.1109/ICADIWT.2014.6814687

Ricardo Moya. (2021). *Tweets Política España*. Kaggle. https://www.kaggle.com/datasets/ricardomoya/tweets-poltica-espaa

Ros, F., Riad, R., & Guillaume, S. (2023). PDBI: A partitioning Davies-Bouldin index for clustering evaluation. *Neurocomputing*, *528*, 178–199. https://doi.org/10.1016/j.neucom.2023.01.043

Saura García, C. (2023). El big data en los procesos políticos: hacia una democracia de la vigilancia. *Revista de filosofía*, *80*, 215–232. https://doi.org/10.4067/S0718-43602023000100215

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press. http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning

Soria-Olivas, E., Sánchez-Montañes Isla, M. A., Gamero Cruz, R., Castillo Caballer, B., & Cano Michelena, P. (2023). *Sistemas de Aprendizaje Automático*. Ra-Ma Editorial.

Strušnik, D., Golob, M., & Avsec, J. (2015). Artificial neural networking model for the prediction of high efficiency boiler steam generation and distribution. *Simulation Modelling Practice and Theory*, *57*, 58–70. https://doi.org/10.1016/j.simpat.2015.06.003

Sundqvist, M., Chiquet, J., & Rigaill, G. (2023). Adjusting the adjusted Rand Index. *Computational Statistics*, *38*(1), 327–347. https://doi.org/10.1007/s00180-022-01230-7

Suresh, P., Shettigar, A., Karunavathi, M., Ajith, & Ramanath Kini, M. G. (2023). Term Frequency Tokenization for Fake News Detection. En J. Hemanth, D. Pelusi, & J. I. Z. Chen (Eds.), *Intelligent Cyber Physical Systems and Internet of Things. ICoICI 2022* (Vol. 3, pp. 1–10). Springer, Cham. https://doi.org/10.1007/978-3-031-18497-0_1

Tari, L., Baral, C., & Kim, S. (2009). Fuzzy c-means clustering with prior biological knowledge. *Journal of Biomedical Informatics*, *42*(1), 74–81. https://doi.org/10.1016/j.jbi.2008.05.009

Velásquez-Gushiken, A. (2023). *Análisis de los métodos de recolección de textos sarcásticos* [Tesis de grado, Universidad de Ingeniería y Tecnología]. https://hdl.handle.net/20.500.12815/307

Vergeer, M., & Hermans, L. (2013). Campaigning on Twitter: Microblogging and Online Social Networking as Campaign Tools in the 2010 General Elections in the Netherlands. *Journal of Computer-Mediated Communication*, *18*(4), 399–419. https://doi.org/10.1111/jcc4.12023

Yule, G. (2004). *El lenguaje.* (3a ed., Vol. 1). Ediciones AKAL.

Apéndices

Apendice A: Dependencias y Configuración del Entorno

Para la correcta ejecución de este estudio y la implementación de los algoritmos y análisis descritos, es fundamental contar con las librerías y versiones específicas enumeradas en este apéndice. Estas dependencias garantizan la compatibilidad del entorno y aseguran el correcto funcionamiento de las herramientas empleadas, especialmente en el procesamiento y análisis del corpus textual.

Entorno de Ejecución

- Versión de Python: 3.10.11

Librerías Utilizadas

A continuación, se detallan todas las librerías y versiones empleadas en este estudio:

Procesamiento del Lenguaje Natural (PLN)

- spacy==3.8.3
- es_core_news_sm==3.8.0
- nltk==3.9.1
- gensim==4.3.3

Vectorización y Modelado

- scikit-learn==1.6.0
- wordcloud==1.9.4

Manipulación y Procesamiento de Datos

- numpy==1.26.4
- pandas==2.2.3

- scipy==1.13.1
- joblib==1.4.2

Visualización

- matplotlib==3.10.0
- seaborn==0.13.2
- umap-learn==0.5.7

Otros Requerimientos

- tqdm==4.67.1
- requests==2.32.3
- smart-open==7.1.0

Cada librería fue seleccionada por su funcionalidad específica, desde la manipulación de datos y el procesamiento del lenguaje natural hasta la generación de visualizaciones y el análisis de clusters.