

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
SANTIAGO - CHILE



“CARACTERIZACIÓN DE HASHTAGS PARA LA ELABORACIÓN DE UNA TIPOLOGÍA”

NAZARETH LOURDES DÍAZ TAPIA

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL EN INFORMÁTICA

Profesor Guía: Dr. Andrés Moreira
Profesor Correferente: Roberto León

Septiembre - 2023

DEDICATORIA

A mamá y papá, mi eterna gratitud por su apoyo y amor incondicional.

AGRADECIMIENTOS

En primer lugar, agradecer a mi núcleo familiar, mamá, papá y hermana, quienes han sido mi mayor apoyo y fuente de amor incondicional. Siempre con palabras de aliento y comprensión. Agradezco infinitamente el aceptar la llegada de mi perro Draco, lo máspreciado en mi vida.

A mi familia materna más cercana, mi Ita, tía Maca y tío Nene, les agradezco por confiar en mis capacidades. En especial, recordar a mi abuelo José, quien siempre está presente en mis pensamientos.

A mi pareja, Mauricio, le agradezco su apoyo a lo largo de estos años, las palabras, los consejos y las sonrisas. También agradecer a su familia, Loreto, María Dolores y Eliseo. Elito, por siempre voy a recordar con mucho cariño tus consejos y tus historias de vida única, muchas gracias.

Mis amigas de la universidad, mis niñas, que siempre han estado en momentos de felicidad, tristeza y estrés, sin ustedes la universidad no habría sido lo mismo. También agradezco a mi grupo de amigos, algunos desde el primer día y otros que se unieron en el camino, por hacer mi experiencia universitaria más entretenida y llevadera.

A los increíbles profesionales que conocí en la universidad, en especial a los profesores del departamento de informática y mención especial al profesor Andrés Moreira, mi profesor guía de esta gran investigación. Por último, agradezco a los profesionales del área que conocí en mis inicios de la vida laboral, quienes me enseñaron mucho y sin saberlo, reavivaron mi pasión por la informática.

RESUMEN

Resumen— Este estudio se centra en la problemática del uso de hashtags en la red social Twitter, donde la creación desordenada de etiquetas similares y la evolución constante de su significado dificultan su interpretación. Para abordar este problema, se propone una investigación exhaustiva que busca desarrollar una tipología de hashtags y comprender su comportamiento en profundidad. La metodología se basa en la extracción y análisis de atributos relacionados con los hashtags, como su frecuencia de uso, estabilidad en el tiempo y co-ocurrencia con otros hashtags. El estudio se apoya en un conjunto de datos recopilados de Twitter y utiliza un algoritmo de clustering para categorizar los hashtags en grupos.

Palabras Clave— Twitter; Hashtags; Clustering; Atributos; Tipología

ABSTRACT

Abstract— This study focuses on the hashtag usage on the social media platform Twitter, where the chaotic creation of similar tags and the constant evolution of their meanings make their interpretation a challenge. To address this problem, an in-depth investigation is proposed with the aim of developing a typology of hashtags and gaining a deeper understanding of their behavior. The methodology relies on the extraction and analysis of characteristics related to hashtags, such as their frequency of use, stability over time, and co-occurrence with other hashtags. The study is supported by a dataset collected from Twitter and uses a clustering algorithm to categorize hashtags groups.

Keywords— Twitter; Hashtags; Clustering; Characteristics; Typology

GLOSARIO

Cronología de inicio o *timeline*: Lugar donde se encuentran los *tweets* y temas que sigue el usuario en orden cronológico.

Hashtag: Son palabras que son escritas con el signo “#” antepuesto y se utiliza para indexar palabras claves o temas en Twitter.

Microblogging: Forma de comunicarse a través de mensajes cortos que pueden ser enviados en una red social o sitios web.

Retweet o RT: Actualmente llamado *Repost*. Forma de compartir un *tweet* ya publicado.

Tweet: Actualmente llamado *Post*. Publicación de máximo 280 caracteres que puede contener texto, imágenes, videos, enlaces y son hechos por los usuarios de Twitter.

Twitter: Red social donde los usuarios comparten sus pensamientos o difunden información a través de los *tweets*. Actualmente llamada X.

Seguidor o *follower*: Cuenta suscrita a los contenidos del usuario.

Seguido: Cuenta a la que un usuario se suscribe

URL: Identificador único que se le asigna a cada recurso disponible en la web.

ÍNDICE DE CONTENIDOS

RESUMEN	IV
ABSTRACT	IV
GLOSARIO	V
ÍNDICE DE FIGURAS	VIII
ÍNDICE DE TABLAS	IX
INTRODUCCIÓN	1
CAPÍTULO 1: DEFINICIÓN DEL PROBLEMA	3
1.1 Propuesta de solución	4
1.2 Objetivos	5
1.2.1 Objetivo general	5
1.2.2 Objetivos específicos	5
1.3 Trabajos relacionados	6
CAPÍTULO 2: MARCO CONCEPTUAL	9
2.1 Twitter	9
2.2 <i>Hashtags</i>	10
2.2.1 Estabilidad en el tiempo	11
2.2.2 Convergencia	11
2.2.3 Co-ocurrencia	12
2.2.4 Frecuencia de uso	12
2.3 ETL	12
2.4 Reducción de dimensionalidad	13
2.4.1 Análisis de componentes principales (PCA)	13
2.4.2 Uniform Manifold Approximation and Projection (UMAP)	15
2.5 Algoritmos de agrupación o <i>clustering</i>	16
2.5.1 Algoritmo <i>k-means</i>	17
CAPÍTULO 3: PROPUESTA E IMPLEMENTACIÓN DE LA SOLUCIÓN	20
3.1 Descripción de los datos	20
3.2 Procesamiento de los datos	23
3.3 Características de <i>hashtags</i>	25
3.3.1 Frecuencia	25
3.3.2 Co-ocurrencia	26
3.3.3 Largo cadena de texto	27
3.3.4 Temporalidad	27
3.3.5 Cantidad de países	28
3.3.6 Dígitos	28

3.3.7	Derivados de frecuencia y tiempo	29
3.3.8	Vectorización de palabras	30
3.4	Implementación de algoritmo	31
3.4.1	Muestra 1	33
3.4.2	Muestra 2	69
CAPÍTULO 4: ANÁLISIS DE RESULTADOS Y DISCUSIÓN		98
CAPÍTULO 5: CONCLUSIONES		104
ANEXOS		108
5.1	Características que aportan al PCA, muestra 1	108
5.2	Características que aportan al PCA, muestra 2	111
REFERENCIAS BIBLIOGRÁFICAS		114

ÍNDICE DE FIGURAS

1	Conexiones en Twitter	10
2	Representación simplex	15
3	Representación de aplicar un algoritmo de clustering	17
4	Pseudocódigo de algoritmo <i>k-means</i>	18
5	Ejemplo de tweet	20
6	Representación componente 1 y componente 2	34
7	Método del codo	35
8	Índice de Davies-Bouldin	35
9	Índice de Calinski-Harabasz	36
10	Estadístico de gap	36
11	Resultado <i>k-means</i> con $k=17$, reducción de dimensiones por PCA	37
12	Dimensiones a partir de UMAP	49
13	Método del codo	49
14	Índice de Davies-Bouldin	50
15	Índice de Calinski-Harabasz	50
16	Estadístico de gap	51
17	Resultado <i>k-means</i> con $k=7$, reducción de dimensiones por UMAP	51
18	Aportes de cada atributo en el componente 1	57
19	Aportes de cada atributo en el componente 2	58
20	Aportes de cada atributo en el componente 3	58
21	Conjunto de dimensiones reducidas a partir de PCA y UMAP	59
22	Método del codo	60
23	Índice de Davies-Bouldin	60
24	Índice de Calinski-Harabasz	61
25	Estadístico de gap	61
26	Resultado <i>k-means</i> con $k=9$, reducción de dimensiones por PCA y UMAP	62
27	Representación de los primeros dos componentes	70
28	Método del codo	70
29	Índice de Davies-Bouldin	71
30	Índice de Calinski-Harabasz	71
31	Estadístico de GAP	72
32	Resultado <i>k-means</i> con $k=9$, reducción de dimensiones por PCA	72
33	Dimensiones a partir de UMAP	79
34	Método del codo	80
35	Índice de Davies-Bouldin	80
36	Índice de Calinski-Harabasz	81
37	Estadístico de gap	81
38	Resultado <i>k-means</i> con $k=7$, reducción de dimensiones por UMAP	82
39	Aportes de cada atributo en el componente 1	88
40	Aportes de cada atributo en el componente 2	89
41	Aportes de cada atributo en el componente 3	90
42	Resultado <i>k-means</i> con $k=7$, reducción de dimensiones por UMAP	91

43	Método del codo	91
44	Índice de Davies-Bouldien	92
45	Índice de Calinski-Harabasz	92
46	Estadístico de GAP	93
47	Resultado <i>k-means</i> con $k=6$, reducción de dimensiones por PCA y UMAP . . .	93
48	Aportes de cada atributo en el componente 4	108
49	Aportes de cada atributo en el componente 5	108
50	Aportes de cada atributo en el componente 6	109
51	Aportes de cada atributo en el componente 7	109
52	Aportes de cada atributo en el componente 8	110
53	Aportes de cada atributo en el componente 9	110
54	Aportes de cada atributo en el componente 4	111
55	Aportes de cada atributo en el componente 5	111
56	Aportes de cada atributo en el componente 6	112
57	Aportes de cada atributo en el componente 7	112
58	Aportes de cada atributo en el componente 8	113
59	Aportes de cada atributo en el componente 9	113

ÍNDICE DE TABLAS

1	Atributos de un tweet.	21
2	Atributos de un tweet.	22
3	Twitter Place.	23
4	10 <i>hashtags</i> con mayor frecuencia	26
5	Palabras similares para covid19	31
6	Resultados cantidad de clusters	98
7	Resultados para muestra 1	102
8	Resultados para muestra 2	102

INTRODUCCIÓN

Las redes sociales han revolucionado la forma en que nos comunicamos y compartimos información en la era digital, y entre ellas, Twitter destaca como una de las plataformas más ampliamente utilizadas en todo el mundo. Con aproximadamente 450 millones de usuarios activos al mes en el año 2022, Twitter se ha convertido en un espacio fundamental para la interacción y difusión de ideas a través de mensajes breves conocidos como tweets. A lo largo de su evolución, Twitter ha ampliado la extensión de estos tweets, pasando de una limitación inicial de 140 caracteres a la actual posibilidad de extenderse hasta 280 caracteres.

Además de su característica distintiva de mensajes cortos, Twitter ha introducido una variedad de elementos multimedia, como imágenes, videos, gifs y enlaces, que enriquecen la experiencia de los usuarios. Sin embargo, uno de los componentes más icónicos y duraderos de esta plataforma son los *hashtags*.

Los *hashtags*, representados por el símbolo numeral o hash (#), permiten etiquetar y clasificar el contenido de los tweets, facilitando su organización y búsqueda. Su éxito en Twitter ha llevado a su adopción en otras redes sociales, convirtiéndolos en herramientas esenciales para la categorización de temas y conversaciones en línea.

No obstante, en este contexto de apertura y descentralización, ha surgido un problema importante relacionado con el uso de los *hashtags*. La creación desenfrenada de etiquetas ha generado una diversidad de *hashtags* similares, pero carentes de una estructura semántica clara. Además, la evolución constante de los *hashtags* en términos de popularidad y significado a lo largo del tiempo dificulta su interpretación precisa.

Este problema se manifiesta en varios aspectos, como la polisemia de los *hashtags*, la presencia de sinónimos y la falta de coherencia en su clasificación y etiquetado debido a la ambigüedad que pueden presentar. Los *hashtags* pueden tener éxito o fracasar de manera rápida y su escritura variada complica su consistencia y comprensión.

Para abordar esta problemática, se propone llevar a cabo una investigación exhaustiva de los *hashtags* en Twitter. El objetivo principal de este estudio es desarrollar una tipología que permita agrupar estos elementos en categorías significativas y comprender su comportamiento en profundidad.

Dado que los *hashtags* presentan características únicas y una jerga específica de Twitter, se adoptará un enfoque innovador que aproveche estas particularidades para desentrañar su significado y clasificarlos en grupos semánticos interconectados. La metodología propuesta se basa en la extracción y análisis de atributos relacionados con los *hashtags*, incluyendo su frecuencia de uso, estabilidad en el tiempo y co-ocurrencia con otros *hashtags*.

Este estudio se apoyará en un conjunto de datos recopilados de Twitter, abarcando desde el año 2019 hasta el 2020, con el fin de analizar la evolución temporal de los *hashtags*. Se

utilizará un algoritmo de clustering para categorizar los *hashtags* en función de los atributos identificados, lo que permitirá identificar patrones y comportamientos de estos elementos.

En el Capítulo 1, se definirá en detalle el problema, se propondrá una estrategia para resolverlo y se establecerán los objetivos de esta investigación. También se revisarán trabajos relacionados que han abordado casos similares en el pasado.

El Capítulo 2 se centrará en el marco teórico, donde se definirán conceptos clave relacionados con Twitter y se introducirán los algoritmos de clustering que se utilizarán en este estudio.

El Capítulo 3 detallará la propuesta de solución, describiendo cómo se procesarán los datos recopilados y cómo se aplicarán los algoritmos mencionados.

En el Capítulo 4, se presentará un análisis exhaustivo de los datos, acompañado de una discusión que mostrará los resultados obtenidos, algunas validaciones y las conclusiones derivadas de los hallazgos.

Finalmente, en el Capítulo 5, se resumirán las conclusiones generales de esta investigación y se planteará las direcciones futuras de trabajo en este campo en constante evolución.

CAPÍTULO 1

DEFINICIÓN DEL PROBLEMA

Twitter es una de las redes sociales más ampliamente utilizadas en el mundo, con aproximadamente 450 millones de usuarios activos al mes en el año 2022. Esta plataforma permite a los usuarios comunicarse a través de mensajes breves denominados tweets, los cuales, en un inicio estaban limitados a 140 caracteres y en la actualidad puede extenderse hasta 280 caracteres. Los tweets pueden dirigirse directamente a otros usuarios o ser publicados en la cronología de inicio o timeline para ser compartidos con la red de seguidores y seguidos en general.

Estos tweets pueden contener una gran variedad de elementos multimedia adicionales al texto como imágenes, videos, gifs y enlaces. Sin embargo, uno de los componentes más icónicos y duraderos de Twitter son los *hashtags*. Los *hashtags* son etiquetas representadas por el símbolo numeral o hash (#) que se antepone a una palabra o frase en el texto del *tweet*. Estas etiquetas se han convertido en piezas fundamentales para la organización y clasificación de contenido en la plataforma, y su éxito causó la extensión a otras redes.

El problema que se presenta en este contexto radica en el uso de los *hashtags*. En las redes sociales, particularmente en Twitter, los *hashtags* se han convertido en una herramienta fundamental para etiquetar y clasificar contenido, lo que facilita la organización y búsqueda de información específica. Sin embargo, debido a su naturaleza abierta y descentralizada, se ha producido un incremento desordenado de *hashtags* lo que puede generar confusión en su significado y propósito.

Este problema se manifiesta en varios aspectos. Primero, la creación sin cesar de *hashtags* por parte de los usuarios ha llevado a una gran diversidad de etiquetas repetidas que pueden referirse a temas similares, pero carecen de estructura semántica clara. Además, se debe considerar la evolución constante de los *hashtags* en temas de popularidad y significado a lo largo del tiempo, lo que dificulta la interpretación precisa.

Los *hashtags* pueden llegar a tener éxito si se propagan de forma adecuada, generando un impacto considerable en la comunidad de usuarios, como también pueden fracasar rotundamente y pasar desapercibidos, lo que refleja la volatilidad de los *hashtags* y la incertidumbre en torno a la efectividad. La popularidad del hashtag puede surgir y decaer rápidamente y el mismo hashtag puede tener el mismo significado en distintos periodos de tiempo, como también existen variantes de un hashtag que se refieren al mismo tema. Además, la polisemia de los *hashtags* y presencia de sinónimos causa un problema adicional: diferentes *hashtags* presentes en el mismo periodo pueden compartir el mismo significado y como también de forma análoga, *hashtags* escritos de la misma forma contienen significado distinto debido al entorno multilingüe y diverso.

Otro desafío es la falta de coherencia en la clasificación y etiquetado de contenidos debido a

la ambigüedad que pueden presentar los *hashtags*, ya que estos se pueden encontrar como acrónimos o simplemente, palabras inventadas por los usuarios. Esto causa una dispersión de temas, causando que sea imposible entender el significado con solo el hashtag.

La libertad en la creación de *hashtags*, variabilidad en el uso y evolución en el tiempo complican la interpretación y categorización correcta de estos elementos, lo que a su vez, puede limitar la utilidad como herramienta de organización.

Los *hashtags* fueron introducidos como forma de clasificar los mensajes de acuerdo a un tema. Actualmente no solo se usan para clasificar sino que también son creados para comentar, alabar, criticar ideas o personas, promover marcas, eventos, breaking news, entre otros.

En cuanto a la estructura, los *hashtags* no hacen distinción entre mayúsculas y minúsculas, no permiten espacios, pueden contener números pero no pueden estar compuestos únicamente por números y solo permite el uso de guión bajo como carácter especial. Esta libertad da lugar a múltiples formas de escribir un mismo hashtag. Por ejemplo, una palabra seguida de un año puede expresarse de distintas maneras como escribiendo la palabra con sólo los dos últimos dígitos del año, el año completo o separados por un guión bajo, entre otras combinaciones posibles, sin olvidar que al no tener restricciones de idioma, se pueden encontrar diversos tipos de acentuaciones gráficas. Esta variabilidad en la escritura de los *hashtags* afecta la consistencia y comprensión de los mismos, y puede influir en su duración y propagación, dependiendo de cómo se difundan.

1.1. Propuesta de solución

Para abordar la problemática mencionada anteriormente, este estudio se propone realizar una caracterización exhaustiva de los *hashtags* en Twitter. El objetivo principal de esta investigación es desarrollar una tipología que permita agrupar estos elementos en categorías significativas y comprender su comportamiento en profundidad.

Dada la falta de restricciones en la creación de *hashtags*, la variabilidad en su uso y la presencia de una jerga específica de Twitter, el análisis convencional mediante métodos de procesamiento de lenguaje natural se vuelve insuficiente. Por lo tanto, se adoptará un enfoque innovador que aproveche las características únicas de los *hashtags* para desentrañar su significado y clasificarlos en grupos semánticos interconectados.

La metodología propuesta se basa en la extracción y análisis de atributos relacionados con los *hashtags*. Algunos de estos atributos son ampliamente conocidos, como la frecuencia de uso, la estabilidad en el tiempo, la co-ocurrencia con otros *hashtags* dentro del mismo tweet, la convergencia entre diferentes *hashtags* y la forma en que pueden moverse entre comunidades de usuarios.

Además de estos atributos, se explorarán nuevas características que puedan derivar de los ya existentes, permitiendo una comprensión más completa de los *hashtags*. Se pondrá un énfasis especial en el análisis de la evolución temporal de los *hashtags*, bajo el supuesto que esta dinámica es importante para su interpretación.

Se implementará un algoritmo de clustering o agrupamiento que permita categorizar los *hashtags* en función de los atributos identificados. Este proceso ayudará a identificar patrones y comportamientos de los *hashtags*, lo que a su vez permitirá asignar categorías significativas.

Es importante mencionar que, para fines de esta investigación, se limitará el análisis a *hashtags* escritos con el alfabeto latino, lo que se logrará mediante el uso de expresiones regulares. Esta restricción facilitará la caracterización y agrupación de los *hashtags* en función de su escritura.

Para llevar a cabo esta investigación, se utilizará un conjunto de datos obtenidos de Internet Archive [arc, 2011], que comprende tweets recopilados diariamente desde el año 2019 hasta el 2020. Estos datos incluyen información valiosa, como el ID del tweet, el autor, la fecha de creación, la ubicación geográfica y, por supuesto, los *hashtags* utilizados en cada tweet. Dada la gran cantidad de datos disponibles, se llevará a cabo una filtración inicial por geolocalización para reducir el volumen de datos sin perder de vista la evolución temporal de los *hashtags*.

1.2. Objetivos

1.2.1. Objetivo general

Elaborar una tipología de *hashtags* en base a la mayor cantidad de atributos posibles encontrados, mediante algoritmos de clustering.

1.2.2. Objetivos específicos

1. Recopilar tweets correspondientes al periodo 2019 a 2020, inclusive, para construir un nuevo conjunto de datos centralizado con la información pertinente.
2. Seleccionar características directas relacionadas a los *hashtags* y generar nuevas, con el fin de obtener la máxima cantidad de atributos posibles.
3. Emplear un algoritmo de clustering sobre un conjunto de datos creado para identificar patrones de uso entre el periodo definido y categorías, en base a las características seleccionadas.

1.3. Trabajos relacionados

En los últimos 15 años, los *hashtags* se han convertido en un símbolo y herramienta de conectividad gracias a los cambios en la tecnología. Se han vuelto ubicuos en los medios y han sido el objeto de numerosos estudios en distintas líneas de interpretación [La Rocca y Boccia Artieri, 2023].

Los *hashtags* son etiquetas que representan un tema y permite a los usuarios de redes sociales contribuir con sus propias opiniones y comentarios con el simple hecho de escribir una cadena de caracteres precedida por el símbolo de almohadilla (#) [La Rocca, 2020]. La invención del hashtag se adjudica a Chris Messina [Zak, 2013], quien introdujo los *hashtags* en Twitter en el año 2007 realizando un *tweet* que decía “*how do you feel about using # (pound) for groups. As in #barcamp [msg]?*” (¿Qué opinan de usar # para grupos como en barcamp?).

No obstante, la popularidad se atribuye a Natan Ritter, un periodista que proporcionó información en tiempo real publicando en Twitter de los incendios de San Diego en octubre de 2007 y tras recibir un mensaje de Messina, comenzó a utilizar *#sandiegofire* lo que ayudó a los ciudadanos a seguir el curso de la información de forma más organizada. Este evento demostró la utilidad de los *hashtags* para obtener información y crear comunidades en torno a eventos específicos. El uso masivo de *hashtags* se consolidó posterior a su introducción en Twitter en el año 2009 como una herramienta para agrupar los *tweets* y registrar temas de manera sencilla. A partir de entonces, los *hashtags* se convirtieron en una pieza fundamental dentro de la plataforma.

[Bruns y Burgess, 2011] expone que la capacidad de Twitter de responder rápidamente a los eventos del momento, genera la creación *hashtags* ad hoc en respuesta a los eventos actuales por los mismos usuarios. Esto permite a los *hashtags* ser utilizados para discutir las noticias de último momento y otros eventos esporádicos. Se debe considerar que no todos los *hashtags* son ad hoc, algunas comunidades pueden surgir desde la anticipación de un evento predecible o aparecer después de que el evento ya haya ocurrido.

Años más tarde, [Bruns *et al.*, 2016], amplió la noción de *hashtags* ad hoc, al lograr asociarlos a eventos mediáticos, como transmisiones de entretenimiento, eventos deportivos o cobertura de elecciones políticas y también, a otra gran categoría de situaciones de eventos repentinos como desastres naturales, ya sea inundaciones, terremotos, o crisis políticas. En este último caso, los *hashtags* surgen de manera espontánea, impulsados por la necesidad de los usuarios de compartir información, seguir noticias a tiempo real y participar en discusiones sobre el tema actual. Un ejemplo moderno es el uso de *hashtags* relacionados con la pandemia de COVID19.

La utilidad de los *hashtags* va más allá de su función de reunir a un público ad hoc en torno a un tema clave. También se utilizan para formar comunidades que participan en actividades compartidas, posiblemente recurrentes como asistir a un evento en vivo o también, utilizar

twitter como un canal secundario para las transmisiones de radio y televisivas; intentar crear y promover memes que se difunden viralmente en redes a nivel local, nacional o global. [Bruns *et al.*, 2016] plantea que un análisis más profundo podría arrojar más luz sobre la amplia gama de usos que son evidentes en Twitter en la actualidad, abarcando desde eventos políticos, eventos deportivos, etiquetas de palabras claves como lugares y términos genéricos así como la rápida respuesta a problemas o temas específicos.

Además de su uso en eventos y situaciones, los *hashtags* se han convertido en símbolos de movimientos sociales y políticos. Los usuarios los emplean para expresar activismo y participación en movimientos políticos y sociales. Ejemplos notables incluyen #MeToo y #BlackLivesMatter

La investigación sobre las características de los *hashtags* se ha enfocado en comprender su comportamiento, difusión y predicción. [Laniado y Mika, 2010] formaliza características relacionadas a los *hashtags*, incluyendo la frecuencia (cantidad de usuarios y mensajes que los utilizan), la especificidad (cuán diferentes es el uso de la misma palabra sin el símbolo #), la estabilidad en el tiempo (si se convierte en parte del vocabulario de los usuarios) y la consistencia. Estas métricas son aplicadas tanto en los mensajes que contienen el hashtag como en la comunidad que lo adopta.

Para abordar la difusión de los *hashtags* [Wang y Zheng, 2014], se debe evaluar varias propiedades asociadas a los *hashtags* en las que se puede encontrar el largo de la cadena de caracteres, la cantidad de tweets que lo contienen, su duración en el tiempo, es decir, el número de días que transcurre entre la primera y última aparición dentro del rango definido en el conjunto de datos; la cantidad de usuarios que lo utilizan por día y de forma adicional, una serie de tiempo que registra las frecuencias diarias correspondientes al hashtag. A fin de predecir la popularidad de nuevos *hashtags*, [Ma *et al.*, 2013] identifica dos tipos de características para poder cumplir el objetivo: las de contenido y las de contexto. Las características de contenido son interesantes ya que derivan intrínsecamente del hashtag como lo son la presencia de dígitos o el número de segmentos de la cadena. Por otro lado, los atributos de contexto se basan en las conexiones formadas por los usuarios e incluye la cantidad de usuarios, cantidad de tweets, menciones y retweets, entre otros.

Con respecto al clustering de *hashtags*, se destaca el estudio de [Romero *et al.*, 2011] que propuso categorizar los *hashtags* en Twitter en ocho temas distintos: políticos, modismos, celebridades, deportes, música, tecnología, películas/TV y juegos. El objetivo era investigar cómo los *hashtags* relacionados con diferentes temas se difunden de manera distinta en Twitter. Los resultados revelaron diferencias significativas en la persistencia y difusión de los *hashtags* según su clasificación temática. Por un lado se observó que los *hashtags* políticos y de juegos emergieron como temas persistentes debido a su frecuente uso y durabilidad en la comunidad. En contraste, los llamados Twitter-Idioms, que son modismos familiares para los usuarios de Twitter, mostraron una alta tasa de adopción inicial pero su persistencia era baja. Bajo este estudio, [Bastos *et al.*, 2021] extrae doce temas con la finalidad de encontrar clasificación para las noticias, lo que se da a partir de una mezcla entre las secciones de los periódicos y los ya definidos anteriormente, lo que resulta en eventos, películas, tec-

nología, deportes, política, celebridades, altruismo, música, juegos, personalidad, lifestyle, modismos.

[Dela Rosa *et al.*, 2011] explora diversas formas de realizar clustering en seis temas predefinidos como noticias, deportes, entretenimiento, ciencia, tecnología, dinero y “solo por diversión”. Algunas de las técnicas aplicadas incluyen LDA, k-means y otros algoritmos de aprendizaje supervisado. Como resultado, se crean 15 clusters que se basan en la observación de que los *hashtags* son indicadores aproximados del tema del *tweet*.

En un enfoque diferente, [Pöschko, 2011] define una red de co-ocurrencia de *hashtags* como una red ponderada que representa la frecuencia con la que los *hashtags* co-aparecen en los *tweets*. Este planteamiento resulta en la identificación de algunas clases como organización, geolocalización, personas, eventos y categoría, que incluye los *hashtags* que no cumplen las características para formar parte de las otras clases.

La revisión de los estudios relacionados proporciona una visión ampliada, siendo estos hallazgos la base para comprender los posibles resultados a obtener. Al explorar las características y diversidad de tipologías de *hashtags* en el propio contexto de esta investigación, contribuirá a una comprensión más profunda del comportamiento de los *hashtags*.

CAPÍTULO 2

MARCO CONCEPTUAL

2.1. Twitter

Twitter es una red social gratuita de microblogging donde se pueden escribir mensajes cortos, y de forma rápida, de no más de 280 caracteres, los llamados *tweets*, los cuales pueden contener texto, imágenes, videos, enlaces, entre otros y, dentro del texto se puede hacer uso de los *hashtags*.

En esta plataforma cualquier persona puede crearse una cuenta y compartir diversas ideas, desde experiencias personales a información relevante. Una de las principales características de esta red social es que la información fluye en tiempo real y logra difundirse rápidamente entre los usuarios, aunque esto muchas veces puede causar una alerta entre las personas por la difusión de noticias falsas. Un caso conocido y cercano fue durante las primeras etapas del COVID-19, la mayoría utilizaba esta red social para informarse y entender lo que ocurría alrededor del mundo, donde se publicó información errónea y se difundió, muchas veces, más rápido que la información correcta.

La comunicación de los usuarios en Twitter se basa en las conexiones que se generan entre seguidores (*followers*) y seguidos. Cuando un usuario sigue a otra persona, se suscribe a sus actualizaciones y estas aparecen en la cronología de inicio. Esta cronología o *timeline* refleja las preferencias del usuario, ya que se muestra el contenido realizado por las cuentas que decidió seguir. Además, cuando el mismo usuario genera un *tweet*, se añade el contenido a su propio perfil y es compartido en la cronología de su red de seguidores. En la figura 1 se representa de forma generalizada el funcionamiento de los usuarios en Twitter y sus conexiones entre los distintos elementos.

Los *tweets* se pueden asociar a una localización, lo que genera un *tweet* que ha sido georeferenciado. Este atributo no aparece en el conjunto total de *tweets*, sino que es una opción que el usuario debe mantener activa dentro de sus ajustes, es por esto que solo alrededor del 1-2 % del total de *tweets* se encuentran geolocalizados ¹. En la interfaz de la plataforma, esta información se muestra como el nombre del lugar, sin embargo, para efectos de este estudio, al momento de realizar un análisis esta será reflejada en puntos de coordenada, que corresponderá a latitud y longitud, por lo que se debe realizar una aproximación real para así, obtener una zona.

¹Advanced filtering with geo data - <https://developer.twitter.com/en/docs/tutorials/advanced-filtering-for-geo-data>

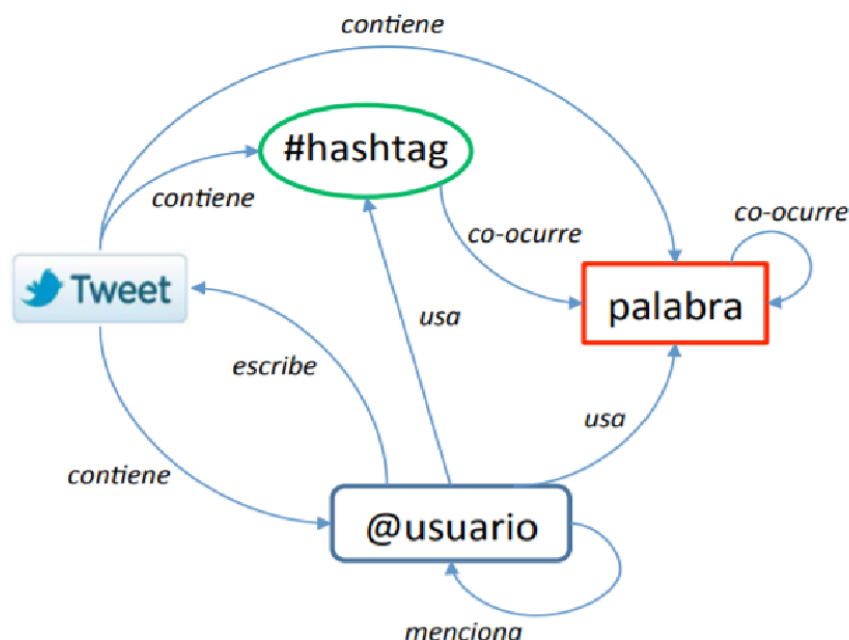


Figura 1: Conexiones en Twitter
Fuente: Obtenido de [Espin-Riofrio y Zumba, 2021]

2.2. Hashtags

Aproximadamente el 40% de los *tweets* contienen *hashtags* y este uso en Twitter lo propone Chris Messina, con la idea de mejorar la contextualización, filtrar el contenido [Messina, 2016] al crear un sistema de canales con etiquetas utilizando el símbolo de numeral o *hash* (#) los que se crearían *ad hoc* y de esa forma, las personas podrán seguir un tema específico y también contribuir en el flujo de la conversación.

Una complejidad asociada a los *hashtags* es que estos tienden a ser ambiguos. Por ejemplo, el nombre de la ciudad que hospeda a un equipo de deportes muchas veces se utiliza para hacer referencia al equipo mismo durante un juego en lugar de indicar la localidad en sí, como ocurre en otras ocasiones [DeMasi *et al.*, 2016]. También para la clasificación, como es el caso del famoso *#superbowl* que puede ser interpretado tanto como un evento como un deporte, y ambas interpretaciones son igualmente válidas.

Antes de encontrar características de los *hashtags*, se debe considerar que existe el *spam*, o correo basura. Teniendo en cuenta el comportamiento dentro de la red, donde un usuario sigue a otro y este mismo, lo sigue de vuelta por cortesía, los *spammer*, o personas que envían el correo basura, siguen cantidades de usuarios aleatorios para así, de la misma forma, obtener más seguidores con el objetivo de que sus *tweets* sean expuestos en la cronología de inicio de los usuarios. Por lo general, abusan de los *hashtags* del momento para poder llegar más allá que solo los seguidores.

En [Sedhai y Sun, 2017] se realiza un análisis para diferenciar los *tweets* que corresponden a *spam* de los que no. Para realizar la primera clasificación se seleccionan palabras claves tales como obtener dinero rápido, contenido de adulto, servicios, regalos, entre otros. Como resultado se encuentra que los *tweets* de *spam* contienen una gran cantidad de *hashtags* y que se enfocan por lo general en los de momento, o los llamados *trending topics*, incluso se podría decir que una porción significativa de los *trending topics* se deben a los *spammers*, otra característica destacada es que se encuentran escritos por lo general en mayúsculas y aparecen en medio del *tweet*, a diferencia de los otros tipos de *tweets* que los ubican usualmente por el final.

Para poder entender a que se refiere una caracterización de *hashtags*, se presentan los siguientes conceptos.

2.2.1. Estabilidad en el tiempo

Muchos de los *hashtags* se espera que se conviertan en parte del vocabulario de los usuarios, es decir, su uso debe mantenerse en el tiempo de tal forma que signifique mucho más que solo un *hashtag* [Bruns y Stieglitz, 2013]. Para medir la estabilidad en el tiempo de un *hashtag*, primero se debe conocer a qué tipo de evento pertenece pues para cada uno serviría distintas unidades de medida.

En el caso de los eventos que duran un período corto de tiempo, como lo son la transmisión de actividades deportivas o programas de televisión, un análisis de minuto a minuto considerando la franja de tiempo que se transmite sería útil; para actividades que consideran un periodo más largo, como por ejemplo las campañas presidenciales o ciertas crisis, tendría sentido establecer métricas por día; y ya para el caso de eventos que perduran en el tiempo, como lo son conflictos militares o publicidades de marcas permanentes, se podría realizar un análisis en base a los *tweets* mes a mes [Weng y Menczer, 2015]. Además no se debe olvidar los *hashtags* que marcan la historia de Twitter y que ya son parte del vocabulario de la comunidad, para esto se podría incluso realizar un análisis de comportamiento más amplio.

2.2.2. Convergencia

Distintos *hashtags* pueden compartir el mismo significado, sin embargo, estos se pueden encontrar escritos utilizando otras palabras, distinto orden e incluso otro idioma. Además, varias veces consisten en acrónimos, palabras concatenadas o abreviaciones a lo que estos también pueden estar refiriéndose al mismo concepto [Stilo y Velardi, 2016]. Cuando esto ocurre, se reduce la efectividad del objetivo de los *hashtags* ya que el flujo de *tweets* se ve afectado y también, el impacto de los *trending topics* y por lo general, existe una etiqueta que destaca sobre el resto, causando así una pérdida de información del flujo total.

2.2.3. Co-ocurrencia

La co-ocurrencia se produce cuando dentro de un mismo *tweet* se utilizan dos o más *hashtags*. En [Wang *et al.*, 2016] se define la red de la co-ocurrencia de los *hashtags* como una red compuesta por nodos que representan los distintos *hashtags* y los arcos, son las instancias de co-aparición de estas etiquetas en un solo *tweet*. Esta corresponde a un grafo con peso en donde el valor de sus aristas indica qué tan frecuentemente estos dos *hashtags* co-ocurren. También, mientras más alto sea el grado del nodo, más alta es la frecuencia que se utiliza una combinación de *hashtags*.

2.2.4. Frecuencia de uso

Los *hashtags* son utilizados con cierta frecuencia por los usuarios y está dada por la cantidad de veces que una misma etiqueta es utilizada por la comunidad de Twitter. Esta característica se puede medir en la cantidad de mensajes publicados que hacen uso de este y también, se puede realizar un estudio en base a la cantidad de usuarios que utilizan la misma etiqueta y así, identificar si son usuarios únicos o repetidos.

Se puede identificar la cantidad de veces de un *hashtag* determinado que utiliza un mismo usuario para entender el aporte que realiza al alcance del *hashtag*, conociendo así, su frecuencia de uso y de esa misma forma se puede llegar a entender la popularidad que pueda alcanzar si este usuario corresponde a una cuenta con una amplia red de seguidores, donde su misma red, seguirá utilizando el *hashtag*.

A partir de este atributo se da cuenta que los *hashtags* con mayor frecuencia de uso, representan el interés mundial de cierto tema en un periodo de tiempo específico.

2.3. ETL

ETL es un acrónimo de extraer, transformar y cargar (en inglés *extract, transform, load*). Como su nombre lo sugiere es el proceso de extraer datos de distintas fuentes y transformarlos de acuerdo a los requerimientos del usuario y posteriormente, cargarlos a un destino específico. El objetivo principal de ETL es hacer que los datos sean más valiosos y usables para los usuarios y organizaciones. Las soluciones con ETL permiten a las organizaciones integrar la información de diferentes fuentes y distintos formatos en un único repositorio lo que facilita la forma de observar los datos, analizar, generar reportes y su uso en machine learning para poder apoyar un proceso de negocios.

Como se mencionó anteriormente, este proceso consta de 3 pasos:

1. **Extraer:** Implica obtener los datos de una o varias fuentes de origen, estos pueden ser

de distintos formatos como base de datos relacionales, XML, JSON, archivos planos, entre otros. Este espacio es una oportunidad de validar la data para que cumpla con la estructura esperada y enviarla de vuelta en caso de existir errores. Las herramientas de ETL automatizan el proceso de extracción creando un flujo de trabajo más eficiente.

2. **Transformar:** Durante esta fase, se convierten los datos al formato requerido aplicando ciertas reglas para así asegurar la calidad y cumplir los requisitos. La transformación es considerada la etapa más importante de un proceso ETL. Puede comprender estos tipos de ajuste:

- Limpieza: elimina las inconsistencias y los valores con errores
- Deduplicación: Identificar y eliminar registros duplicados
- Revisión de formato: Se convierten los datos en un formato coherente. Un ejemplo simple es el dato tipo fecha.
- Clasificación: Se organiza acorde a su tipo
- Otras tareas: Reglas adicionales pueden ser aplicadas.

3. **Cargar:** En la fase final, los datos transformados se cargan en su almacenamiento destino, que puede ser un simple archivo plano o un *data warehouse*. Esta información puede ser cargada en su totalidad o de forma progresiva por lotes.

2.4. Reducción de dimensionalidad

2.4.1. Análisis de componentes principales (PCA)

El análisis de componentes principales (PCA) es una técnica fundamental en el procesamiento de datos, utiliza los principios matemáticos para transformar un número de posibles variables correlacionadas a un pequeño número de variables llamadas componentes principales. El objetivo principal del PCA es reducir las dimensiones del espacio observado y hacerlas independientes. La reducción se obtiene al crear una nueva combinación lineal de las variables que caracterizan los objetos en estudio. Estas combinaciones resultantes, son los componentes principales [Maćkiewicz y Ratajczak, 1993].

El PCA se emplea sobre todo en análisis exploratorios de datos y para construir modelos predictivos. Se utiliza comúnmente para reducción de dimensionalidad proyectando cada punto de datos en solo los primeros componentes principales para obtener datos de menor dimensión conservando tanta variación de los datos como sea posible. Para esto, se genera una matriz de covarianza que señala cómo cada par de atributos se relaciona entre sí.

Supongamos que existe una muestra con n puntos cada uno con p características $(X_1, X_2, X_3, X_4, \dots, X_P)$ es decir, el espacio muestral tiene p dimensiones. La técnica de análisis de componentes principales permite encontrar un número de atributos tal que $z < p$

que explican aproximadamente lo mismo que las p variables originales, en función de la covarianza. Donde antes se necesitaban p valores para caracterizar a cada punto, ahora es suficiente z valores; estas z variables son los componentes principales.

Para aplicar PCA existe una serie de pasos:

1. **Estandarizar los datos:** Antes de aplicar análisis de componentes principales, es importante estandarizar el conjunto de datos. El objetivo es que cada atributo tenga una media de 0 y varianza de 1. Esto es importante ya que el PCA es sensible a las diferencias de escalas entre los atributos y en el caso que existiera mucha diferencia entre las escalas, esos valores más grandes van a predominar. Esto se realiza restando la media de cada atributo a todos los valores de ese atributo y luego dividir por la desviación estándar.

$$x_{is} = \frac{x_i - \mu_x}{\sigma_x} \quad (1)$$

Donde x_{is} representa el atributo x_i estandarizado, μ_x la media de la columna del atributo y σ_x la desviación estándar.

2. **Matriz de covarianza:** Una vez que la data está estandarizada, se procede a calcular la matriz de covarianza. Esta matriz muestra cómo cada par de atributos se relacionan entre sí. Si hay d atributos, se obtiene una matriz $d \times d$

$$\sigma_{jk} = \frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \mu_j)(x_{ik} - \mu_k) \quad (2)$$

Donde x_j y x_k representan los dos atributos y σ_{jk} la covarianza entre ellos. Como la matriz ya fue estandarizada, la matriz de covarianza Σ se puede calcular de la siguiente forma:

$$\Sigma = \frac{1}{n-1} X^t X \quad (3)$$

3. **Vectores y valores propios:** Los vectores y valores propios son elementos claves del pca. Los vectores propios determinan la dirección en el espacio de los nuevos atributos, es decir, son los componentes principales, y los valores propios, determinan su magnitud. En otras palabras, los valores propios explican la varianza de la data junto a los nuevos atributos, lo que significa que explican que tanta varianza está incluida. El vector propio v se calcula como:

$$\Sigma v = \lambda v \quad (4)$$

En la ecuación λ es un escalar conocido como el valor propio.

4. **Seleccionar componentes principales:** Se ordenan de forma descendente los valores propios a partir de su magnitud. Para seleccionarlos, se elige los k primeros componentes que corresponden a los valores propios más altos. La elección de k depende de cuánta varianza se va a retener, por lo general se elige un valor que explique al menos

el 70-80 % de la varianza total. Esta selección se puede realizar a partir de un gráfico utilizando:

$$\frac{\lambda_j}{\sum_{j=1}^d \lambda_j} \quad (5)$$

5. **Proyección:** Se proyectan los datos originales en el nuevo espacio definido por los componentes principales. Este nuevo conjunto es de dimensión reducida pero que conserva la mayor parte de la información original relevante. Se puede calcular a partir de:

$$Z = XW \quad (6)$$

2.4.2. Uniform Manifold Approximation and Projection (UMAP)

[McInnes *et al.*, 2018] describe de forma teórica el algoritmo en función de complejos simpliciales difusos. Desde el punto de vista computacional, sólo es posible para el esqueleto uno, que como última instancia podría describirse como un grafo ponderado. Un complejo simplicial es un medio para construir espacios topológicos a partir de componentes combinatorios simples, en un lenguaje menos técnico, se podría decir que son como bloques de construcción para armar espacios complicados de manera simple. Para entender se debe proporcionar algunos bloques simples llamados “simplex”. Geométricamente, un simplex de dimensionalidad k es llamado k -simplex. En la figura 2 se puede ver una representación de algunos simplex, como el 0-simplex al 4-simplex. En la definición anterior de UMAP, cuando se habla de esqueleto 1, se refiere a 1-simplex.

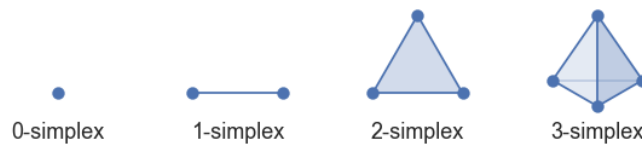


Figura 2: Representación simplex

Fuente: Documentación UMAP ([McInnes *et al.*, 2018]).

El objetivo principal de este algoritmo es mapear los datos de alta dimensionalidad a un espacio de baja dimensionalidad preservando la estructura topológica de los datos, generalmente se reduce a dos o tres dimensiones, lo que puede ser más fácil de visualizar y analizar.

Al igual que otros algoritmos basados en grafos de k -vecinos, UMAP se puede describir en dos fases. En la primera, se construye el grafo ponderado donde se representan las relaciones entre los puntos y en la segunda, se calcula el diseño de baja dimensión del grafo, buscando conservar la estructura topológica del grafo construido en la primera fase.

En términos de complejidad computacional, se considera un algoritmo rápido y escalable por lo que es adecuado para conjuntos de datos grandes y con gran dimensionalidad. Para ejecutar el algoritmo se debe tener en cuenta algunos parámetros importantes:

- **Número de vecinos:** Este parámetro influye en la estructura del grafo que se construye en la primera iteración. Los valores más pequeños se centran en la estructura local, considerando las relaciones cercanas entre los datos. Por otro lado, los valores más grandes representan una estructura más general, quitando detalles finos.
- **Mínima distancia:** Controla la distancia mínima entre los puntos en el espacio resultante de baja dimensión. Un valor más bajo permite que los puntos se agrupen más cerca del espacio proyectado, es decir, grupos más densos y compactos. Por el contrario, con valores más altos se evita que los puntos se agrupen cerca, lo que se centra en la estructura general.
- **Métrica:** Se define la métrica que se utiliza para medir la similitud entre los puntos. Esta puede ser distancia euclidiana, coseno, entre otras.
- **Número de componentes:** El número de componentes o dimensión, se refiere a la cantidad de dimensiones en la proyección final después del proceso de reducción. Por lo general, se eligen dos o tres para lograr visualizar los datos, sin embargo, no existen limitaciones de este número.

2.5. Algoritmos de agrupación o *clustering*

Una de las principales técnicas de aprendizaje no supervisado es el *clustering*. Un algoritmo de agrupación se caracteriza por particionar un conjunto en subconjuntos con el objetivo de encontrar grupos de instancias, llamados *clusters*. En tal caso cada subconjunto corresponde a un *cluster* y sus objetos pertenecientes poseen características similares entre sí y por el lado contrario, son diferentes a las de otros grupos. En la figura 3 se representa el funcionamiento de un algoritmo de *clustering* a modo general. Para este caso, los *hashtags* serán el conjunto de datos a trabajar.

A grandes rasgos un proceso de aprendizaje automático no supervisado, significa que las etiquetas de los datos no se encuentran presentes y se clasifican a partir de sus características. Además, es un método de aprendizaje por observación más que aprender por ejemplos ya conocidos.

El *clustering* es una de las herramientas más utilizadas durante el análisis exploratorio de datos (EDA) para conocer un primer acercamiento a la estructura de los datos y puede ser útil ya que puede derivar en el descubrimiento de información no conocida anteriormente. Se utiliza como un preprocesamiento para luego poder utilizar otros algoritmos según corresponda.

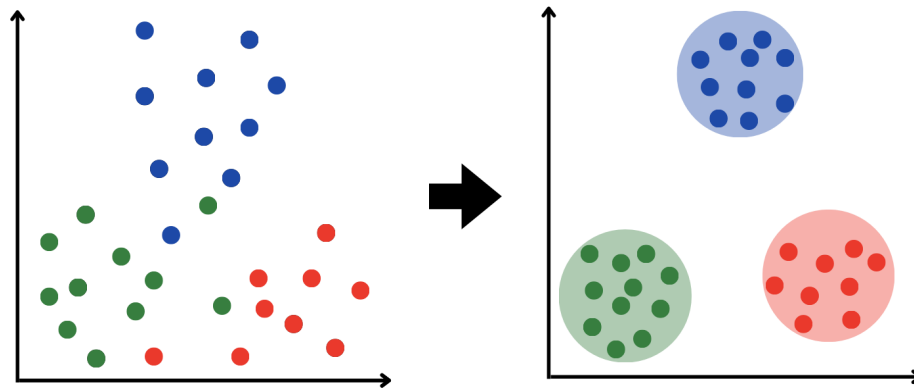


Figura 3: Representación de aplicar un algoritmo de clustering
Fuente: Elaboración propia.

Existen diversos tipos de algoritmos de *clustering* y cada uno selecciona los *clusters* de manera distinta, para este estudio se utilizará *k-means*, ya que al realizar pruebas preliminares, era el menos costoso computacionalmente.

2.5.1. Algoritmo *k-means*

El algoritmo *k-means* busca instancias centradas en un punto determinado, llamado centroide. Este algoritmo es un método muy utilizado por su facilidad de implementación y su eficiencia desde el punto de vista computacional. El objetivo es encontrar y asignar a cada centroide un grupo en base a la similitud de sus características.

Para poder ejecutar este algoritmo, se debe tener a consideración su funcionamiento [Han et al., 2011]. Además, en la figura 4 se puede visualizar un pseudocódigo de *k-means*.

1. Se eligen la cantidad de *clusters* deseados, es decir, la cantidad de segmentaciones de los datos.
2. Para la asignación de puntos, se calcula la distancia de cada punto a cada centroide y de esta forma, se agrupa con el que se encuentre más próximo.
3. La cantidad de centroides corresponde a la definida de *clusters* por lo que se establece esta cantidad en el espacio y sus coordenadas se inicializan de forma aleatoria.
4. Cuando todos los puntos se encuentran asignados, se recalcula los centroides de manera que vuelvan a ser el centro del *cluster*.
5. Se repite el proceso de asignar (3) y calcular (4) hasta llegar al criterio definido para detener el algoritmo.

Input

K : cantidad de clusters

D : conjunto de datos con n elementos

Output

conjunto de K clusters.

K-means algorithm

elegir K objetos de D para los centroides iniciales.

repeat

asignar cada objeto a un cluster basado en la distancia más cercana al centroide

actualizar el centroide de cada cluster

until se alcance el criterio de convergencia

Figura 4: Pseudocódigo de algoritmo *k-means*

Fuente: Elaboración propia (basado en [Mahesh, 2020]).

Para los criterios de detención estos pueden ser los siguientes:

- Los centroides dejan de cambiar, esto quiere decir que después de múltiples iteraciones los centroides no cambian su posición por lo que se asume que se ha llegado al punto de convergencia.
- Los puntos ya no cambian de grupo, al realizar la asignación y cálculo, los puntos dejan de moverse entre *clusters*.
- Se puede delimitar fijando un número de iteraciones máximo.

Para saber a qué agrupamiento corresponde un punto, la distancia comúnmente utilizada en este algoritmo corresponde a la distancia euclidiana entre dos puntos [Han *et al.*, 2011], en este caso el punto p y el centroide c_i , que se verá representado de forma general como:

$$dist(x, y)^2 = \sum_{i=1}^n (x_i - y_i)^2 = ||x - y||_2^2 \quad (7)$$

Donde x e y son el punto y el centroide respectivamente, representando $dist(p, c_i)$. Existe la llamada inercia de *cluster* que corresponde a la suma de los errores al cuadrado entre los objetos dentro del *cluster* (C_i) y el centroide, definida como:

$$E = \sum_{i=1}^k \sum_{p \in C_i} dist(p, c_i)^2 \quad (8)$$

Donde E representa la suma de los errores de todos los objetos del *dataset*, p es el punto de un objeto y c_i el centroide del *cluster* C_i . Esta función objetivo intenta que los k *clusters* resultantes se encuentren lo más separados posibles.

Como se menciona, el algoritmo *k-means* requiere como único hiperparámetro el valor de k , es decir, la cantidad de *clusters* a generar. Para seleccionar este valor, existen diversas heurísticas, se define a grandes rasgos las que se utilizarán a lo largo de este trabajo:

- **Método del codo:** Se enfoca principalmente en minimizar la suma de las distancias entre cada uno de los puntos y el centroide respectivo en el *cluster*. A medida que se incrementa el valor de k , esta distancia tiende a disminuir ya que los puntos se encuentran más cerca del centroide. Sin embargo, existe un punto de inflexión, conocido como “el codo”, que representa el número óptimo de *clusters*. Este método es útil pero puede ser ambiguo muchas veces ya que el codo no siempre se encuentra bien definido en el gráfico.
- **Índice de Calinski-Harabasz:** Esta técnica evalúa la calidad del *cluster* basándose en su compacidad y separación. Un valor más alto indica *clusters* más densos y bien separados. En el gráfico se busca el punto que maximiza la función.
- **Índice de Davies-Bouldin:** Mide la similitud promedio de cada *cluster* con el más similar a este. Mientras menor sea la similitud, mejor separados están los *clusters*. A diferencia de otros índices, una disminución en este señala una mejora en el *clustering*.
- **Estadístico de gap:** Compara la varianza intra-*cluster* con una distribución de referencia. Busca el valor en el que K maximiza la diferencia entre estas varianzas.

CAPÍTULO 3

PROPUESTA E IMPLEMENTACIÓN DE LA SOLUCIÓN

3.1. Descripción de los datos

Los datos a analizar se obtienen directamente desde Internet Archive. Esta es una biblioteca digital sin fines de lucro que provee acceso gratuito al material para investigadores, historiadores y estudiantes. Su misión es proveer acceso universal a todo el conocimiento. Respecto a los datos de Twitter disponibles en Internet Archive, estos fueron obtenidos desde el flujo real de Tweets utilizando la API de Twitter y corresponde al 1 % de los tweets diarios. Para esta investigación se recolectaron datos desde el 2019 al 2020, inclusive, lo que se refleja en aproximadamente 1500 millones de tweets.

A grandes rasgos, este dataset se encuentra organizado por mes y año, y cada uno de estos, contiene archivos comprimidos en formato .tar. Dentro existe una gran cantidad de archivos en formato .json, comprimidos en archivos .bzip, donde se encuentra el flujo de tweets correspondiente.

Para entender, se debe conocer el modelo de los datos que maneja Twitter. Lo que se entrega es una serie de datos semi-estructurados en formato JavaScript Object Notation (JSON). Se puede observar un ejemplo de Tweet en la figura 5

```
{
  "created_at": "Fri Sep 18 18:36:15 +0000 2020",
  "id": 1307025659294674945,
  "text": "Here's an article that highlights the updates in the new Tweet payload v2 https://t.co/oeF3ZHeKQQ",
  "truncated": false,
  "entities": {
    "hashtags": [],
    "symbols": [],
    "user_mentions": [],
    "urls": [
      {
        "url": "https://t.co/oeF3ZHeKQQ",
        "expanded_url": "https://dev.to/twitterdev/understanding-the-new-tweet-payload-in-the-twitter-api-v2-1fg5",
        "display_url": "dev.to/twitterdev/und...",
        "indices": [ 74, 97 ]
      }
    ]
  },
  "source": "<a href='\"https://mobile.twitter.com/\"' rel='\"nofollow\"'>Twitter Web App</a>",
  "in_reply_to_status_id": 1304102743196356610,
  "in_reply_to_user_id": 2244994945,
  "in_reply_to_screen_name": "TwitterDev",
  "user": {
    "id": 2244994945,
    "name": "Twitter Dev",
    "screen_name": "TwitterDev",
    "location": "127.0.0.1",
    "description": "The voice of the #TwitterDev team and your official source for updates, news, and events, related to the #TwitterAPI.",
    "url": "https://t.co/3ZX3TNiZCY",
    "followers_count": 513958,
    "friends_count": 2039,
    "created_at": "Sat Dec 14 04:35:55 +0000 2013",
    "geo": null,
    "coordinates": null,
    "place": null,
    "contributors": null,
    "is_quote_status": false,
    "retweet_count": 11,
    "favorite_count": 70,
    "favorited": false,
    "retweeted": false,
    "lang": "en"
  }
}
```

Figura 5: Ejemplo de tweet
Fuente: Obtenido de Twitter Developer Web

El objeto principal es el Tweet que corresponde a un JSON que contiene más de 20 atributos en su jerarquía más alta y además es objeto padre de varios objetos hijos, siendo estos, actualmente, *User* (usuario) que describe al autor del tweet; *Entities* (entidades) que encapsula arreglos de *hashtags*, menciones de usuarios, URLs, *cashtags* y media; Extended entities (entidades extendidas), *Place* (lugar) en caso de que el tweet se encuentre geolocalizado; extended Tweet (tweet extendido) que este objeto aparece cuando el tweet supera los caracteres permitidos inicialmente; Retweeted status que indica cuando un tweet ha sido retwitteado. Además, existen otros atributos relevantes para este estudio que son la fecha de creación, el identificador único, coordenadas y el idioma lo que se puede observar en la tabla 2 junto a su descripción.

Tabla 1: Atributos de un tweet.

Fuente: Elaboración Propia.

Atributo	Descripción
created_at	Fecha y hora en formato UTC que el tweet fue creado
id	Representación en entero del identificador único del tweet.
text	El texto del tweet en formato UTF-8
coordinates	Representa la ubicación geográfica del tweet reportado por el usuario. Se encuentra en formato geoJSON, se muestra longitud y latitud, en ese orden.
place	Indica que el tweet esta asociado a un lugar. Es un objeto.
retweeted_status	Contiene una representación del tweet original que fue retwitteado
entities	Entidades que son analizadas fuera del texto del tweet. Es un objeto.
lang	Indica el identificador del lenguaje en formato BCP47 correspondiente al idioma del tweet identificado automaticamente
extended_tweet	Contiene el texto del tweet completo y las entidades correspondientes.

En el 2017 fueron introducidos los tweets de 280 caracteres por lo que se creó un nuevo objeto llamado Extended Tweet, con el objetivo de encapsular los mensajes más largos y de esa forma no destruir las miles de aplicaciones que utilizan y analizan los objetos ya existentes. Para mantener esta compatibilidad el campo text original aun existe pero solo logra abarcar hasta los 140 caracteres, en el caso que un tweet sobrepase esta cantidad el texto se verá incompleto y truncado. Dentro de este objeto existe un campo full_text que contiene el mensaje completo así como también, existe un objeto de entidades que contiene los *hashtags* correspondientes. Este nuevo atributo se identifica a partir del atributo truncated,

que si su valor corresponde a true, los campos de texto y *hashtags* deberán ser analizados a partir del `extended_tweet`.

Los retweets son representados con el patrón “RT @” en el atributo de tweet. Sin embargo, al existir este atributo en la metadata, se puede encontrar las características asociadas al tweet original dentro del objeto `retweeted_status` lo que es fundamental considerar al momento de realizar esta investigación ya que, puede existir más de un registro que contenga un retweet al mismo tweet lo que causaría información duplicada por lo que es importante conocer el identificador único del tweet.

Las entidades proporcionan metadatos y contexto adicional acerca del contenido que ha sido publicado. El objeto encapsula arreglos de distintos elementos que se incluyen en los tweets como *hashtags*, menciones a usuarios, enlaces, símbolos, encuestas de twitter, multimedia incrustada, entre otros. En los metadatos del tweet, siempre se incluye el objeto de entidades con el mínimo de *hashtags*, URLs, menciones a usuarios y símbolos, incluso si ninguna de estas entidades son parte del texto, en ese caso los arreglos están vacíos. Este objeto es importante para los desarrolladores cuando consumen datos de Twitter ya que se puede ir directamente a esta sección del JSON para obtener información y no buscar de forma explícita dentro del texto. Como se mencionó, los *hashtags* son parte de este objeto los que se representan con un arreglo donde cada elemento corresponde a un objeto (Tabla 3) que contiene el texto del *hashtag* presente en el tweet y su posición.

Tabla 2: Atributos de un tweet.

Fuente: Elaboración Propia.

Atributo	Descripción
indices	Arreglo de enteros que indica el inicio y fin del texto del <i>hashtag</i> . El primer entero representa la ubicación del símbolo # y el segundo entero, la ubicación del primer carácter después del <i>hashtag</i> .
text	El texto del <i>hashtag</i> sin considerar el carácter #.

Algunos tweets se encuentran georreferenciados, esto significa que se encuentran asociados a una localización geográfica y existen tres formas principales de realizar esta acción: referencias geográficas dentro del tweet, georreferenciados por el mismo usuario, la ubicación fijada por el usuario en su perfil. Aproximadamente el 1-2 % de los tweets se encuentran geolocalizados usando alguno de los métodos mencionados anteriormente. En cuanto a esta investigación, solo se tendrá en cuenta la segunda forma, esta puede ser asignada al utilizar la interfaz usuaria de Twitter o cuando se publica a través de la API. Dentro del JSON puede ser identificado como un punto específico representado por latitud y longitud (coordenadas) o como el objeto Place. Cuando se encuentra un punto de coordenadas, se representa la localización exacta del tweet pero este tipo de geolocalización no contiene información

contextual asociada como lo hace el objeto Place, lo que se puede observar en la Tabla 2. Este objeto siempre está presente cuando el tweet es geolocalizado mientras que el atributo coordinates solo está presente cuando se asigna una ubicación exacta. Los tweets asociados a una ubicación no son necesariamente emitidos exactamente desde esa ubicación pero es una aproximación ya que cuando los usuarios deciden utilizar esta opción, se les presenta una lista con los posibles candidatos para ubicaciones de Twitter.

Tabla 3: Twitter Place.

Fuente: Elaboración Propia.

Atributo	Descripción
place_type	El tipo de ubicación representado por este lugar
name	Nombre acortado del lugar.
full_name	Nombre completo del lugar.
country_code	Código de país
country	Nombre del país

3.2. Procesamiento de los datos

Para abordar el procesamiento de la gran cantidad de datos obtenidos, se emplea el potencial dado por Talend Open Studio, una herramienta de integración de datos de código abierto. Esta plataforma permite la unificación, transformación y consolidación de datos provenientes de diversas fuentes en un único destino de manera altamente eficiente. Una de las características destacables de Talend radica en su interfaz intuitiva, donde los componentes se arrastran y se sueltan en el lienzo de diseño, conectándolos mediante enlaces que representan los flujos de datos. Este mecanismo facilita el modelado de transformaciones de datos, generando un código en Java de manera automatizada en lugar de requerir la escritura de código manual.

La organización de los datos a trabajar se estructura en cuatro niveles jerárquicos de ficheros, lo cual provee un marco coherente para el proceso de análisis. En primer lugar, los datos se distribuyen en carpetas correspondientes a años y meses específicos, lo que facilita la segmentación temporal. Dentro de estas carpetas se alberga archivos comprimidos, cada uno representando un día del mes en formato ZIP. Estos archivos comprimidos, a su vez, contienen carpetas que almacenan los archivos JSON en formato comprimido BZIP. Esta estructura proporciona una forma eficiente de segmentar y acceder a los datos de manera secuencial, permitiendo un enfoque escalonado en el preprocesamiento.

La etapa de procesamiento y limpieza de datos representa un punto primordial en el flujo de trabajo, donde se llevan a cabo transformaciones significativas para asegurar la relevancia de los datos obtenidos. En esta etapa, se aplica una serie de operaciones que permite

transformar los datos en bruto a un conjunto coherente y pulido, de esta forma podrá ser utilizado para el análisis.

Uno de los primeros pasos es la normalización de *hashtags*, dado que los *hashtags* no son sensibles a mayúsculas y minúsculas, se convierte todo el texto del *hashtag* a minúsculas. Esta normalización asegura que los *hashtags* con términos idénticos se agrupen correctamente, independientemente de cómo se encuentren escritos originalmente. En el siguiente paso se procede a eliminar los ID duplicados presentes en los tweets. La duplicación de ID es común cuando existen retweets por lo que al eliminarlos, se obtiene una representación más precisa dado que los retweets no aumentan la cantidad de *hashtags*.

Para continuar el proceso, se realizan filtros selectivos. Primero, los tweets que carecen de información de ubicación geográfica se descartan y seguidamente, los que no posean *hashtags* también son desechados, debido a que estos elementos son fundamentales para el análisis y comprensión de los patrones, lo que ayuda a reducir el ruido en los datos y garantiza que los tweets seleccionados sean relevantes para los objetivos de la investigación.

Dado que los archivos se leen en una codificación específica, en este caso corresponde a ISO-8859-15, se implementa una comprobación adicional para eliminar los tweets que contengan caracteres no latinos. Esto incluye caracteres de idiomas como por ejemplo chino, coreano, ruso o arabe, que podrían causar problemas en la interpretación y análisis. La limpieza de estos caracteres no sólo asegura la consistencia de la información, sino que también asegura la calidad y precisión de los resultados posteriores.

En resumen, la etapa de procesamiento y limpieza no sólo transforma los datos para que sean más manejables y coherentes, sino que también refina el conjunto de datos para garantizar que solo los tweets relevantes se incluyan en el análisis a realizar.

La fase de generación de output final representa el término del procesamiento de datos. En esta etapa los resultados obtenidos se representan de manera estructurada y coherente, listos para ser explorados a partir de los objetivos definidos.

Después de aplicar las transformaciones y filtraciones pertinentes, se produce un conjunto de datos pulidos que incluye únicamente los atributos esenciales y relevantes para temas de esta investigación. Esto implica elementos cruciales como ID del tweet, el texto del tweet en sí, la geolocalización asociada, *hashtags* utilizados, urls compartidas y las menciones de usuarios.

Una vez organizados en un formato final, los datos se representan en un archivo con formato CSV (Valores separados por comas) donde cada línea representa un tweet individual y contiene información que permite contextualizar y comprender el contenido, lo que se traduce en un total de 5492927 tweets y 12746772 *hashtags* totales.

Este archivo final, al estar libre de duplicados, caracteres no deseados y elementos irrelevantes, sienta las bases para continuar y realizar un análisis más detallado lo que se convierte

en una herramienta valiosa.

3.3. Características de *hashtags*

Se pueden encontrar diversas características para los *hashtags*, tanto de contenido como contextuales. Esta investigación se centrará exclusivamente en las características intrínsecas de los *hashtags*, sin considerar un análisis léxico o semántico. La razón detrás de esta decisión radica en la complejidad derivada de la diversidad de idiomas presentes en Twitter, lo que hace que la identificación precisa de significados sea compleja. Las características a explorar incluyen la frecuencia de uso, dinámica temporal, geolocalización, presencia de dígitos y la representación vectorial mediante Word Embedding. Cada uno de estos atributos ofrece una perspectiva única de cómo los *hashtags* son utilizados en la plataforma. A través de ejemplos concretos obtenidos del conjunto de datos creado se podrá ilustrar de mejor manera cada una de las características encontradas.

3.3.1. Frecuencia

La frecuencia de los *hashtags* es una métrica esencial para entender el comportamiento de los usuarios en la plataforma y la relevancia de los temas de conversación. A medida que los usuarios utilizan estos *hashtags* en sus tweets para hablar de distintos temas, los *hashtags* se convierten en etiquetas que categorizan un tópico. A través de esto, las estadísticas de Twitter pueden rastrear cuantas veces se utiliza un *hashtag* en todos los tweets publicados.

En la interfaz de Twitter, los usuarios pueden encontrar los Trending Topics (temas del momento) que muestran los *hashtags* más utilizados del momento. Esta función permite conocer los temas de mayor relevancia en tiempo real, siendo esto una representación de la intensidad con la que se utilizan diferentes *hashtags* en la plataforma y esta intensidad, se mide mediante el conteo de frecuencias.

La popularidad de un *hashtag* se basa en la cantidad de veces que ha sido utilizado en el conjunto de tweets. Esta popularidad está estrechamente relacionada con los temas que se están discutiendo en el momento dado, lo que puede incluir eventos repentinos o recurrentes.

Para cuantificar esta métrica, Twitter realiza un seguimiento del número de veces que se ha utilizado un mismo *hashtag* en los tweets de los usuarios. A partir del output generado, se debe considerar que los retweets no aumentan la frecuencia de los *hashtags*, debido a que se toma en cuenta como un solo evento al considerar el ID correspondiente al tweet original, independientemente de cuántas veces se comparta. Además, si un usuario repite varias veces el mismo *hashtag* en un solo tweet, esto tampoco aumenta su frecuencia. En la tabla 4 se puede observar los 10 *hashtags* con mayor frecuencia encontrados.

Tabla 4: 10 *hashtags* con mayor frecuencia

Fuente: Elaboración Propia.

Hashtag	Frecuencia
covid19	44785
coronavirus	24631
love	16278
bbmastopsocial	16129
photography	14774
bts	13845
peckpalitchoke	13357
travel	13284
covid_19	12949
art	12655

3.3.2. Co-ocurrencia

En el estudio respecto a la co-ocurrencia de *hashtags*, se analizan dos aspectos principales: la cantidad de veces que un *hashtag* co-ocurre con otros *hashtags* y la cantidad de *hashtags* distintos con los que co-ocurre, sin repetir. Es importante destacar que algunos *hashtags* pueden no contar con co-ocurrencias, lo que significa que nunca aparecen junto a otros *hashtags* en los tweets. En estos casos, se asigna un valor igual a 0 para ambos casos de la co-ocurrencia.

Para determinar la co-ocurrencia de un *hashtag* de manera cuantificada, se realiza el siguiente proceso:

1. En primer lugar se recopilan todos los *hashtags* presentes en cada tweet.
2. Cada conjunto de *hashtag* se transforma en una lista y se ordena alfabéticamente, para así asegurar un recuento exacto y evitar errores de duplicación.
3. Después de ordenar las listas idénticas, se agrupan junto a la frecuencia de repetición, es decir, se identifican las listas de *hashtags* que se repiten en múltiples tweets.
4. Para la cantidad de co-ocurrencias, se registra cuántas veces se observa que un *hashtag* aparezca con otros en los tweets, independientemente cuáles sean los otros *hashtags* o si existen repeticiones.
5. Para la cantidad de *hashtags* con los que co-ocurre un *hashtag* en particular, se considera con cuántos *hashtags* únicos diferentes se mencionan junto a este en los tweets. Esto logra entregar información sobre la variedad de temas relacionados con el *hashtag* central y su relevancia en un contexto más amplio.

Cabe mencionar que un *hashtag* al co-ocurrir con otros, puede impactar directamente en su popularidad ya que, puede aumentar la visibilidad relacionada con el tema central del *hashtag*, por lo que muchas veces es utilizado como estrategia por los usuarios para ampliar el alcance dentro de la plataforma, lo que también muchas veces se puede relacionar con el spam. Si bien la co-ocurrencia puede ser analizada mediante matrices, para este caso se vuelve computacionalmente costosa debido a la gran cantidad de datos y la variabilidad en la frecuencia de cada uno de los *hashtags*, es por esto que se decide utilizar este enfoque para calcular esta métrica y así, poder obtener información relevante sobre las interacciones de los *hashtags* en la plataforma.

3.3.3. Largo cadena de texto

En este análisis, otra característica encontrada, se relaciona con la cantidad de caracteres presentes en cada *hashtag*. El largo de la cadena de texto puede tener un impacto significativo en su utilidad y a su vez, comprensión dentro de los tweets.

Desde la perspectiva de los tweets, los *hashtags* con más caracteres no siempre son buenos, o “económicos” puesto que los tweets poseen un límite de caracteres por lo que los *hashtags* más largos pueden consumir una gran parte de este mensaje, lo que podría ser problemático, sin embargo, muchas veces pueden proporcionar una gran cantidad de información ya que con solo leer el *hashtag* se puede entender de lo que se está hablando.

Por otro lado, los *hashtags* que por lo general contienen 2 o 3 caracteres no proporcionan suficiente información por sí solos para lograr entender su contexto. La mayoría de las veces estos *hashtags* son abreviaturas o acrónimos que necesitan de un contexto adicional o conocimiento previo para comprender el significado.

3.3.4. Temporalidad

Se realiza un análisis de la temporalidad de cada *hashtag*. Para normalizar las fechas y facilitar el análisis, se asignan valores numéricos del 1 al 731, correspondientes a fechas desde el 1 de enero del 2019 hasta el 30 de diciembre del 2020.

Para esto, a partir del archivo obtenido de tweets, se crea un objeto JSON para cada registro de *hashtag*. En este objeto se incluye la fecha en que se publicó en formato “YYYY-MM-DD” (sin considerar la hora) como llave del objeto y como valor, se encuentra su frecuencia asociada a la fecha. En función de esto, se logra obtener 3 características relacionadas netamente al tiempo:

- **Primera fecha de aparición:** Indica la fecha en que se utilizó el *hashtag* por primera vez dentro del rango de tiempo definido.

- **Última aparición:** Siguiendo lo anterior, muestra la fecha más cercana antes de la finalización del periodo de análisis.
- **Rango temporal:** Representa el rango en días que estuvo vigente el *hashtag*, es decir, la diferencia entre la última y primera fecha de aparición dentro del conjunto de datos.

3.3.5. Cantidad de países

Como se ha mencionado anteriormente, cada tweet considerado cuenta con su información de geolocalización. Esta información se representa mediante coordenadas geográficas o el nombre del lugar que incluye detalles como el nombre exacto del lugar, el país al que pertenece y el código de país correspondiente. Para garantizar la uniformidad y estandarización en el análisis, se utiliza el código de país en lugar del nombre ya que se ha observado que el nombre del país o lugar exacto puede variar según el idioma de preferencia del usuario u otros factores. Los códigos de país, basados en la norma ISO 3166-2², son únicos y no cambian con respecto al idioma, lo que se convierte en una elección sólida para la identificación de regiones para un *hashtag*.

A cada *hashtag*, se le asigna un valor numérico que representa la cantidad de países en los que se ha twitteado. Este valor puede indicar la actividad dentro de una región específica a nivel país y si se utiliza principalmente en solo una región específica o en varias. Esta métrica es útil para entender la diversidad y el alcance de un *hashtag*. Entiéndase que mientras más países estén asociados a un *hashtag*, mayor será su alcance y por ende su capacidad para llegar a una mayor audiencia.

3.3.6. Dígitos

La presencia de dígitos dentro de un *hashtag* puede aportar contexto adicional. Puede proporcionar información temporal como por ejemplo fechas, horas, ediciones de eventos específicos, lanzamientos de productos, entre otros; también puede ser utilizado para ciertas dinámicas conocidas dentro de la comunidad de Twitter como por ejemplo #Top10.

Esta característica es representada con una variable binaria para describir la presencia de dígitos dentro del *hashtag*. Si posee al menos un dígito, se asigna un valor de 1; en caso contrario, se asigna el valor 0.

²<https://www.iso.org/>

3.3.7. Derivados de frecuencia y tiempo

A partir de las características de la frecuencia y tiempo, se generan dos vectores de frecuencias: uno diario y otro mensual, los cuales serán utilizados para encontrar otras características.

El vector de frecuencias diarias tiene una longitud de 731 elementos, representando cada día dentro del periodo establecido. Si un *hashtag* no se utiliza en un día determinado, su frecuencia se asigna como 0 para así lograr mantener la dimensión del vector. De forma análoga, el vector de frecuencias mensuales posee una longitud de 24 elementos donde cada uno representa los meses del periodo. Se emplea el mismo mecanismo explicado anteriormente para asignar valores al vector.

En función de los vectores generados, se puede extraer nuevas características asociadas a cada *hashtag*:

- **Máxima frecuencia:** Se refiere al valor más alto en el vector de frecuencias diarias, se representa como la frecuencia relativa, calculada dividiendo la máxima frecuencia absoluta del *hashtag* entre su frecuencia total.
- **Fecha de máxima frecuencia:** Indica la fecha en que ocurrió la máxima frecuencia de un *hashtag* en términos porcentuales o decimales respecto al período total. Esto se obtiene del vector diario, identificando el día asociado a la máxima frecuencia.
- **Media diaria y mensual:** Las medias para cada caso se calculan a partir de los vectores de frecuencias diarias y mensuales, respectivamente. Representan el equilibrio entre los datos, proporcionando una medida de tendencia central para cada período.
- **Coefficiente de variación diario y mensual:** Indica la dispersión de los datos y se calcula tanto para el vector diario como para el mensual. Se obtiene dividiendo la desviación estándar entre la media, para cada caso con sus variables correspondientes.
- **Pendiente:** La pendiente se deriva del vector mensual, lo que permite definir la relación entre las variables y estimar una tasa de variación promedio en el tiempo. Esto puede proporcionar información sobre la tendencia de uso de un *hashtag* a lo largo de los meses en el rango definido.
- **Razón de coeficiente de variación:** Se obtiene a partir de la comparación entre el coeficiente de variación mensual y diario calculados anteriormente. Se utiliza para evaluar la variabilidad relativa de los dos aspectos distintos de los *hashtags*, en este caso se refiere a la frecuencia diaria y mensual.

3.3.8. Vectorización de palabras

En el procesamiento de lenguaje natural (NLP), Word Embedding se refiere a un conjunto de modelos y técnicas donde las palabras del lenguaje natural se representan como vectores de números reales en un espacio multidimensional. Dentro de esta categoría existen diversos algoritmos, entre ellos Word2Vec, es una de las técnicas más conocidas y efectivas para lograr la vectorización.

Word2Vec es un algoritmo que emplea modelos de redes neuronales para aprender asociaciones de palabras a partir de un gran corpus de texto. Una vez que el modelo está entrenado, puede detectar palabras sinónimas. Cada palabra distinta se representa mediante un vector de números y estos vectores, están diseñados de manera que la similitud semántica entre palabras se pueda medir mediante funciones matemáticas, como coseno entre vectores.

Para entrenar un modelo Word2Vec se utiliza un conjunto de datos que generalmente consiste en un gran número de textos. Para este contexto, se utilizará la librería Gensim Word2Vec en Python para entrenar un modelo utilizando las co-ocurrencias de los *hashtags* en todos los tweets del conjunto de datos, lo que conforma el corpus del modelo, esto quiere decir que, a través de las co-ocurrencias el modelo aprende, el algoritmo observa cómo los *hashtags* se relacionan entre sí en función de cómo aparecen juntos en los tweets, independiente del contexto específico del tweet. De forma implícita, es que si dos *hashtags* tienden a aparecer juntos en múltiples tweets, es probable que estén relacionados en términos de significado o temas.

Esto permite que el modelo capture las relaciones semánticas entre los *hashtags*. En palabras simples, si dos o más *hashtags* co-ocurren en tweets, el modelo aprenderá que estos *hashtags* tienen una similitud. Por ejemplo, para simplificar, si dos *hashtags* como perro y mascota co-ocurren con frecuencia, el modelo aprenderá que tienen una relación semántica bajo el contexto de los tweets presentes en el conjunto de datos, a pesar de no conocer el contenido, y así para cada conjunto de co-ocurrencias. Esta relación semántica deriva directamente de cómo los usuarios utilizan los *hashtags* en sus publicaciones.

Como resultado, cada palabra, en este caso *hashtags*, será representada como un vector de dimensión 100. A partir de esta representación, mediante el coseno entre vectores, es posible encontrar la similitud entre los *hashtags*. Cuanto más cercano se encuentren los vectores, más relacionadas están las palabras. Por ejemplo, esta cercanía podría significar *hashtags* similares en términos de tema, lo que sería útil para categorizar. Esta técnica es valiosa para analizar *hashtags* ya que permite relacionar los *hashtags* entre sí y entender su similitud.

Considerando el *hashtag* con mayor frecuencia correspondiente a covid19, a continuación se muestra en la tabla 5 las palabras con mayor similitud de acuerdo al modelo entrenado.

Tabla 5: Palabras similares para covid19

Fuente: Elaboración Propia.

Hashtag	Distancia
covid_19	0.889
covid	0.884
covid2019	0.828
covid__19	0.825
coronagünlükleri	0.776
coronavirus	0.772
beatcoronavirus	0.771
inverclydecares	0.766
actfl	0.762
corvid_19	0.758

3.4. Implementación de algoritmo

Una vez que hemos recopilado todas las características, generamos un archivo CSV que contiene 2435827 filas. La primera columna de este archivo representa el *hashtag*, y las columnas siguientes contienen las características, separadas por punto y coma. Para este estudio, se decide eliminar los datos que aparecen solo una vez, es decir, los que tienen una frecuencia de 1. Esta decisión se basa en el hecho de que muchas de las otras características dependen de la frecuencia, lo que podría generar valores ruidosos. Esto resulta en un conjunto de datos para trabajar que contiene 800773 *hashtags* únicos.

Antes de aplicar algún algoritmo, es esencial comprender cómo se encuentran distribuidos los valores en un conjunto de datos y cómo los valores extremos pueden influir en los resultados. Para lograr esto, se utilizan herramientas como los percentiles y el rango intercuartílico (IQR).

Los percentiles permiten dividir el conjunto de datos en partes iguales, lo que ayuda a determinar qué valores se encuentran por debajo o encima de ciertos valores. Por ejemplo, el percentil 25, que corresponde de igual forma al primer cuartil, indica que el 25 % de los valores son iguales o menores que el cuartil 1 (Q1).

El IQR es una medida que se calcula a partir de la diferencia entre el tercer cuartil y el primer cuartil, esto da como resultado la representación de la dispersión de los valores en la mitad central de los datos, por ende, ayuda a determinar cuánto se extienden los datos en esta parte central.

La detección de valores atípicos o *outliers*, se basa en el IQR. Se utilizan dos límites:

- Inferior: restando 1.5 veces el IQR a Q1

- Superior: sumando 1.5 veces el IQR al Q3

Cualquier valor que se encuentra debajo del límite inferior o encima del límite superior, se considera un *outlier*. Esto se debe a que la mayoría de los *outliers* tiende a encontrarse significativamente alejados de la mayoría de los datos, lo que perjudica en los resultados del análisis.

En esta investigación, se aplica el procedimiento de detección de valores atípicos al conjunto de datos, para cada una de las columnas, priorizando la característica de la frecuencia.

Para este caso específico, se decide utilizar un subconjunto que comprenda únicamente los valores correspondientes a los valores atípicos. La razón detrás de esta elección radica en la naturaleza diversa de los *hashtags* que estamos analizando. Dado que existe una amplia variedad de *hashtags* en nuestro conjunto de datos, también observamos una gran dispersión entre ellos en términos de frecuencia y relevancia.

Es importante destacar que, en este contexto, los valores extremos o los valores atípicos se convierten en recursos valiosos para nuestro análisis. Estos valores atípicos representan los *hashtags* que destacan de manera excepcional en términos de uso o popularidad. Al enfocarnos en estos valores extremos, estamos dirigidos hacia los *hashtags* más relevantes e importantes de nuestro conjunto de datos. Esto nos permite centrar nuestros esfuerzos en comprender y analizar a fondo aquellos temas que captaron la atención y generan interacciones significativas en la comunidad, lo que resulta fundamental para nuestros objetivos de investigación. Por lo tanto, al utilizar este enfoque, podemos identificar de manera eficaz las tendencias más destacadas y las conversaciones de mayor impacto en nuestra área de estudio.

Se toma la decisión de considerar dos muestras para llevar a cabo el análisis tras realizar una serie de iteraciones en el conjunto de datos para detectar valores atípicos. Cada iteración generó un nuevo subconjunto que se analizó visualmente para entender cómo se distribuían los puntos de datos en el espacio. Esta elección se basa en el objetivo de realizar un agrupamiento y encontrar patrones basados en las características. Una de las muestras es relativamente pequeña en cantidad de datos, pero destaca debido a la relevancia de los *hashtags* que contiene en el conjunto de datos general. Estos *hashtags* se consideran puntos atípicos en el conjunto de datos general.

Inicialmente, se opta por trabajar con esta muestra más reducida debido a su capacidad para proporcionar un análisis detallado y exhaustivo de los puntos de datos que la componen. Al focalizar la atención en una muestra selecta y relevante, podemos identificar y comprender con mayor profundidad las tendencias y patrones emergentes y de esta forma encontrar categorías a partir de sus características y comportamiento. Estos puntos destacados, a pesar de su cantidad limitada, son esenciales para arrojar luz sobre aspectos cruciales del estudio en curso.

El plan es expandir posteriormente el análisis a un conjunto de datos más grande, lo que

nos permitirá evaluar la consistencia y robustez de los hallazgos obtenidos en la muestra pequeña. La segunda muestra recopila los mismos datos que la primera muestra, pero además incluye datos más cercanos a los valores típicos del conjunto general, aunque aún se consideran atípicos. Hemos elegido esta muestra porque creemos que contiene una mezcla de datos más generalizados, lo que nos permitirá realizar un análisis en profundidad y comparar los resultados con la primera muestra. A pesar de que no es tan grande como el conjunto de datos original, contiene más información y variedad. La mayor diversidad y dispersión entre los datos nos permitirá explorar detalles más específicos que podrían estar relacionados con una comunidad o características particulares, proporcionando una visión más completa del comportamiento de los datos.

Para ambas muestras, se procede a estandarizar los atributos de la muestra, esto ajusta cada elemento de la matriz para que tenga una media de cero y una desviación estándar de uno.

Con el objetivo de reducir la dimensionalidad de los atributos, se aplicarán dos métodos distintos: PCA y UMAP. Estos métodos se aplicarán por separado y también se realizará una combinación de ambos para analizar cómo influyen en los resultados y si son consistentes entre sí. Una vez obtenidos estos resultados, se procederá a aplicar el algoritmo de clustering K-Means.

3.4.1. Muestra 1

La muestra número 1 que se va a explorar consta de 986 datos y comprende 17 atributos numéricos, además se encuentran los vectores de palabras correspondientes para cada uno de estos puntos.

1. Caso PCA:

Al aplicar el Análisis de Componentes Principales (PCA) a los atributos numéricos, se logra que un total de 9 componentes principales sean capaces de explicar el 96 % de la varianza total. Esto significa que al utilizar estos 9 componentes, se captura la mayor parte de la variabilidad en los datos y que al retener sólo estos componentes principales, se puede reducir la dimensionalidad de los datos sin perder mucha información. Esto se descompone en:

- Componente 1: 39.36 %
- Componente 2: 23.15 %
- Componente 3: 9.82 %
- Componente 4: 7.55 %

- Componente 5: 5.12 %
- Componente 6: 4.58 %
- Componente 7: 4.03 %
- Componente 8: 3.54 %
- Componente 9: 2.85 %

En la figura 6 se puede observar la representación gráfica de los primeros dos componentes, los cuales consideran más del 50 % de la varianza total.

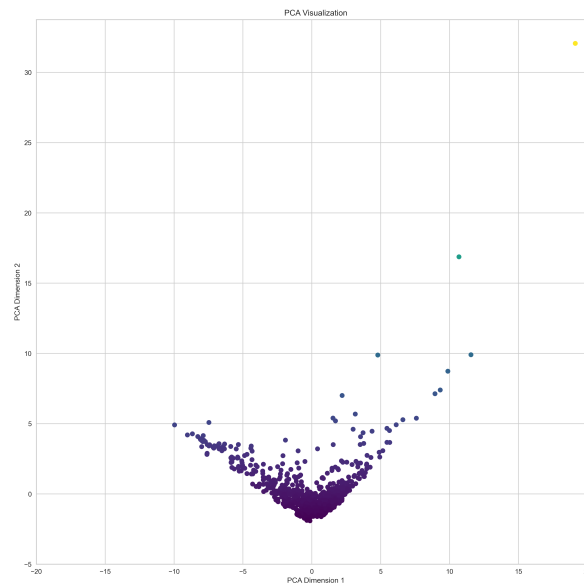


Figura 6: Representación componente 1 y componente 2
Fuente: Elaboración a partir de código

En el caso de los vectores de palabras, se encuentra que se requieren 33 componentes principales para explicar el 95 % de la varianza total. Estos análisis se realizan por separado en ambos conjuntos, ya que las características de los atributos numéricos y la vectorización de palabras están relacionadas entre sí. Una vez que ambos conjuntos se encuentren en un espacio de dimensionalidad reducida, se generará un vector en común que unirá ambos resultados. Este vector será utilizado en el proceso de clustering.

Antes de aplicar el algoritmo de *k-means*, es importante considerar que uno de sus hiperparámetros clave es el número de centroides o clusters deseados. Para determinar el número óptimo de clusters, existen varias heurísticas, cada una con sus propias sugerencias.

- El método del codo sugiere un número óptimo de 10 clusters

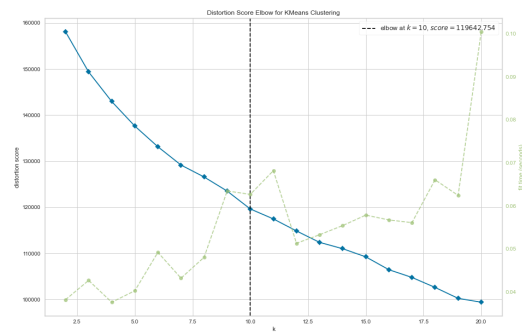


Figura 7: Método del codo
Fuente: Elaboración a partir de código

- el índice de Davies-Bouldin sugiere 17

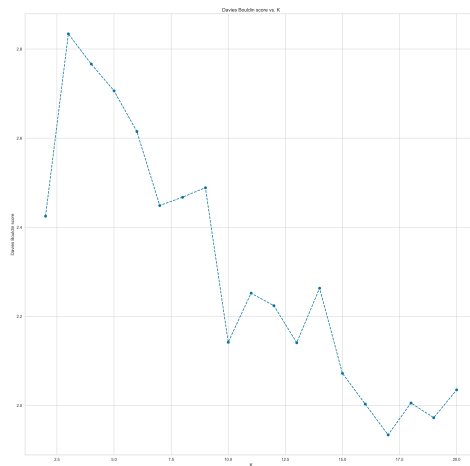


Figura 8: Índice de Davies-Bouldin
Fuente: Elaboración propia a partir de código

- el índice de Calinski-Harabasz sugiere 2

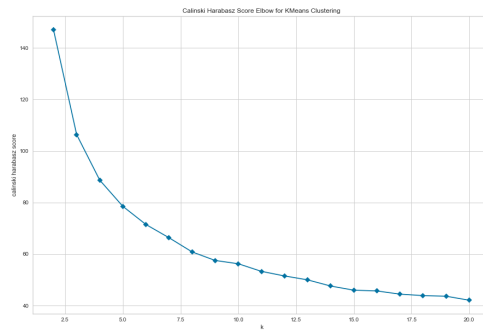


Figura 9: Índice de Calinski-Harabasz
Fuente: Elaboración a partir de código

- el estadístico de gap sugiere 20.

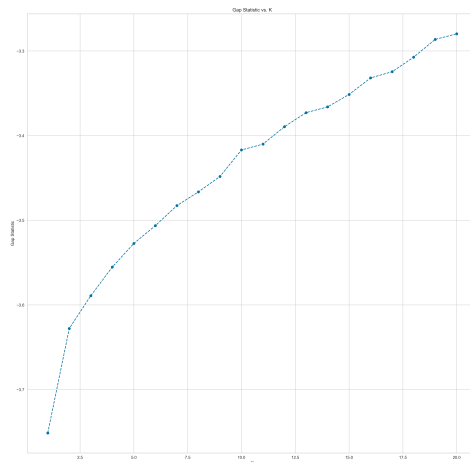


Figura 10: Estadístico de gap
Fuente: Elaboración a partir de código

Sin embargo, en este primer experimento se elegirá la cantidad sugerida por el índice de Davies-Bouldin, ya que con un número reducido de datos se espera una mejor comprensión del comportamiento de los datos con un mayor número de clusters.

Al aplicar el algoritmo de clustering con un k igual a 17, se obtiene la siguiente representación: (figura 11):

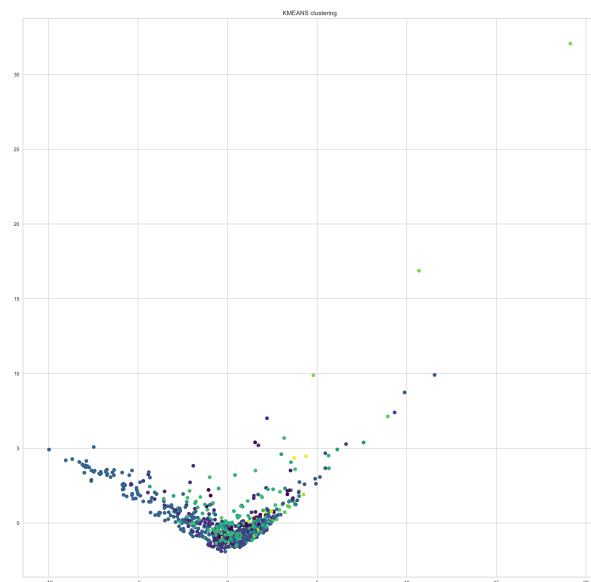


Figura 11: Resultado *k-means* con $k=17$, reducción de dimensiones por PCA
Fuente: Elaboración a partir de código

Cluster 1

En este cluster, se tiene un total de 41 elementos que se relacionan principalmente con días de la semana y localidades hispanohablantes, entre otros que se relacionan con la pandemia. Los *hashtags* más cercanos al centroide incluyen palabras como *hoy*, *domingo*, *felizmiércoles*, *caracas*, *felizviernes*, *felizmiercoles*, *pandemia*, *felizdomingo*, *argentina*, *méxico*.

Al analizar este cluster, se identifican varios valores clave con respecto a su significado en relación al cluster. Esto incluye la cantidad de *hashtags* que co-ocurren, la frecuencia de uso, la cantidad de co-ocurrencias, la última fecha de aparición y el rango de uso. Interesantemente, solo tres de estos *hashtags* tienen un rango de menos de un año y son exactamente los que se relacionan con la temática de la pandemia.

Además, se observa que los valores con la varianza más alta en este cluster son la cantidad de *hashtags* que co-ocurren, la frecuencia y la cantidad de co-ocurrencias. Esto sugiere que algunos de estos *hashtags* pueden estar altamente relacionados entre sí y se utilizan con frecuencia en conjunto. Todos los *hashtags* en este cluster están en español y algunos de ellos presentan variaciones con tildes y sin tildes, de hecho, dentro de los elementos más cercanos al centroide se puede notar que existen dos *hashtags* iguales solo que uno lleva tilde. Asimismo, estos *hashtags* son utilizados en más de 20 países, lo que podría indicar su amplia difusión en la comunidad hispanohablante y latinoamericana.

Curiosamente, se reconoce un *hashtag* en portugués que guarda similitud con los *hashtags* en español en este cluster. Esto podría deberse a la similitud lingüística entre el español y el

portugués, lo que lleva a una posible confusión en su categorización.

En cuanto a la temática de los *hashtags*, la mayoría de ellos se relacionan con días de la semana, lo que sugiere que están siendo utilizados para etiquetar publicaciones específicas según el día. Además, algunos *hashtags* señalan lugares, principalmente países latinos, con la excepción de España, pero que comparten la característica de ser hispanohablantes. Brasil, aunque no es hispanohablante, forma parte de Latinoamérica, lo que podría explicar su inclusión en este cluster.

Cluster 2

Compuesto por un total de 85 datos, y los *hashtags* más cercanos al centroide incluyen palabras como *boston*, *sandiego*, *legal*, *nashville*, *sanfrancisco*, *ad*, *securityguard*, *securityofficer*, *seattle* y *projectmgmt*. Al analizar este cluster, observamos que comparte características similares con el Cluster 1. Las características con la media más alta en este caso son la cantidad de *hashtags* que co-ocurren, la frecuencia de uso y la cantidad de co-ocurrencias. También destacan la última fecha de aparición y el rango, que son atributos distintivos de este clúster.

La varianza más alta se encuentra en la cantidad de *hashtags* con los que co-ocurren, la frecuencia y la cantidad de co-ocurrencias, lo que indica una mayor dispersión en estos atributos dentro de este cluster.

En términos de idioma, todos los *hashtags* en este cluster están en inglés. Se mencionan lugares en Estados Unidos, como *newyorkcity*, *atl*, *vegas*, así como varios lugares en general. También se incluyen lugares de Reino Unido y Canadá, como *liverpool* y *toronto*. En cuanto a temas de entretenimiento, se hacen menciones a la WWE y Star Wars. También se menciona *gameofthrones*, que fue el programa de televisión más comentado a nivel global, y su otra forma de referirse, *got*. Además, se hacen referencias a deportes como la NBA, NFL, *golf*, *superbowl*, *baseball* y *basketball*, así como términos relacionados con el trabajo, como *projectmgmt*, *businessmgmt*, *manufacturing*, *retail*, *sales*, *transportation*, *securityguard*, *nursing* y *hr*.

Este cluster abarca una amplia variedad de temas, desde lugares en Estados Unidos y otros países hasta entretenimiento, deportes y términos laborales. Las características clave se centran en la co-ocurrencia, frecuencia y co-ocurrencias de *hashtags*, lo que indica la diversidad de conversaciones que ocurren en torno a estos temas en las redes sociales.

Cluster 3

Este cluster consta de un total de 121 instancias y abarca una amplia gama de temas e intereses. Los *hashtags* más cercanos al centroide incluyen palabras como *perdu*, *mewlions*, *kellyservices*, *kuzgun*, *sanditon*, *saigai*, *hazalkaya*, *savesanditon*, *kellyjobs* y *quake*. Al exa-

minar este cluster, se destacan ciertas características clave: La frecuencia es el valor con la media más alta, lo que sugiere que los *hashtags* de este cluster reflejan temas populares en la plataforma. Además, la última fecha de aparición, el rango, la cantidad de *hashtags* que co-ocurren y la cantidad de co-ocurrencias también son características distintivas de este cluster. En cuanto a la varianza, se observa una dispersión significativa en estas mismas características, lo que sugiere una amplia diversidad en la forma en que se utilizan estos *hashtags*.

Se mencionan *hashtags* relevantes en eventos culturales y deportivos a nivel mundial, como *bbb20* (Big Brother Brasil 2020), *ucl* (UEFA Champions League) y *cwc19* (Cricket World Cup 2019), lo que indica que se trata de conversaciones sobre eventos ampliamente comentados a nivel global debido a que estas 3 instancias fueron de lo más comentado en el año 2020 y 2019, respectivamente.

También se incluyen *hashtags* relacionados con movimientos sociales, como *endsars*, *endsarsnow* y *endswat*, que se asocian con el movimiento contra la brutalidad policial en Nigeria, así como *endpolicebrutality*, que podría estar relacionado con un movimiento similar en Estados Unidos.

Se alude a ciertas temáticas correspondientes al continente asiático como ciertas bandas coreanas, series de Turquía e India. Esto refleja la diversidad de intereses culturales y de entretenimiento.

Cluster 4

Se encuentran 93 *hashtags* que abordan una amplia gama de temas relacionados con viajes y aventuras. Los *hashtags* más cercanos al centroide incluyen términos como *shotoniphone*, *roadtrip*, *view*, *nofilter*, *portugal*, *traveling*, *river*, *rain*, *travelling* y *ocean*.

Uno de los atributos más notables de este cluster es que la cantidad de *hashtags* que co-ocurren tiene la media más alta. Esto sugiere que los usuarios suelen utilizar múltiples *hashtags* en sus publicaciones para aumentar su visibilidad y para proporcionar contexto adicional sobre su experiencia de viaje o aventura. Los *hashtags* tienden a enfocarse en describir el entorno natural, las actividades y las emociones asociadas con el viaje.

En segundo lugar, aunque en menor medida, se encuentran la frecuencia de uso y la cantidad de co-ocurrencias. Esto indica que estos *hashtags* son utilizados con cierta regularidad y que tienden a aparecer en publicaciones relacionadas entre sí. Además, la última fecha de aparición, el rango de uso y la media mensual también son características importantes en este cluster, lo que sugiere que las conversaciones etiquetadas con estos *hashtags* son constantes en el tiempo, es decir, estos *hashtags* pueden ser utilizados en cualquier momento.

La varianza más alta se observa en la cantidad de *hashtags* con los que co-ocurren. Esto se explica por la naturaleza de los temas de viajes y aventuras, donde las personas suelen

utilizar una amplia variedad de etiquetas para describir detalladamente sus experiencias. Estos *hashtags* son versátiles y pueden utilizarse en diferentes contextos de viaje.

Dentro de este cluster, los *hashtags* reflejan la pasión por los viajes, la naturaleza y la aventura. Incluyen términos como *beach*, *explore*, *nature*, *landscape*, *summer*, *vacation*, *sun*, *sunrise*, y *sunset*, que capturan aspectos escénicos. También hay *hashtags* relacionados con la fotografía, como *shotoniphone*, *iphone*, *nikon*, *nofilter*, y *vsco*, que destacan las características de la fotografía.

En este cluster se puede encontrar una comunidad apasionada por compartir sus experiencias de viajes. Los *hashtags* utilizados permiten a los usuarios describir detalladamente sus momentos y lugares. Desde lugares como *#Portugal* hasta actividades como los *#roadtrips*, estos *hashtags* capturan la diversidad y la belleza del mundo.

Cluster 5

Este cluster está formado por un total de 56 *hashtags* que se centran principalmente en el ámbito de las fotografías, las redes sociales y la moda. Los *hashtags* más cercanos al centroide incluyen términos como *red*, *me*, *pic*, *mood*, *instapic*, *instaphoto*, *likesforlikes*, *photoshoot*, *pink* y *picture*. Al analizar este cluster, se destacan varias características clave: La varianza más alta se encuentra en la cantidad de *hashtags* que co-ocurren, lo que sugiere que estos *hashtags* suelen utilizarse en combinación con otros para aumentar la visibilidad de las publicaciones y generar un mayor alcance en las redes sociales. En segundo lugar, aunque en menor medida, se encuentra la frecuencia y la cantidad de co-ocurrencias. Esto indica que estos *hashtags* se utilizan con cierta regularidad y tienden a aparecer en publicaciones relacionadas entre sí.

En cuanto a las medias más altas, nuevamente la cantidad de *hashtags* que co-ocurren ocupa el primer lugar, seguido por la frecuencia y la cantidad de co-ocurrencias. Esto confirma la importancia de la combinación de *hashtags* para aumentar la visibilidad en las redes sociales.

Dentro de este cluster, se observa el uso de *hashtags* diseñados para mejorar la visibilidad y el alcance, como *like4likes*, *followme*, y *like*. Estos *hashtags* son estrategias comunes para atraer más atención a las publicaciones y obtener más interacciones en redes sociales. Además, se encuentran términos relacionados con la moda, como *shopping* y *model*.

Un detalle interesante es la presencia del *hashtag* *love*, que es uno de los más tuiteados después de los relacionados con la pandemia de COVID-19. Este *hashtag* es bastante genérico y se utiliza en una variedad de contextos y momentos. La mayoría de estos *hashtags* se utilizan desde el primer día del período definido y continúan siendo utilizados hasta el final, lo que demuestra su persistencia y popularidad. Esto incluye el *hashtag* *tbt* (*throwback Thursday*), que es ampliamente conocido y utilizado en las redes sociales para publicar recuerdos o fotos del pasado.

En pocas palabras, los *hashtags* aquí presentes son utilizados para aumentar la visibilidad de las publicaciones y generar interacción, incluso, a partir de los *hashtags* más cercanos al centroide se podría asumir. Algunos de estos *hashtags* son genéricos y se utilizan en una variedad de contextos, mientras que otros están más centrados en la descripción de situaciones específicas. En su conjunto, estos *hashtags* son herramientas comunes para aquellos que desean aumentar su presencia en las redes sociales y generar actividad en sus cuentas y publicaciones.

Cluster 6

Con un total de 181 *hashtags* que se relacionan con una variedad de eventos temporales y temas populares en redes sociales. Los *hashtags* más cercanos al centroide incluyen términos como *rocaafazenda*, *verzuz*, *lockdownhouseparty*, *roçafaazenda*, *thebachelor*, *drugfreeindia*, *thelastdance*, *gobackmodi*, *happybirthdayntr*, y *savegujanre*. Algunas características que destacan por su valor en la media son la frecuencia, que tiene un valor notablemente alto en comparación con el resto de atributos, le siguen en importancia la última fecha de aparición, la cantidad de *hashtags* que co-ocurren, la cantidad de co-ocurrencias y el rango de uso. También se incluye la característica de la primera fecha de aparición.

En términos de varianza, nuevamente la frecuencia ocupa el primer lugar, seguida de la cantidad de *hashtags* que co-ocurren y la cantidad de co-ocurrencias. Estos valores altos de varianza indican que estos *hashtags* varían en la frecuencia de uso y la interacción con otros *hashtags* en diferentes contextos y momentos. Dentro de este cluster, se destaca el *hashtag* *afazenda12*, que se refiere a un programa de televisión brasileño que fue el segundo programa más comentado en la televisión a nivel global en 2020. Además, se mencionan *hashtags* relacionados con eventos y temas populares en redes sociales, como *avengersendgame* (que generó un gran número de tweets en 2019), *eurovision*, *forabolsonaro*, *debates2020*, *loveisland*, *oscars*, *8m*, *apruebo*, *captainmarvel*, *chileenmarcha*, *elections2020*, *endgame*, *notredame*, y *copaamerica*.

También se incluyen *hashtags* relacionados con fechas y eventos temporales, como año nuevo (*#happynewyear2019*), halloween (*#halloween2020*), día de san valentín (*#happyvalentinesday*) y días de la semana, que se utilizan en el contexto de *goodmorning*, como por ejemplo *goodmorningfriday*, *goodmorningtuesday*.

Este cluster abarca una variedad de temas temporales y populares en las redes sociales. Incluye *hashtags* relacionados con programas de televisión, eventos deportivos, fechas especiales y eventos noticiosos. Estos *hashtags* son utilizados con alta frecuencia y varían según el momento y el contexto, lo que refleja la naturaleza cambiante y efímera de las conversaciones dentro de la plataforma y la relevancia de los temas populares en redes sociales en diferentes momentos.

Cluster 7

Un total de 28 *hashtags* que se centran principalmente en temas relacionados con la tecnología, la innovación y los empleos en el campo tecnológico. Los *hashtags* más cercanos al centroide incluyen términos como *startup*, *digital*, *digitalmarketing*, *technology*, *stem*, *jobs*, *innovation*, *bigdata*, *tech*, y *machinelearning*.

Las características con la media más alta en este cluster son la cantidad de *hashtags* que co-ocurren, la frecuencia de uso y la cantidad de co-ocurrencias. Estos valores sugieren que los *hashtags* en este cluster son utilizados de manera frecuente y en combinación con otros para describir y contextualizar conversaciones relacionadas con la tecnología. Además, se observan valores altos para el rango de uso, la última fecha de aparición y la media mensual, lo que indica que estos *hashtags* mantienen su relevancia a lo largo del tiempo y se utilizan consistentemente.

En cuanto a la varianza, la cantidad de *hashtags* que co-ocurren destaca claramente, lo que sugiere que estos *hashtags* varían en la forma en que se utilizan para agregar información o contexto a los tweets. La frecuencia y la cantidad de co-ocurrencias también presentan varianza, pero a un nivel similar.

Dentro de este cluster, se destacan conceptos relacionados con la tecnología, como *startup*, *digitalmarketing*, *technology*, *stem* (ciencia, tecnología, ingeniería y matemáticas), *jobs*, *innovation*, *bigdata*, *tech*, y *machinelearning*. Estos *hashtags* sugieren una fuerte conexión con el mundo de la tecnología y la innovación, así como con oportunidades laborales en este campo.

También se mencionan empresas de tecnología líderes como *apple*, *amazon*, *google*, y *facebook*, lo que indica un enfoque en las principales empresas tecnológicas y su impacto en la industria.

Además, se encuentran conceptos relacionados con la innovación y la computación, como *ai* (inteligencia artificial), *bitcoin*, *blockchain*, *iot* (Internet de las cosas), *100daysofcode*, y *machinelearning*. A grandes rasgos, este cluster se centra en temas de tecnología e innovación. Se generan comunidades interesadas en la tecnología, computación y ciencia, lo que genera una relevancia constante de estos *hashtags* a lo largo del tiempo.

Cluster 8

Este cluster consta de un total de 34 *hashtags* que se asocian principalmente en promocionar un estilo de vida saludable y positivo. Los *hashtags* más cercanos al centroide incluyen términos como *goals*, *power*, *passion*, *wellness*, *positivevibes*, *gratitude*, *healthy*, *yoga*, *lifestyle*, y *happiness*.

Los *hashtags* en este cluster están estrechamente relacionados con la salud, el bienestar, la motivación y la positividad. Se incluyen términos relacionados con el deporte, como *health*, *gym*, *fitness*, *yoga*, *sport*, *wellness*, y *natural*. Además, se mencionan conceptos relacionados con la felicidad, como *happy*, *goals*, *positivevibes*, y *gratitude*.

También se observan *hashtags* que reflejan una actitud positiva hacia la vida, como *jesus*, *god*, *success*, *quotes*, *inspiration*, *blessed*, y *goodvibes*. Estos *hashtags* pueden estar relacionados con la espiritualidad y la búsqueda de la felicidad y el éxito.

Se utilizan en publicaciones que buscan inspirar y motivar a otros a llevar un estilo de vida más saludable y feliz. Los *hashtags* reflejan la importancia de la salud física y mental, la búsqueda de objetivos y la gratitud por la vida.

Estos *hashtags* son recurrentes y suelen utilizarse en combinación con otros para contextualizar situaciones o proporcionar más información sobre el contenido de la publicación. Este cluster representa una comunidad creada a partir de un estilo de vida saludable donde las personas comparten sus experiencias y actividades.

Cluster 9

Se compone de 33 puntos que se refiere a música y entretenimiento audiovisual. Los *hashtags* más cercanos al centroide incluyen términos como *musician*, *pop*, *singer*, *trap*, *jazz*, *spotify*, *show*, *guitar*, y *studio*.

Los *hashtags* en este cluster consideran una amplia gama de estilos musicales, se incluyen términos que van desde *musician* y *pop* hasta *trap* y *jazz*, lo que refleja la diversidad de géneros y preferencias musicales de la comunidad.

Además de la música, este cluster también se relaciona con servicios de transmisión de música y videos, como *spotify*, *youtube* así como con la producción musical en *studio*. También se mencionan *hashtags* relacionados con el entretenimiento multimedia, como *tv*, *videos*, *musica*, *juegos*, y *podcasts*.

Algunos *hashtags* específicos, como *soundcloud*, *spotify*, y *twitch*, indican la importancia de estas plataformas en la vida de la comunidad. Estas plataformas pueden utilizarse para compartir música, transmitir en vivo y disfrutar de contenido multimedia en línea.

En conjunto, estos *hashtags* sugieren un enfoque en el entretenimiento audiovisual en general. Se podría interpretar como una comunidad interesada en la música, las plataformas de transmisión de música, los videos musicales y otros tipos de contenido multimedia, como programas de televisión y juegos.

Cluster 10

Conformado por un total de 22 *hashtags* que están directamente relacionados con fandoms de bandas coreanas y premiaciones en las que estas bandas suelen estar nominadas o presentes. Los *hashtags* más cercanos al centroide incluyen términos como *taehyung*, *boywithluv*, *jimin*, *jungkook*, *v*, *pcas*, *bbmas*, *blackpink*, *mamavote*, y *mgmavote*.

Los *hashtags* en este cluster están estrechamente relacionados con bandas coreanas populares, como BTS y Blackpink, así como con eventos de premiaciones a artistas. Los miembros de estas bandas, como *taehyung*, *jimin*, *jungkook*, y *v*, son mencionados con frecuencia en los *hashtags*.

Se observa la presencia de *hashtags* relacionados con premiaciones de música, como *pcas* (People's Choice Awards), *bbmas* (Billboard Music Awards), *mamavote* (Mnet Asian Music Awards Vote), y *mgmavote* (MGM Grand Awards Vote), *iheartawards*. Estos *hashtags* se utilizan para promover la votación de fanáticos en estos eventos y para mostrar apoyo a sus bandas favoritas.

Este cluster representa una comunidad unida por las bandas coreanas y su participación en eventos de premiación. Se refleja la importancia de las redes sociales, en particular Twitter, como una plataforma usualmente utilizada para fandoms de distintos artistas.

Cluster 11

Este cluster tiene 125 *hashtags* que aluden principalmente a los días de la semana, rutinas diarias y eventos especiales. Los *hashtags* más cercanos al centroide incluyen términos como *today*, *tuesday*, *tuesdayvibes*, *facts*, *wednesday*, *thankyou*, *thursday*, *sundayvibes*, *power*, y *weekndvibes*.

Los *hashtags* en este cluster se utilizan para contextualizar publicaciones en función del día de la semana o de eventos especiales. Esto incluye *hashtags* diarios muy utilizados en la comunidad, como *mondaymotivation*, *tuesdaythought*, *wednesdaywisdom*, *sundayfunday*, *saturday*, *throwbackthursday*, y *tgif* (Thank God It's Friday). También se mencionan días de la semana en general y *hashtags* relacionados con las dinámicas de Twitter, lo que sugiere que estos *hashtags* se utilizan para seguir las tendencias diarias y participar en conversaciones relacionadas con los temas del día.

Dentro de este cluster, se incluyen *hashtags* relacionados con celebraciones anuales, como *valentinesday*, *easter*, y *merrychristmas*, lo que indica la importancia de marcar eventos especiales y festivos en las redes sociales.

Se observa, igualmente, la presencia de *hashtags* relacionados con el COVID-19, como *quarantine*, *lockdown*, *stayathome* y *stayhome*, lo que sugiere que estos *hashtags* se utilizaron

en su momento para comentar sobre la rutina diaria durante la pandemia. Además, se mencionan *hashtags* generales utilizados en Twitter, como *rt* (*retweet*) y *retweet*, lo que se podría interpretar como una forma de promover el alcance de la publicación.

En términos generales, se refleja una serie de *hashtags* utilizados para marcar días de la semana junto a rutinas diarias, considerando que una de las características destacadas es los *hashtags* con los que co-ocurren, y teniendo en cuenta la aparición de *hashtags* relacionados a la pandemia, se puede deducir que se utilizaban en conjunto para expresar situaciones diarias durante el confinamiento.

Cluster 12

Este cluster representa principalmente eventos políticos y sociales de relevancia a través de los 69 *hashtags* que lo componen. Los *hashtags* más cercanos al centroide incluyen términos como *biden*, *bidenharris*, *joebiden*, *metoo*, *fakenews*, *trumpvirus*, *votehimout*, *gapol*, *moscowmitch*, y *buildthewall*.

La mayoría de los elementos cercanos al centroide hacen referencia a la situación política y las polémicas elecciones que ocurrieron en Estados Unidos entre Joe Biden y Donald Trump. Los *hashtags* relacionados con estas elecciones incluyen *biden*, *bidenharris*, *joebiden*, *impeachtrump*, y *votehimout*.

También se observan *hashtags* relacionados con eventos sociales y movimientos, como *metoo*, que se refiere al movimiento *#MeToo* contra el acoso sexual, y *blacklivesmatter* (BLM), que se relaciona con el movimiento por la justicia racial. Además, se mencionan los nombres de personas como *breonnataylor* y *georgefloyd*, que se han convertido en símbolos de la lucha contra la brutalidad policial y la discriminación racial.

El *hashtag* *brexit* hace referencia al proceso de salida del Reino Unido de la Unión Europea, que fue un tema de gran relevancia política en su momento. Adicionalmente, se mencionan *hashtags* relacionados con elecciones en otros países, como *auspol*, que se refiere a las elecciones presidenciales en Australia.

Estos *hashtags* son temporales y aparecen en momentos específicos en respuesta a eventos políticos o sociales. Por ejemplo, el *hashtag* *blm* ganó relevancia después de la trágica muerte de George Floyd en 2020, lo que desencadenó protestas en todo el mundo. Otros *hashtags*, como *auspol* y *trump2020* o *biden2020*, están relacionados con elecciones y aparecen durante el proceso electoral o campañas políticas.

Este cluster refleja el uso de Twitter como plataforma para discutir eventos políticos y sociales de actualidad, utilizando los *hashtags* para poder etiquetar y ser partícipe de las conversaciones del momento, de esta forma se aprecia la creación de *hashtags* ad hoc en respuesta a los eventos significativos en tiempo real.

Cluster 13

Este cluster se compone de un total de 59 *hashtags* que abordan temas relacionados con la creatividad, el arte y el entretenimiento. Los *hashtags* más cercanos al centroide incluyen términos como *cool*, *creative*, *blue*, *awesome*, *color*, *colors*, *baby*, *amazing*, *designer*, y *halloween*.

Una de las particularidades de todos los *hashtags* de este cluster es que se encuentran en inglés y que además, ninguno cuenta con dígitos. Se mencionan mascotas, incluyendo *perros* y *gatos*, así como animales en general. Los amantes de los animales utilizan *hashtags* como *dogs* y *cats* para compartir fotos y momentos especiales con sus mascotas.

El arte y la creatividad son temas prominentes en este cluster. Se incluyen *hashtags* relacionados con diferentes formas de expresión artística, como *art*, *film*, *painting*, *creative*, *hand-made*, *illustration*, *design*, *digitalart*, *sketch*, e *ink*. Estos *hashtags* pueden ser utilizados por artistas para compartir sus obras y también, descubrir nuevas creaciones.

Se mencionan *hashtags* relacionados con festividades y celebraciones, como *halloween*, *christmas*, y *birthday*. Estos *hashtags* se utilizan para marcar publicaciones relacionadas con eventos festivos y ocasiones especiales.

El entretenimiento también es una parte importante de este cluster. Se incluyen *hashtags* relacionados con películas, cómics, anime, y cosplay. Los aficionados a la cultura pop y el entretenimiento utilizan *hashtags* como *marvel*, *horror*, *comics*, *anime* y *cosplay* para participar en conversaciones y compartir su pasión por estos temas.

Cluster 14

El decimocuarto cluster, aunque pequeño en cantidad de elementos, es de gran relevancia debido a su enfoque en la pandemia de COVID-19 que marcó el año 2020 en todo el mundo. Este cluster está compuesto por solo tres *hashtags*: *coronavirus*, *covid19*, y *covid_19*. A pesar de su número limitado, estos *hashtags* representan uno de los temas más críticos y debatidos en las redes sociales durante el año 2020.

Los elementos cercanos al centroide, además de los tres que son parte, como *stayhome*, *lockdown*, *quedateencasa*, y *stayathome*, indican que existen otros *hashtags* que se relacionan con las medidas de prevención y confinamiento que se adoptó en todo el mundo para combatir la propagación del Covid, no obstante, los tres elementos están aún más estrechamente relacionados.

Covid19 y sus variantes, como *covid_19* fueron los *hashtags* más utilizados en Twitter en todo el mundo durante la pandemia, se estima que estos *hashtags* se twitearon alrededor de 400 millones de veces en 2020. Esto refleja la magnitud de la conversación masiva sobre

la pandemia y cómo las redes sociales se convirtieron en un medio crucial para compartir información, noticias, y experiencias relacionadas con el COVID-19. La inclusión de *coronavirus* en este cluster también es significativa, ya que este término se utilizó inicialmente para referirse al virus.

Cluster 15

Compuesto por 21 *hashtags*, se centra en el tema de la comida y abarca una variedad de elementos relacionados con la gastronomía y la experiencia culinaria. Los *hashtags* más cercanos al centroide incluyen términos como *dinner*, *lunch*, *delicious*, *pizza*, *breakfast*, *home-made*, *chocolate*, *foodphotography*, *restaurant*, y *yummy*.

Los *hashtags* en este cluster hacen referencia a una amplia diversidad de alimentos y comidas, como *pizza*, *chocolate*, *coffee*, *wine*. Algunos *hashtags* como *delicious* y *yummy* se utilizan para expresar satisfacción y placer al disfrutar de la comida.

Etiquetas de la comida como *homemade* y *restaurant* señalan experiencias distintas de cocinar en casa a comida en restaurantes, lo que podría ser utilizado junto a otros *hashtags* para encasillar el lugar de procedencia del alimento.

También, se mencionan momentos específicos del día en los que se disfruta de la comida, como *breakfast* (desayuno), *lunch* (almuerzo) y *dinner* (cena). Además *foodphotography* es un *hashtag* utilizado por aquellos que disfrutaban capturando visualmente sus platos y compartiendo imágenes de sus comidas en las redes sociales.

A modo resumen, este cluster refleja conceptos vinculados a la comida, lo que genera una comunidad que aprecia la gastronomía y experiencias relacionadas.

Cluster 16

Con una pequeña cantidad de elementos, se centra en el tema de bienes inmuebles y propiedades. Aunque hay solo ocho elementos en este cluster, cada uno de ellos está relacionado con aspectos específicos relacionados con bienes y propiedades. Se incluye terminos como *property*, *design*, *realstate* y *homes*, lo que hace referencia a las propiedades inmobiliarias.

El término *realstate* es un tema central en este cluster. Este *hashtag* se utiliza para hablar sobre el mercado inmobiliario, transacciones de propiedades y asuntos relacionados con la compra y venta. Se mencionan términos específicos relacionados con lugares de vivienda, como *apartment*, *flat* y *residential*. Estos *hashtags* se utilizan para describir el tipo de propiedades y los lugares de residencia.

Los *hashtags* aquí presentes se utilizan para discutir aspectos relacionados con la compra y

venta de bienes inmuebles, además de su tipo y arquitectura.

Cluster 17

Este cluster se caracteriza por su enfoque en cantantes y grupos musicales asiáticos, especialmente aquellos provenientes de Tailandia y Corea del Sur. Los *hashtags* cercanos al centroide incluyen términos relacionados con artistas y bandas asiáticas.

La presencia de *hashtags* relacionados con cantantes y grupos musicales asiáticos sugiere que este cluster está dedicado a la promoción y el apoyo de la música asiática en las redes sociales, lo cual ha ganado popularidad a nivel mundial en el último tiempo. Los fans de estos artistas utilizan estos *hashtags* para mostrar su entusiasmo y expresar su apoyo.

La mención de cantantes tailandeses en este cluster sugiere la popularidad de artistas de Tailandia en las redes sociales. Estos cantantes pueden tener una base de fans global que utiliza *hashtags* específicos para seguir y promocionar su música. La inclusión de *got7* hace referencia a la banda de K-pop GOT7, siendo estos grupos los que tienen una gran base de fanaticada en redes sociales.

Esto refleja la creciente influencia y popularidad de la música asiática en la escena global y cómo los fans utilizan las redes sociales para conectarse y expresar su entusiasmo.

2. Caso UMAP

Para poder aplicar el algoritmo de UMAP, se requiere asignar valores a los hiperparámetros requeridos como: la cantidad de vecinos para cada punto, que se recomienda entre el rango de 2 a 100, lo que genera un balance entre la estructura global y local; la mínima distancia entre los puntos, que señala que tan compacto serán los clusters y la métrica a utilizar. Considerando lo que representa cada parámetro, se realiza distintas iteraciones entre el rango de vecinos señalados para observar y decidir cuál será la cantidad adecuada para este conjunto de datos. Cuando el número es pequeño, no se logra visualizar cómo los puntos se conectan entre sí y quedan muchos de estos desconectados y dispersos en el espacio. A medida que aumenta el valor, la estructura global comienza a notarse y los datos se logran relacionar entre sí. Con un valor igual a 30, se obtiene una vista general bastante buena sin dejar de lado los detalles de la vista local. En el caso de la mínima distancia, mientras más cerca del 0, más aglomerados se encuentran los puntos lo que es útil para poder encontrar agrupaciones y por ende, para este caso. Bajo este contexto, los valores se definen de la siguiente forma: número de vecinos igual a 30, distancia mínima de 0.001 y métrica euclídeana, que es por defecto y la menos recursos computacionales requiere. Estos valores de hiperparámetros serán iguales para los experimentos posteriores que requieran del algoritmo de UMAP.

Para este caso, se realizará una reducción de dimensionalidad para los atributos numéricos y

los vectores de palabras, con un valor de 2 componentes para cada uno. Una vez reducidos, se fusionan en un vector común. En la figura 12 se puede observar cómo se comportan estas dos componentes:

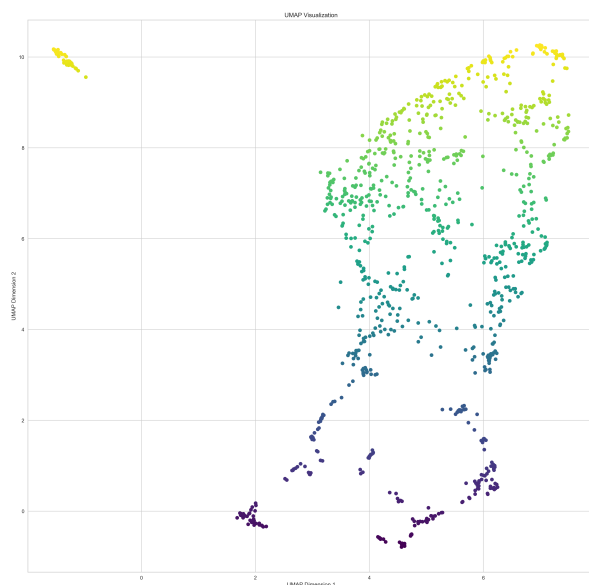


Figura 12: Dimensiones a partir de UMAP
Fuente: Elaboración a partir de código

Se consideran resultados a partir de las siguientes métricas para decidir el número de clusters a considerar:

- El método del codo sugiere un número óptimo de 7 clusters

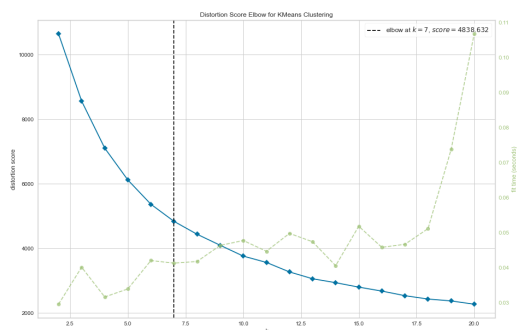


Figura 13: Método del codo
Fuente: Elaboración a partir de código

- el índice de Davies-Bouldin sugiere 3 clusters

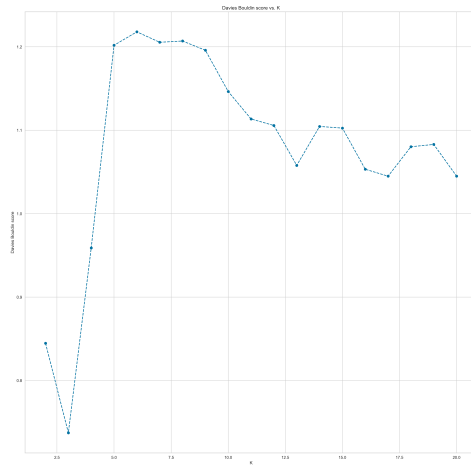


Figura 14: Índice de Davies-Bouldin
Fuente: Elaboración a partir de código

- el índice de Calinski-Harabasz sugiere 2 clusters

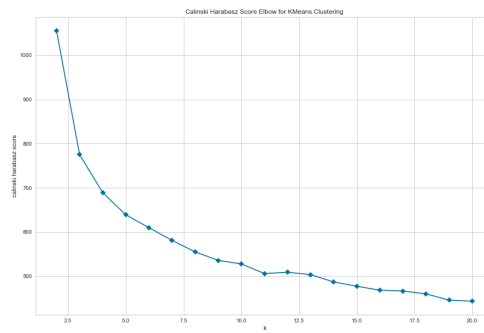


Figura 15: Índice de Calinski-Harabasz
Fuente: Elaboración a partir de código

- el estadístico de gap sugiere 7 clusters

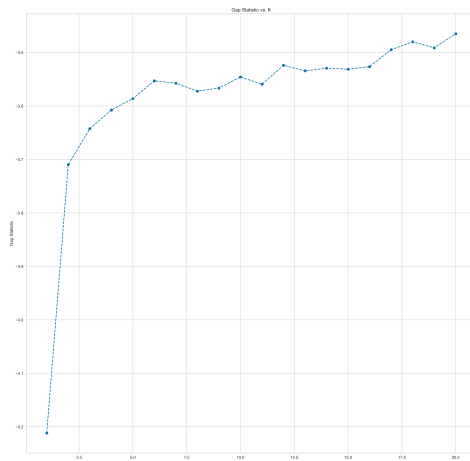


Figura 16: Estadístico de gap
Fuente: Elaboración a partir de código

Para este caso, se selecciona un óptimo de 7 clusters para aplicar el algoritmo K-Means. Esto a partir de lo señalado por la métrica del codo y del estadístico de GAP lo que se ve reflejado en la figura 38.

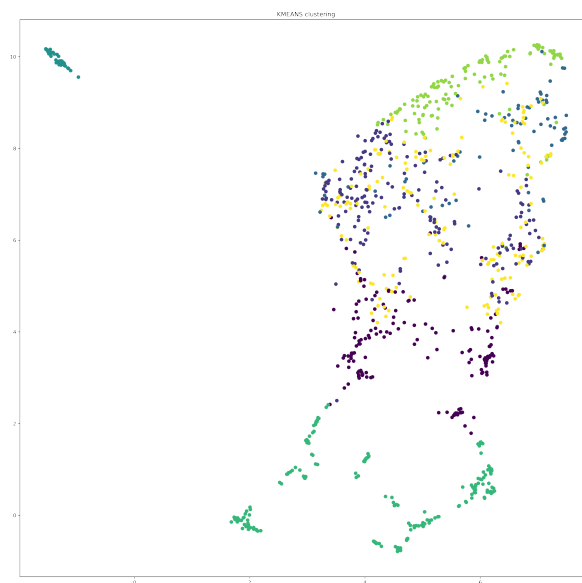


Figura 17: Resultado *k-means* con $k=7$, reducción de dimensiones por UMAP
Fuente: Elaboración a partir de código

Cluster 1

El primer cluster está compuesto por 162 *hashtags* que abarcan una variedad de temas, y se destacan algunas características clave en su conjunto: el atributo con la media más alta

en este cluster es la frecuencia, lo que indica que estos *hashtags* se utilizan con bastante regularidad en las redes sociales. Esto podría sugerir que los temas relacionados con estos *hashtags* son de interés continuo para la comunidad de Twitter. Los *hashtags* cercanos al centroide del cluster son *popezaap*, *redebbb*, *aldenrichards*, *mainemendoza*, *forabolsonaro*, *facciamorete*, *demetözdemir*, *nontfam*, *txhsfb*, *equiparacionya*.

Dentro de este cluster, se mencionan eventos y premios importantes, como *bbmas* (Billboard Music Awards), *kcamexico* (Kids' Choice Awards México), *premiosmtvmiaw* (Premios MTV Miaw), *amas* (American Music Awards) y un evento deportivo *copaamerica*. También se hacen referencias a BTS, la famosa banda de K-pop, incluyendo nombres de sus integrantes. Estos *hashtags* pueden estar relacionados con discusiones sobre los eventos en sí o el reconocimiento a artistas y celebridades.

En cuanto a la política, se mencionan *hashtags* relacionados como *biden*, *bidenharris* y *joebiden*, lo que sugiere discusiones sobre la política estadounidense y las elecciones más recientes. Además, se abordan temas de justicia social, ya que se mencionan *hashtags* relacionados con la brutalidad policial en Nigeria.

El cluster también incluye programas de televisión que se encuentran dentro de los más comentados, como *redebbb* y *loveisland*, lo que sugiere que parte de este cluster podría estar relacionado con las comunidades que utilizan Twitter como una especie de segunda pantalla para seguir el programa y la conversación que se genera en torno.

Finalmente, es importante destacar que los *hashtags* en este cluster tienen diferentes duraciones, algunos terminan al final del período definido, mientras que otros están relacionados con eventos específicos que ocurrieron en un momento determinado. Esto refleja la diversidad de temas que se abordan en este cluster y el interés constante de la que van desde el entretenimiento y la música hasta la política y los derechos humanos, lo que se puede observar en forma generalizada a partir de los elementos cercanos al centroide.

Cluster 2

Comprende un total de 188 elementos y parece estar relacionado con una variedad de temas y contenido genérico. Los *hashtags* cercanos al centroide de este cluster incluyen palabras clave como *iphone*, *awesome*, *black*, *men*, *girl*, *hair*, *bike*, *street*, *holidays* y *ootd* (Outfit Of The Day). Estos *hashtags* sugieren que las publicaciones en este grupo abarcan una amplia gama de temas y no están necesariamente centradas en un tema específico. Los atributos destacados en este cluster incluyen la cantidad de *hashtags* que co-ocurren, la cantidad de co-ocurrencias, la frecuencia, la máxima frecuencia y la última fecha. Esto indica que las publicaciones son variadas en contenido y pueden incluir una mezcla de *hashtags* relacionados con diferentes temas y eventos.

Se mencionan lugares como *city*, *garden*, *mountains* y *bar*. También se hacen menciones a

países como *Brazil*, *Spain* y *Russia*. Esto sugiere que las publicaciones pueden estar relacionadas con viajes y lugares visitados, lo que agrega un elemento de exploración y turismo al cluster.

Se observan referencias a fotografías, mascotas y animales, así como a comida. Esto indica que las publicaciones pueden incluir contenido visual, ya sea de mascotas, paisajes o platos de comida. Además, se mencionan deportes como *soccer* y *golf*. Algunos de los *hashtags* en este cluster son bastante genéricos y se utilizan junto con otros para proporcionar contexto o información adicional sobre la publicación. Esto refleja la versatilidad de las publicaciones en este grupo, que pueden abordar una amplia variedad de temas y momentos.

La inclusión de *hashtags* como *tattoos*, *street* y *shopping* propone una variedad de intereses, desde el arte corporal hasta la moda y la vida urbana. En conclusión, estos *hashtags* parecen ser atemporales y pueden utilizarse en cualquier momento. Pueden utilizarse para proporcionar contexto al combinarse entre sí y representar una variedad de temas.

Cluster 3

Este cluster consta de 82 *hashtags* y está asociado a temas de actualidad y eventos relevantes en el momento. Los *hashtags* cercanos al centroide de este cluster incluyen palabras clave como *mufc*, *cdnpoli*, *mumbai*, *foodwaste*, *hyderaba*, *breaking*, *unitedkingdom*, *pakistan*, *brexit* y *fakenews*.

El atributo más relevante en este cluster es la frecuencia, lo que significa que las publicaciones en este grupo se centran en temas que alcanzan una gran popularidad en el momento. La cantidad de *hashtags* con los que co-ocurren en las publicaciones es otro atributo importante, lo que indica que se tienden a utilizar múltiples *hashtags* en las publicaciones para ampliar el alcance y el contexto de sus mensajes. La cantidad de co-ocurrencias de estos *hashtags* también es significativa, lo que sugiere una alta interacción y conexión entre las palabras clave utilizadas en las publicaciones.

Se mencionan *hashtags* como *blacklivesmatter* y *blm* (un movimiento social que ganó relevancia tras el asesinato de George Floyd), y por el lado de la política, *brexit* (un proceso político relacionado con el Reino Unido) y *auspol* (proceso de elecciones en Australia).

Las temáticas en este cluster incluyen temas como *covid19* y *coronavirus*, lo que indica que las publicaciones están relacionadas con la pandemia de COVID-19, siendo estos *hashtags* los más utilizados en el 2020. Adicionalmente, se mencionan otros elementos asociados a la pandemia como *lockdown*, *stayhome*, *quarantine*. De igual forma, se observa la presencia de *hashtags* relacionados con lugares, lo que podría indicar que las personas utilizaban estas ubicaciones geográficas junto con otros *hashtags* para contextualizar y comentar sobre la pandemia de manera localizada.

En el marco de la incertidumbre y la gran propagación de información al respecto de la pan-

demia, la inclusión del *hashtag* *fakenews* también podría relacionarse con la difusión de noticias falsas y la forma en que las personas utilizaban este *hashtags* para etiquetar y señalar contenido no verificado o engañoso y de esa forma combatir la desinformación. Algunos *hashtags* relacionados a programas de televisión que se encuentran en este cluster son *gameofthrones* y *bbb20* (*Big Brother Brasil 2020*) los cuales fueron los más comentados a nivel mundial en el año 2019 y 2020, respectivamente.

Estos *hashtags* reflejan eventos y temas de actualidad relevantes en un momento dado. Es importante destacar que muchos de estos *hashtags* son generados de manera *ad hoc*, es decir, surgen en respuesta a eventos o situaciones específicas, como nuevos programas de televisión, resurgimientos de movimientos sociales o elecciones políticas.

Cluster 4

El cuarto cluster, compuesto por 40 *hashtags*, se centra en los días de la semana y expresiones asociadas a ellos. Estos *hashtags* se utilizan tanto en inglés como en español y son ampliamente reconocidos en las redes sociales. Los elementos cercanos al centroide son *thursdaythoughts*, *thursdaymotivation*, *mondaymotivation*, *wednesdaywisdom*, *wednesdaymotivation*, *sundayfunday*, *saturdaymorning*, *saturdaymotivation*, *saturdaythoughts*, *sundaymotivation*.

En este conjunto, encontramos *hashtags* que corresponden a cada día de la semana, como *thursdaythoughts*, *mondaymotivation*, *felizjueves*, y *felizmiercoles* e incluso, a partir de los *hashtags* más cercanos al centroide, se puede deducir este tipo de comportamiento. Además de los días de la semana, también se incluyen expresiones clásicas de Twitter, como *tgif* (*Thank God It's Friday*) y *throwbackthursday*. Estas expresiones son comunes en la plataforma y se utilizan para marcar el inicio del fin de semana o para compartir recuerdos del pasado.

Lo notable de este cluster es que estos *hashtags* son temporales y se utilizan principalmente en el día correspondiente de la semana. Por ejemplo, *thursdaythoughts* se utiliza para compartir pensamientos o reflexiones los jueves, mientras que *saturdaymorning* se relaciona con publicaciones relacionadas con la mañana del sábado.

Estos *hashtags* son altamente recurrentes y tienen una frecuencia que supera los 1000 usos para cada uno, lo que indica que son ampliamente utilizados. Sirven como una forma de marcar el tema de las publicaciones en la plataforma en esos días específicos de la semana.

Cluster 5

Este cluster contiene un total de 209 *hashtags* y está relacionado con una variedad de eventos y temas temporales. Los atributos más destacados en este cluster son la frecuencia, la

cantidad de *hashtags* que co-ocurren y la cantidad de co-ocurrencias.

Los *hashtags* incluidos en este cluster abarcan una amplia gama de eventos y temas. Algunos de los *hashtags* destacados son *oscars*, *cwc19* (la Copa del Mundo de Críquet 2019), *f1*, *bbb19* y *afazenda12* (programas de televisión brasileños), *grammys* y *fifawwc* (la Copa Mundial Femenina de la FIFA). Estos *hashtags* están relacionados con eventos deportivos y de entretenimiento que ocurren en momentos específicos y fueron de los más comentados a nivel global en el periodo.

También se mencionan *hashtags* relacionados con eventos trascendentales en la sociedad, como *georgefloyd*, *justiceforgeorgefloyd* y *breonnataylor*, que aluden a los asesinatos racistas ocurridos en Estados Unidos, y *endsars*, que se relaciona con protestas contra la brutalidad policial en Nigeria.

Se incluyen *hashtags* relacionados con el COVID-19, como *wearamask*, *quedateencasa*, *quarantine* y *socialdistancing*, lo que indica que las conversaciones en este cluster también abordan la pandemia y las medidas relacionadas.

Además, se observan *hashtags* relacionados con fechas específicas de celebraciones, como *8m*, *valentinesday*, *christmas2020* y *halloween2020*. Esto sugiere que se comparten publicaciones relacionadas con festividades y eventos con fechas específicas.

El cluster muestra una mezcla de *hashtags* en diferentes idiomas, lo que sugiere que las publicaciones pueden provenir de personas que hablan distintas lenguas y están interesadas en eventos y temas globales.

En general, este cluster se caracteriza por temas que se discuten con regularidad en un periodo de tiempo específico. Los *hashtags* utilizados en estas publicaciones proporcionan contexto y describen las conversaciones en torno a estos eventos y temas.

Cluster 6

Este cluster contiene un total de 134 instancias. Los *hashtags* que están más cerca del centroide de este cluster incluyen palabras como *cute*, *lifestyle*, *beauty*, *me*, *selfie*, *amazing*, *architecture*, *blessed*, *model*, y *home*. Estos *hashtags* sugieren que este cluster se relaciona con temas de estilo de vida, belleza, autorretratos y arquitectura, posiblemente compartiendo fotos personales o de lugares interesantes.

Uno de los atributos más destacados en este cluster es la cantidad de *hashtags* con los que co-ocurren en las publicaciones. Esto indica que las personas en este cluster tienden a usar múltiples *hashtags* para describir sus fotos o estados. Además, la frecuencia y la cantidad de co-ocurrencias de estos *hashtags* son también características importantes.

En el cluster se puede encontrar temas como fotografía, viajes, vacaciones, ejercicio y vida sa-

ludable. Los *hashtags* relacionados con lugares y actividades sugieren que las publicaciones pueden estar vinculadas a experiencias de viaje, ya que se mencionan *hashtags* relacionados con vacaciones y lugares. Además, los *hashtags* relacionados con el ejercicio y la vida saludable indican un interés en el bienestar físico.

Las publicaciones en este cluster parecen ser genéricas y versátiles, ya que pueden describir una amplia variedad de contenidos, como fotografías, estados de ánimo, actividades, situaciones, etc.

En pocas palabras, este cluster se caracteriza por publicaciones genéricas que abarcan una amplia gama de temas, desde fotografía y viajes hasta ejercicios y música. Ayudan a describir las actividades y experiencias de la comunidad en aspectos de su tiempo libre y estilo de vida.

Cluster 7

Compuesto por 171 *hashtags*, abarca una amplia variedad de temas. Aunque no se puede identificar una categoría específica que englobe a todas las instancias, se pueden destacar varios temas y características comunes en este cluster. Los *hashtags* más cercanos al centro de corresponden a *win*, *share*, *tech*, *movie*, *lasvegas*, *online*, *show*, *amazon*, *digital*, *house*.

Uno de los temas centrales es la tecnología y las empresas de tecnología, como se evidencia con *hashtags* como *tech*, *digital*, *innovation*, *ai*, *bitcoin*, *machine learning*, y *bigdata*.

Además de la tecnología, también se mencionan lugares en los Estados Unidos y Canadá, como *lasvegas*, *atlanta*, *dallas*, *chicago*, y *florida*. Estos *hashtags* podrían estar relacionados con eventos o lugares específicos en estas regiones.

El entretenimiento también es un tema presente en este cluster, con *hashtags* como *movie*, *gaming*, *book*, y menciones de plataformas de música y videos.

En cuanto a bienes inmuebles, se hacen referencias a *apartment*, *flat* y *realstate*, lo que indica un posible interés en propiedades.

En general, este cluster parece estar compuesto por *hashtags* genéricos que no se centran en eventos específicos, sino que se utilizan de manera recurrente para agregar contexto y más información a las publicaciones. Refleja una diversidad de intereses que abarcan desde la tecnología y el entretenimiento hasta propiedades, lo que lo convierte en un conjunto versátil utilizado por diversas comunidades.

3. Caso PCA+UMAP

Para poder seleccionar los componentes para utilizar el algoritmo UMAP para los atributos numéricos y el vector de palabras, se debe utilizar los resultados obtenidos por el PCA aplicado al conjunto de datos en 3.4.1. Según esos resultados, los atributos numéricos se explican con sólo 9 componentes de 17, lo que rescata un 97 % de la varianza total. Se analizará la relevancia de las características por cada uno de los componentes para así poder tomar una decisión para poder aplicar UMAP.

El primer componente explica el 39,36 % de la varianza. El atributo que más destaca es el coeficiente de variación diario. En un nivel bastante cercano entre sí, le sigue el coeficiente de variación mensual, máxima frecuencia, rango, *hashtags* que co-ocurren, cantidad de co-ocurrencias y cantidad de países, lo que se observa en la figura 18. La pendiente mensual es la característica con el menor valor.

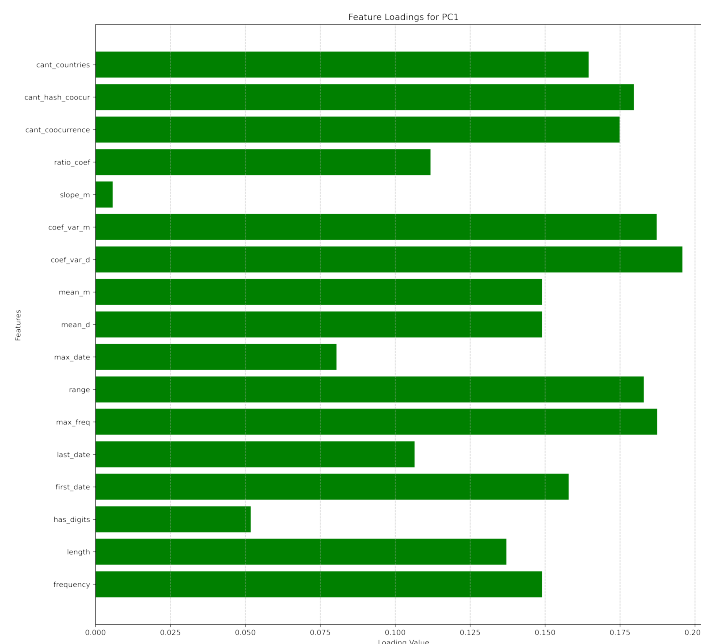


Figura 18: Aportes de cada atributo en el componente 1

Fuente: Elaboración a partir de código

El segundo componente principal tiene un valor de 23,15 % y que en conjunto del primer componente, suman el 62 % de la varianza total. Se observa en la figura 19 que varios de los atributos que en el componente número 1 se consideraban destacados, comienzan a bajar su magnitud así como también el largo de la cadena y la cantidad de *hashtags* que co-ocurren, disminuyen a la mitad. Por otro lado, el caso de la pendiente mensual que comienza a tomar fuerza.

El tercer componente solo explica un 9.82 % de la varianza total. En este caso, atributos que

tenían una gran relevancia en el primer componente, ahora poseen un porcentaje bastante bajo en comparación, y viceversa lo que se observa claramente en la figura 20.

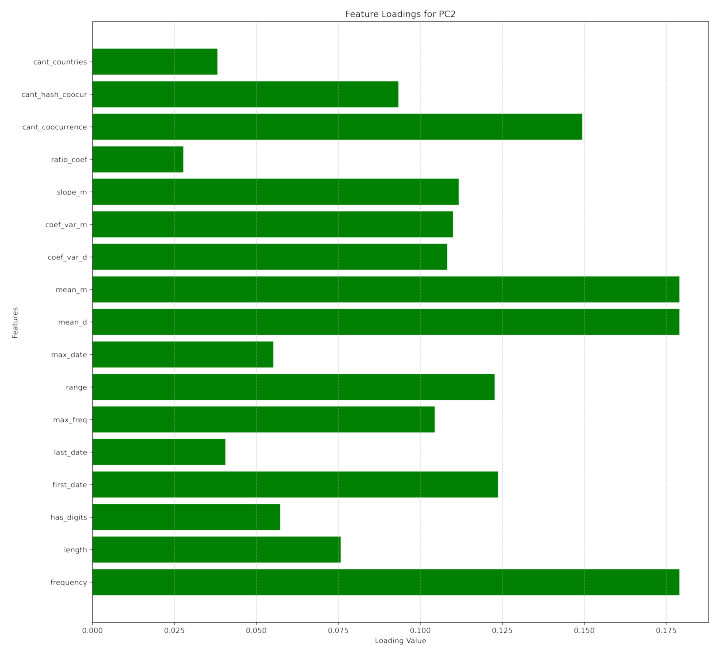


Figura 19: Aportes de cada atributo en el componente 2
Fuente: Elaboración a partir de código

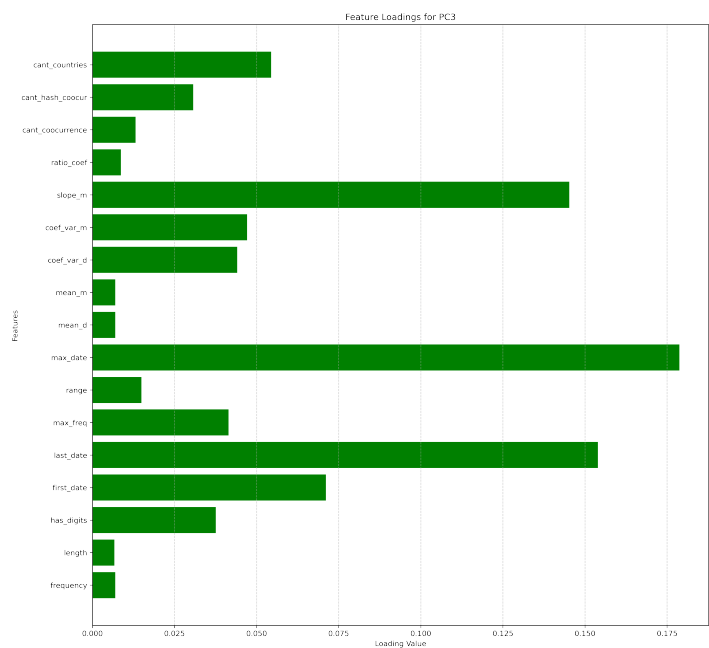


Figura 20: Aportes de cada atributo en el componente 3
Fuente: Elaboración a partir de código

A partir de este análisis se decide trabajar con solo 2 componentes para aplicar UMAP una vez hecho el análisis de componentes principales. Los gráficos con el aporte de cada atributo para las componentes restantes, se pueden encontrar en el Anexo 5.1.

Posterior a aplicar el algoritmo de UMAP sobre el conjunto de datos de dimensión ya reducida por el algoritmo de PCA, se obtiene la siguiente representación (figura 21):

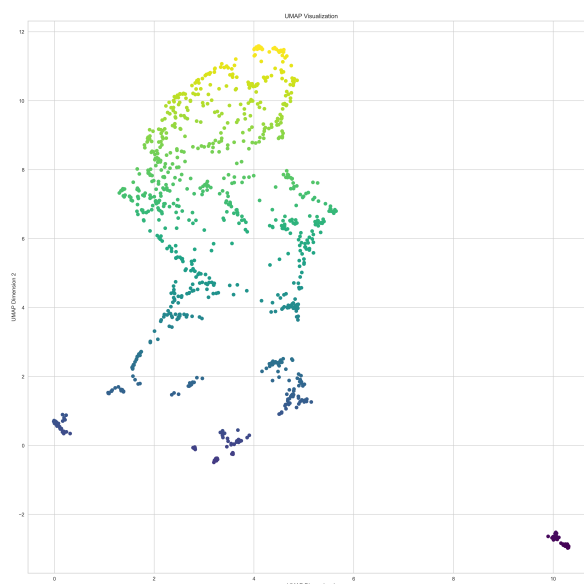


Figura 21: Conjunto de dimensiones reducidas a partir de PCA y UMAP
Fuente: Elaboración a partir de código

Respecto al vector de palabras, se obtiene que este logra explicar el 95 % de la varianza total a través de 43 componentes. Se debe recordar que este vector tiene una dimensión de 100. Para simplificar este caso, se considerarán sólo las componentes que comprendan un valor mayor o igual al 5 %. Como resultado, se obtiene que se trabajará con 6 componentes para el caso del vector de palabras.

Antes de aplicar el algoritmo de K-Means al conjunto de datos con dimensión reducida, se comparan los resultados de distintas métricas:

- El método del codo indica un óptimo de 8 clusters

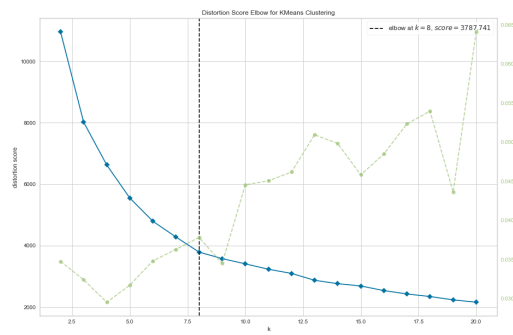


Figura 22: Método del codo
Fuente: Elaboración a partir de código

- el índice de Davies-Bouldin sugiere 3 clusters

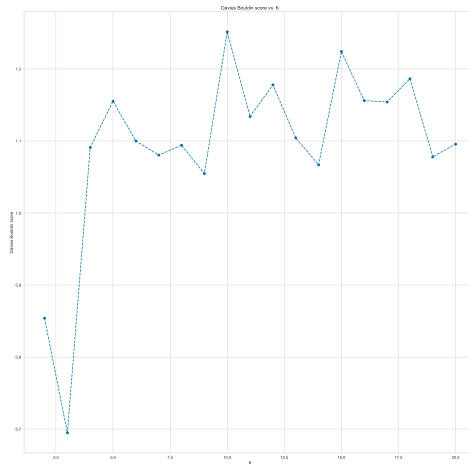


Figura 23: Índice de Davies-Bouldin
Fuente: Elaboración a partir de código

- el índice de Calinski-Harabasz señala como mejor valor 2 clusters

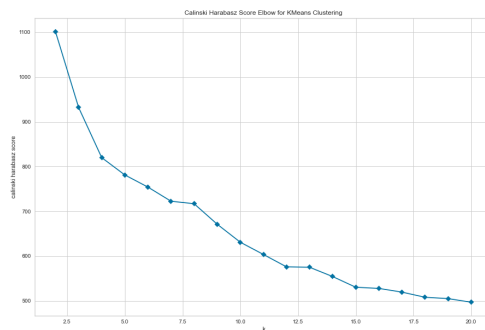


Figura 24: Índice de Calinski-Harabasz
Fuente: Elaboración a partir de código

- el estadístico de gap sugiere 9 clusters.

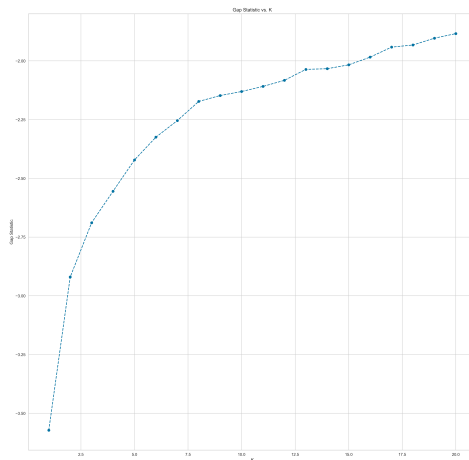


Figura 25: Estadístico de gap
Fuente: Elaboración a partir de código

Se selecciona el valor resultante por el estadístico de GAP, esto para aprovechar el tamaño del conjunto y de esa forma, poder observar de mejor forma las categorías, lo que es representado en la figura 26.

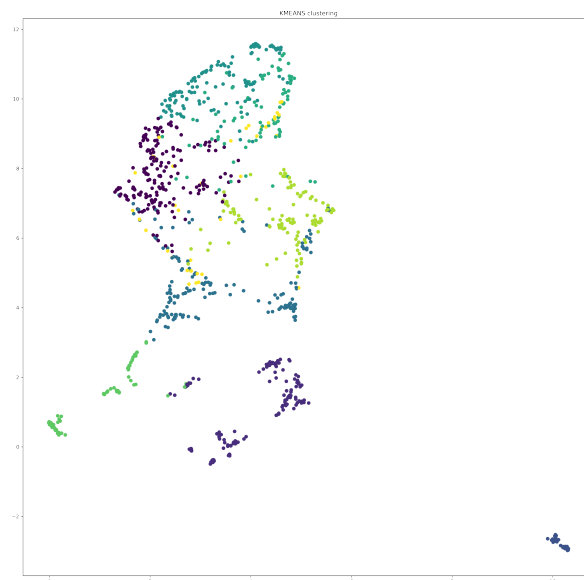


Figura 26: Resultado *k-means* con $k=9$, reducción de dimensiones por PCA y UMAP
Fuente: Elaboración a partir de código

Cluster 1

Con un total de 180 datos, abarca una amplia variedad de temas relacionados con el estilo de vida, el entretenimiento y los pasatiempos. Los *hashtags* cercanos al centroide de este cluster incluyen palabras clave como *loveislove*, *vintage*, *pink*, *funny*, *kids*, *event*, *morning*, *trans*, *golf* y *awesome*.

Se mencionan lugares como *Brazil*, *Portugal*, *Dallas*, *Disney* y *Dubai*, lo que sugiere que las publicaciones pueden incluir referencias a viajes y destinos turísticos. Además, se hacen menciones a deportes como *golf* y *soccer*, así como conceptos relacionados con otros deportes como *NBA* (baloncesto), *boxing* y *sports* de forma genérica.

El cluster también incluye *hashtags* relacionados con el tiempo libre y el estilo de vida, como *familytime* y *weekendvibes*. Esto indica que las publicaciones pueden centrarse en actividades recreativas en tiempo libre.

Conceptos relacionados con la comida diaria, como *breakfast*, *dinner* y *lunch*, así como términos de comida como *pizza*, *restaurant*, *chocolate*, *cafe*, *delicious* y *beer*, sugieren que el cluster también pueden incluir contenido relacionado con la gastronomía y las experiencias culinarias.

Los *hashtags* relacionados con viajes, tiempo libre, estilo de vida y entretenimiento como *goodvibes*, *cycling*, *painting* y *blogger* indican que se puede asociar a vida cotidiana e intereses personales. La inclusión de *hashtags* relacionados con mascotas sugiere que las publica-

ciones pueden incluir contenido sobre animales de compañía y su importancia en la vida de las personas.

Aunque no se define claramente una categoría específica, este cluster parece abordar temas y conceptos que pueden ser atemporales y versátiles, lo que permite adaptarse a diferentes contextos y preferencias de los usuarios.

Cluster 2

Consta de 145 elementos y abarca temas de actualidad. Los *hashtags* cercanos al centroide, como *savegujaratnre*, *atmanirbharbharat*, *lrd_male*, *lockdownhouseparty*, *excludeduk*, *farmersprotest*, *fridaylivestream*, *ntrbirthdaycdp*, *verzuz* y *happybirthdayntr*, reflejan la diversidad de conversaciones.

El atributo más destacado en este cluster es la frecuencia, con un valor notablemente alto de 1176. Esto sugiere que las publicaciones en este grupo están relacionadas con eventos o temas que se discuten con gran intensidad y frecuencia en un momento específico. Se observan *hashtags* relacionados con los días de la semana, como *goodmorningmonday*, *goodmorningtuesday*, *goodmorningwednesday*, *goodmorningthursday*, *goodmorningfriday* y *goodmorningsaturday*. Aunque no son de actualidad, siguen un patrón temporal respecto a los días de la semana, por lo que no son aleatorios en el tiempo.

Algunos *hashtags* en este cluster están relacionados con eventos de actualidad, como *endsars*, que se refiere a las protestas contra la brutalidad policial en Nigeria, y *breonnataylor*, *georgefloyd* y *justiceforgeorgefloyd*, que hacen referencia a los nombres de personas afro-americanas asesinadas por la policía en Estados Unidos. También se mencionan eventos como *eurovision*, *oscar*s y *superbowl*, lo que indica un interés en eventos de entretenimiento a nivel internacional.

El *hashtag* *notredame* fue ampliamente utilizado en relación con el incendio de la catedral de Notre Dame en París, lo que lo convierte en un tema de noticias de alto perfil. *Demdebate* también se destacó como un *hashtag* relacionado con noticias en 2019. En el contexto de la pandemia de COVID-19, se mencionan *hashtags* como *wearamask*, *quarantinelif*e, *cuarentena*, *socialdistancing*, *stayathome* y *stayhomestaysafe*, lo que sugiere una discusión sobre las medidas en la pandemia.

Otros *hashtags* son más locales y están relacionados con temas como películas, actores, movimientos sociales y conflictos regionales, como *prayforganja* y *stoparmenianterrorism*, que se utilizan en conjunto para visibilizar un movimiento social.

En este grupo los temas están vinculados a eventos y conversaciones en tiempo real, y los *hashtags* utilizados reflejan la diversidad de intereses alrededor del mundo.

Cluster 3

Este cluster consta de 40 *hashtags* y está enfocado con días de la semana. Las instancias cercanas al centroide incluyen *hashtags* como *wednesdaymotivation*, *mondaymotivation*, *thursdaymotivation*, *fridaymotivation*, *thursdaythoughts*, *fridaythoughts*, *fridayfeeling*, *mondaythoughts*, *sundaythoughts* y *wednesdaythoughts*.

Lo más destacado en este cluster son las características relacionadas con la cantidad de *hashtags* que co-ocurren, la frecuencia y la cantidad de co-ocurrencias. Estos *hashtags* son recurrentes y adaptables, lo que significa que los usuarios los utilizan en combinación con otros tipos de *hashtags* para dar contexto a su contenido y aumentar su visibilidad en las redes sociales.

Los *hashtags* representan los días de la semana en español e inglés, así como dos de ellos en turco. Esto refleja la diversidad lingüística de los usuarios y su enfoque en diferentes culturas.

Algunos de los *hashtags* en este cluster son conocidos como *day-of-the-week recurring Twitter hashtags* como *wednesdaywisdom*, *throwbackthursday*, *fridayfeeling* y *sundayfunday*. Estos *hashtags* son conocidos por la comunidad de Twitter y se utilizan de manera recurrente según el día correspondiente de la semana. Son flexibles y se adaptan al contenido que se desea publicar, lo que los convierte en parte de la cultura de la plataforma.

Cluster 4

Contiene un total de 159 instancias y las instancias cercanas al centroide incluyen *hashtags* como *popezaap*, *aldenrichards*, *mainemendoza*, *brandenburg*, *redebbb*, *parthsamthaan*, *demetözdemir*, *kellyservices*, *kellyjobs* y *kasautiizindagiikay*.

Las características con valores más altos de la media en este cluster son la frecuencia, la cantidad de *hashtags* que co-ocurren, la última fecha, el rango y la cantidad de co-ocurrencias. Esto sugiere que las publicaciones en este grupo son activas y se centran en temas y eventos que están en el centro de atención en un momento específico.

El cluster abarca temas relacionados con la música, como nombres de integrantes de BTS y referencias al K-pop. También se menciona *avengersendgame*, que fue la película más tuiteada en la historia y está relacionada con el éxito de Avengers: Endgame. Además, se hace referencia a *ucl* en el ámbito deportivo, siendo este el evento deportivo más discutido en el 2019.

Se observan *hashtags* relacionados con el movimiento *MeToo*, que es un movimiento social que busca concienciar sobre el acoso sexual y el abuso de poder. También se mencionan *hashtags* relacionados con la política estadounidense, como *impeachtrump*, que se refiere a los procedimientos de juicio político contra el expresidente Donald Trump, así como *hashtags*

relacionados con el gobierno actual, como *joebiden* y *bidenharris*.

En este cluster, se encuentran nombres de actores y actrices asiáticos, lo que puede sugerir un interés en la industria del entretenimiento asiático.

Los *hashtags* *kellyservices* y *kellyjobs* hacen referencia a una empresa estadounidense que ofrece servicios de reclutamiento y búsqueda de empleo.

Se mencionan ceremonias de premiación relacionadas con la música, como *bbmas* (*Billboard Music Awards*) y *premiosmtvmiaw* (*Premios MTV MIAW*).

En resumen, el cluster aborda temas relacionados con la cultura pop, la política estadounidense, el entretenimiento y la música. La mayoría de estos *hashtags* están relacionados con temas temporales y eventos específicos que estaban en tendencia en ese momento.

Cluster 5

Este cluster comprende un total de 133 datos y está relacionado con un estilo de vida diverso y activo. Los *hashtags* cercanos al centroide de este cluster incluyen palabras clave como *lifestyle*, *blessed*, *smile*, *goodmorning*, *friends*, *selfie*, *fun*, *style*, *happybirthday* y *beauty*.

Lo que destaca en este cluster es la cantidad de *hashtags* que co-ocurren y la frecuencia con la que se utilizan en conjunto. Estos *hashtags* están relacionados con actividades y elementos de un estilo de vida activo y variado.

Los *hashtags* en este cluster están asociados a varios aspectos del estilo de vida, como pasatiempos, viajes, lugares, escapadas y la compañía de amigos. También se mencionan distintos deportes y actividades físicas, como *football* y *gym*.

En general, los *hashtags* en este cluster representan una variedad de estilos de vida que incluyen actividades, comidas, tipos de comida, entrenamientos, deportes y entretenimiento, como música y videos. Algunos ejemplos de estos *hashtags* son *summer*, *food*, *dance* y *beach*.

Se mencionan varios lugares, localidades y países, lo que sugiere que las publicaciones pueden incluir referencias a destinos turísticos y experiencias de viaje. Además, se hacen referencias a tipos de música y actividades de pasatiempo, como la fotografía. Este cluster refleja la diversidad de intereses y actividades en la vida de las personas, así como las emociones relacionadas con estos momentos.

Los *hashtags* en este cluster son atemporales y genéricos, ya que se utilizan en diferentes momentos y contextos elegidos por los usuarios. A menudo, se combinan con otros *hashtags* para formar un contexto más completo y compartido con comunidades que comparten intereses similares.

Cluster 6

Con los 85 *hashtags* que comprenden este cluster, se enfoca en temas de actualidad y eventos significativos que ocurrieron durante el período estudiado. Los elementos cercanos al centroide incluyen *hashtags* como *pakistan*, *hyderabad*, *unitedkingdom*, *giletsjaunes*, *food-waste*, *brexit*, *mumbai*, *auspol*, *cdnpoli* y *srilanka*.

Lo que resalta en este cluster es la alta frecuencia de uso de los *hashtags*, lo que indica que estos temas eran discutidos con regularidad durante el período analizado. Además, destaca la cantidad de *hashtags* que co-ocurren y la cantidad de co-ocurrencias, lo que sugiere que estos temas estaban interconectados.

En este cluster se encuentra el *hashtag* con la mayor frecuencia que corresponde a *covid19*, siendo este y sus variantes los más utilizados en el 2020 como *coronavirus*, *covid_19*, *covid*, entre otros y también, algunos ligados como *staysafe*, *lockdown* y *stayhome*, siendo este el tercero más comentado en el 2020.

Se hacen referencias a varios países, como *venezuela*, *argentina*, *chile* y *china*, lo que sugiere discusiones sobre eventos y situaciones específicas en estos lugares o que puedan ser utilizados junto a otros *hashtags* para localizar un tema.

El *hashtag* *gameofthrones* es mencionado, el cual fue la serie más comentada a nivel mundial dentro del periodo.

Los *hashtags* *blacklivesmatter* y *blm* resurgen en este cluster debido a las protestas y movimientos sociales relacionados con los asesinatos raciales a manos de la policía en Estados Unidos.

Otros temas políticos, como *brexit*, *auspol* y *cdnpoli*, están relacionados con procesos políticos en el Reino Unido, Australia y Canadá, respectivamente.

El *hashtag* *endsars* se destaca como un tema principal en este cluster, relacionado con las protestas en Nigeria contra la brutalidad policial y el movimiento para poner fin a la unidad especial de la policía nigeriana conocida como SARS.

Además, se mencionan *hashtags* como *bbb20* (*Big Brother Brasil 2020*), siendo el programa de televisión más discutido en el 2020 y *climatechange* y *hongkong*, que se encuentran entre los *hashtags* más utilizados para comentar noticias en todo el mundo.

Este cluster representa una recopilación de temas de actualidad y muchos se generan de manera repentina en respuesta a los eventos.

Cluster 7

El cluster consta de 84 puntos de datos que se centran en eventos específicos, tanto en el ámbito televisivo como en el deportivo y el político. Los *hashtags* cercanos al centroide incluyen términos como *got7realthai*, *massgodntrbdayin100days*, *sanremo2019*, *bbb19*, *quebec_immigration_injustice*, *loksabhaelections2019*, *afcon2019*, *teamgot7*, *trndnl* y *nigeria-decides2019*. Algunos de los temas más destacados incluyen:

- Eventos televisivos, como los premios *Oscars2019* y *Grammys*, que son ampliamente comentados debido a su importancia en la industria del entretenimiento.
- Elecciones presidenciales en Estados Unidos, con *hashtags* como *biden2020* y *trump2020*, que representan a los principales candidatos de los partidos demócratas y republicanos, respectivamente. También se hacen referencias a elecciones en India y Nigeria.
- Programas de televisión brasileños populares, como *Afazenda12* y *BBB19*, que alcanzan un nivel global.
- Eventos deportivos destacados, como *CWC19*, *SB19*, *CopaAmerica* y *FIFAWWC*,
- Eventos de premiaciones musicales, como *TeenChoice* y *KCA*, donde los ganadores son elegidos a través de votaciones populares. Estos eventos a menudo generan comunidades de seguidores en Twitter, especialmente en el caso de bandas coreanas, como *teambts*, *choiceinternationalartist*, *teamgot7*, *teamexo* y *mtvlafandombtsarmy*.

Es importante destacar que estos *hashtags* están asociados con eventos que tienen un rango temporal definido, como eventos anuales o nuevas temporadas de programas. Aunque varían en términos de cuándo ocurren, son frecuentes y generan discusiones significativas mientras están en curso.

Cluster 8

Este cluster comprende un total de 121 *hashtags* que se concentran en una variedad de temas relacionados con el tiempo libre, el entretenimiento y las emociones al respecto. Los *hashtags* cercanos al centroide incluyen términos como *memories*, *creative*, *comingsoon*, *memes*, *natural*, *book*, *support*, *interiordesign*, *positivevibes* y *poetry*.

Se incluyen distintas temáticas como:

- Actividades relacionadas con el tiempo libre, como la lectura de libros (*book*), la creación de contenido creativo (*creative*) y la expresión artística a través de la poesía (*poetry*).

- Emociones y actitudes positivas, como *positivevibes* con respecto a un hecho.
- Mascotas, con *hashtags* como *dogs*, *cat*, *catsoftwitter*, *birds* y *adoptdontshop*, que indican la participación en conversaciones sobre mascotas y cuidado de animales.
- Otros temas de entretenimiento, como *comics*, *gaming* y *quotes*, que abordan intereses variados, desde cómics y videojuegos hasta citas y reflexiones inspiradoras.
- Hashtags relacionados con contenido viral en las redes sociales, como *memes*, *trending* y *viral*, que indican un interés por el humor y el contenido popular.

Los *hashtags* utilizados en este cluster son recurrentes y a menudo se utilizan en conjunto con otros para formar un contexto completo. Reflejan una variedad de intereses y pasatiempos de las personas en la plataforma.

Cluster 9

Este cluster está compuesto por 39 *hashtags* que se encuentran ligados a temas de tecnología e industria. Los *hashtags* cercanos al centroide incluyen términos como *digital*, *tech*, *bigdata*, *machinelearning*, *iot*, *digitalmarketing*, *finance*, *blockchain*, *startup* y *manufacturing*.

En este conjunto de *hashtags*, se destacan temas relacionados con la computación y las nuevas herramientas tecnológicas. Esto sugiere un interés en la innovación y el avance de la tecnología.

También se hacen referencias a trabajos relacionados con la ciencia y la tecnología, lo que indica la presencia de profesionales y entusiastas del campo STEM (ciencia, tecnología, ingeniería y matemáticas).

Sin embargo, este cluster no se limita únicamente a trabajos tecnológicos, ya que también incluye *hashtags* relacionados con áreas como *nursing*, *hr*, *transportation* y *customerservice*.

Además, se mencionan términos asociados a la industria y al marketing de empresas, ya sea en el ámbito digital o no. Esto indica un enfoque en estrategias de negocios y la promoción de productos y servicios en línea.

Se destacan varios términos asociados a la tecnología, como *iot*, *machinelearning*, *bigdata*, *bitcoin* y *blockchain*, lo que demuestra un interés en áreas tecnológicas avanzadas y tendencias actuales.

Es probable que estos *hashtags* generen comunidades en línea con intereses comunes en tecnología y empleo en el campo de la ciencia y la tecnología. Al utilizar estos *hashtags*, las publicaciones pueden aumentar su visibilidad y permitir que personas interesadas encuentren noticias, artículos y discusiones relacionadas con estos temas de interés.

3.4.2. Muestra 2

La muestra a explorar se compone por 29263 datos con 17 características numéricas cada uno, además existe una vectorización de palabras para cada *hashtag* correspondiente.

1. Caso PCA

Empleando análisis de componentes principales en la muestra número 2, se obtiene que con 9 componentes se explica el 96 % de la varianza total, lo que significa que al mantener esta cantidad de dimensiones, no se pierde mucha información. Estos 9 componentes pueden ser descritos como:

- Componente 1: 36.96 %
- Componente 2: 24.54 %
- Componente 3: 10.32 %
- Componente 4: 7.53 %
- Componente 5: 5.80 %
- Componente 6: 5.49 %
- Componente 7: 3.57 %
- Componente 8: 3.28 %
- Componente 9: 2.51 %

En la figura 27 se puede observar la representación de los primeros dos componentes, los cuales consideran más del 50 % de la varianza total:

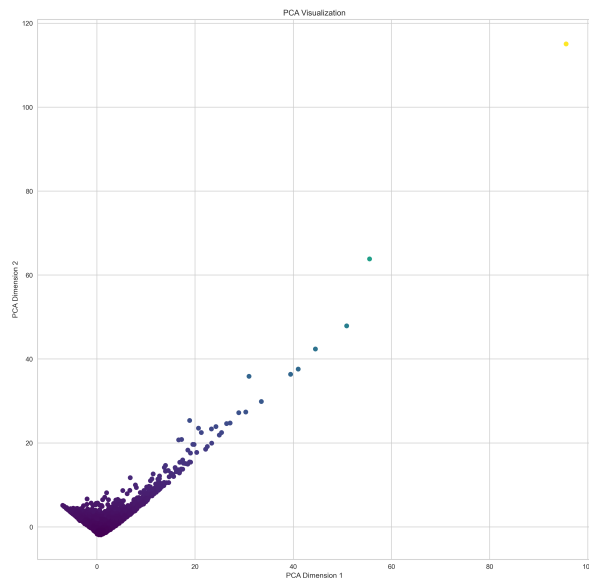


Figura 27: Representación de los primeros dos componentes
Fuente: Elaboración a partir de código

El vector de palabras, al utilizar PCA, se obtiene un total de 43 componentes que describen el 95 % de la varianza de los datos. Con los nuevos vectores de dimensiones reducidas, se fusionan en uno común para poder realizar el clustering.

Para obtener el número óptimo de clusters para el algoritmo de K-Means, se utilizaran las siguientes heurísticas que arrojaron los siguientes resultados:

■ Método del codo: Señala 9 clusters

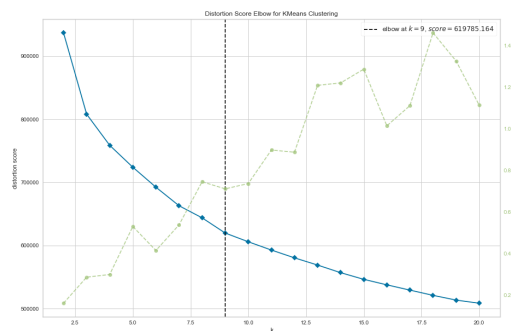


Figura 28: Método del codo
Fuente: Elaboración a partir de código

■ Índice de Davies-Bouldin: sugiere 3 clusters

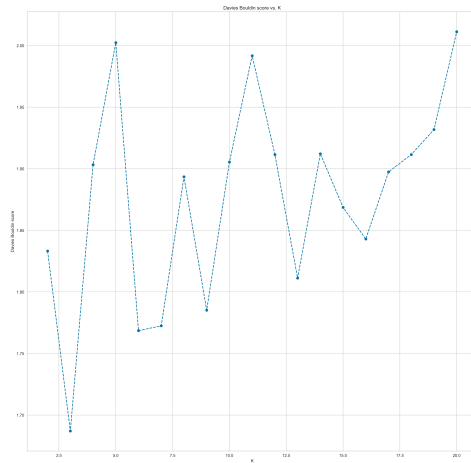


Figura 29: Índice de Davies-Bouldin
Fuente: Elaboración a partir de código

- **Índice de Calinski-Harabasz:** también sugiere como óptimo 3 clusters

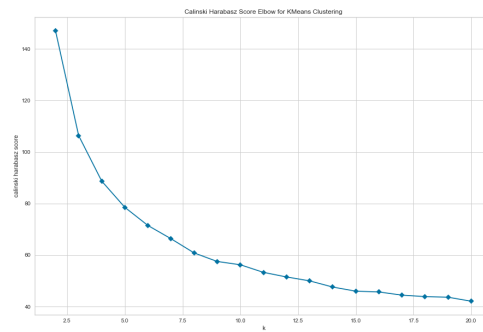


Figura 30: Índice de Calinski-Harabasz
Fuente: Elaboración a partir de código

- **Estadístico de GAP:** propone 20 clusters

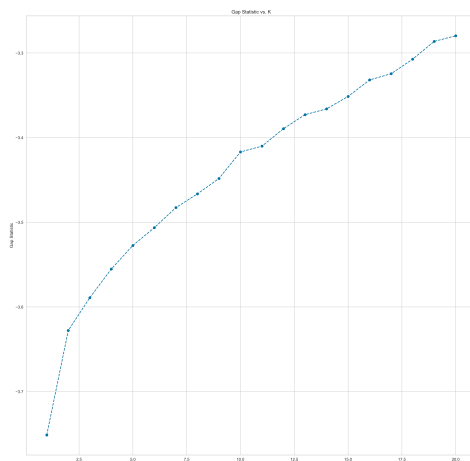


Figura 31: Estadístico de GAP
Fuente: Elaboración a partir de código

Se seleccionará el resultado dado por el método del codo, esto debido a la gran cantidad de datos y la naturaleza de estos, de forma visual, el resultado del clustering para un valor de k igual a 9 se observa en la figura 32.

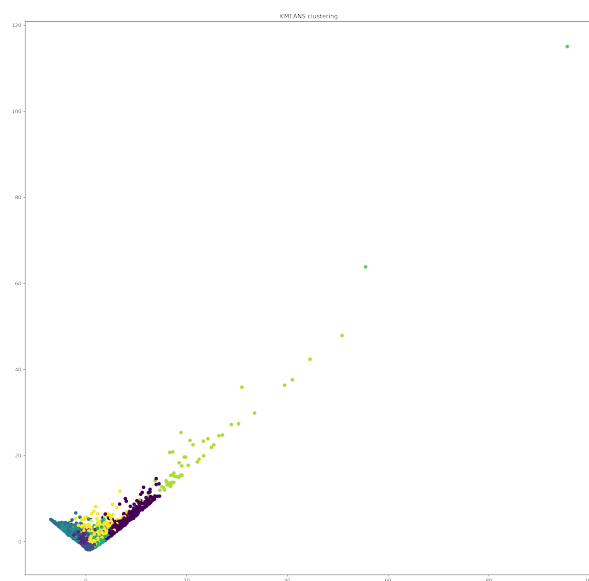


Figura 32: Resultado k -means con $k=9$, reducción de dimensiones por PCA
Fuente: Elaboración a partir de código

Cluster 1

Este cluster de *hashtags*, que contiene un total de 731 elementos, abarca una amplia gama de temas y elementos genéricos que se utilizan en diversas situaciones en las redes sociales. Los

hashtags cercanos al centroide son *awesome, magic, latepost, mood, newyear, instamood, day, lovemyjob, time, enjoy*; estos reflejan una serie de categorías y temas. En este cluster, se encuentran *hashtags* relacionados con:

- **Vacaciones y Viajes:** Expresiones como *vacation, wanderlust, tour* y *trip* sugieren un interés en compartir experiencias de viaje y aventuras.
- **Destinos y Lugares:** Se mencionan lugares específicos, tanto países como ciudades, como *Australia, Berlin, Brasil, Chicago, China*, y *Disneyland*. Además, se mencionan tipos de lugares como *beach* y *city*.
- **Días de la Semana:** Se incluyen *hashtags* relacionados con días de la semana, cómo *wednesdaywisdom, throwbackthursday, fridayfeeling, sundayfunday*, y *tgif*, que suelen utilizarse para compartir contenido específico de cada día
- **Comida:** Se mencionan *hashtags* relacionados con comida y restaurantes, como *coffee, yummy, cocktails, pizza, restaurant*, y *breakfast*.
- **Tecnología:** Se incluyen *hashtags* relacionados con tecnología y desarrollo, como *ai, machinelearning, blockchain*, y *javascript*.
- **Estilos de Vida:** Se mencionan *hashtags* relacionados con maquillaje (*makeup*), moda (*moda*), vida nocturna (*nightlife*), y fiestas (*party*).
- **Entretenimiento:** Se hace referencia a la popular serie *Game of Thrones (gameofthrones)*, que generó conversaciones a nivel mundial, siendo esta la más comentada en el 2019. También se mencionan géneros musicales como el *rap, pop, house, edm, kpop* así como plataformas de transmisión de contenido audiovisual como *Netflix, Twitch* y *Spotify*
- **Deportes:** Se incluyen *hashtags* relacionados con deportes como *golf, football, soccer, basketball* y *cycling*.
- **Fotografía:** Se mencionan *hashtags* que indican con qué tipo de dispositivo se tomó una fotografía, como *shotoniphone* y *canonphotography*, así como ediciones y estilos de fotos como *vsco, photoshop*, y *vintage*.
- **Interacciones Sociales:** Se incluyen *hashtags* utilizados para aumentar interacciones y seguidores, como *l4l, followback, followforfollow*

En conclusión este cluster agrupa *hashtags* genéricos y versátiles que se utilizan en diversas situaciones y contextos, los cuales pueden ser empleados para describir, etiquetar y contextualizar una amplia variedad de temas y momentos personales, lo que convierte en *hashtags* atemporales para ser utilizados en cualquier momento.

Cluster 2

Este cluster posee una gran cantidad de elementos, exactamente 5299. Los *hashtags* más cercanos al centroide son *boycottfrance*, *mlscupplayoffs*, *renewannewithane*, *digitalnigeria*, *la-radiodelhincha*, *dalitlivesmatter*, *jaymaharashtra*, *alertanacional*, *badboysforlife*, *silnirazem*.

Entre esos *hashtags*, *boycottfrance* surgió como movimiento de protesta en respuesta a ciertas políticas gubernamentales en Francia y ha sido utilizado por activistas. En el ámbito deportivo se encuentra *mlscupplayoffs* que se ha utilizado para seguir los *playoffs* de la MLS. Para el caso de la *copaamerica*, se encuentran ciertas variantes con tildes y detalles específicos como el año por ejemplo *copaamerica2019*. Asimismo, siguiendo esta línea, dentro de este cluster hay *hashtags* asociados a otros eventos deportivos como la copa mundial de rugby (*rugbyworldcup*), la copa del mundo de cricket (*cricketworldcup2019*) y el superbowl (*superbowl2019*, *superbowl53* y *superbowlliii*). Estos *hashtags* se utilizan para comentar sobre los eventos, sin embargo, no son los *hashtags* oficiales en su mayoría pero de igual forma representan el evento a pesar de no ser los de mayor frecuencia.

En un contexto cultural, ocurre lo mismo para el caso de eurovision (*eurovision2019* y *eurovisionagain*) y también, para el caso del movimiento BLM que se utilizan otros *hashtags* como *justiceforbreonnataylor*, *justiceforfloyd* y *justiceforgeorge* junto al oficial.

En relación a los *hashtags* que hacen referencia a fechas exactas del año, como *01mayo* o *04jul*, es interesante notar que estos se utilizan solo una vez al año en la fecha correspondiente, y en este caso, en los años 2019 y 2020. Al corresponder a días precisos del año, su comportamiento en relación al rango es similar, ya que se repiten cada 365 días. Además, en este cluster también se incluyen *hashtags* relacionados con aniversarios y conmemoraciones, como *10anniversary* o *50yearsofmarriage*. Aunque estos elementos no representan eventos conocidos, siguen un patrón de temporalidad y suelen incluir al menos un dígito como característica distintiva.

En cuanto a la pandemia de COVID-19, se observa que este cluster contiene *hashtags* que combinan el término *covid* o *coronavirus* con nombres de países o localidades. Esto sugiere que los usuarios utilizaron estos *hashtags* como una forma de difundir información de manera más localizada y generar comunidades en torno a temas relacionados con la pandemia en regiones específicas.

También se hacen presentes *hashtags* relacionados con el estallido social en Chile del 2019, como *chiledesperta*, *chileresiste*, *nuevaconstitucionparachile*, y *apruebo* lo que sugiere una conciencia y apoyo a los acontecimientos políticos y sociales en Chile.

Dado el tamaño del cluster y la variedad presente, es desafiante encontrar una categoría generalizada que englobe todos los elementos. Además, la presencia de *hashtags* de pocos caracteres, caracteres no relacionados con el español o el inglés y derivaciones de otros *hashtags*, que muchas veces comprende un error de escritura, complica aún más la comprensión de algunos elementos, sin embargo la mayoría presente surge respecto a un evento tempo-

ral.

Cluster 3

El cluster que engloba una gran cantidad de *hashtags*, con un total de 12,821, refleja una diversidad de temas e intereses. Entre los más cercanos al centroide del cluster se encuentra *greenday*, *oneokrock*, *einstein*, *kuwtk*, *thequeen*, *omega*, *bentley*, *coldplay*, *trolls*, *werbung*

Algunos de los aspectos más destacados de este cluster incluyen:

- **Signos zodiacales:** Es interesante notar que los signos zodiacales son un tema recurrente en este cluster. Los *hashtags* relacionados con los signos del zodiaco, como *virgo*, *leo* y otros, puede indicar discusiones sobre astrología y horóscopos en sus publicaciones.
- **Entretenimiento:** Se mencionan series de televisión populares, como *greysanatomy*, *theflash*, *thegooddoctor*, y *thevoice*, lo que indica que las conversaciones en este cluster pueden estar relacionadas con programas de televisión. Además, se hacen referencias a bandas musicales como *greenday*, *coldplay* y *ironmaiden*, lo que sugiere una discusión sobre música y artistas.
- **Películas populares:** *Hashtags* como *jurassicworld* y *jurassicpark* indican conversaciones sobre películas populares, en particular la franquicia de Jurassic Park.
- **Diversidad lingüística:** Se observa que los *hashtags* en diferentes idiomas están presentes en este cluster, lo que sugiere una diversidad de usuarios de diferentes regiones del mundo. Esto puede deberse a que los temas de interés y las conversaciones abarcan una audiencia global.

Dada la diversidad y la cantidad de elementos en este cluster, no es posible encontrar una categoría única que los englobe a todos. En su lugar, este cluster refleja la variedad de intereses y temas que son relevantes para los usuarios de las redes sociales durante el período analizado.

Cluster 4

Con un total de 2629 elementos, en este cluster se encuentran *hashtags* bastante largos que pueden complicar la comprensión a primera vista. La media de longitud de los *hashtags* en este cluster es la más alta de los cluster correspondientes a este caso. Los *hashtags* que más cerca están del centroide son *ssrcaseisnotsuicide*, *biharrejectsnitish*, *saverailwaysavenation*, *immediatecbiforssr*, *modisurrenderstochina*, *imrankhanpubgkhola*, *mahagovtcbiforssr*, *indiawithdrswamyforssr*, *billclintonisapedo*, *arrestanuragkashyap*.

La presencia de *hashtags* relacionados con Aldub representa el 6 %, esto representa a una famosa pareja de estrellas de televisión filipina formada por *aldenrichards* y *mainemendoza*. Aunque esta historia se hizo conocida hace varios años, aún mantiene una gran popularidad en Filipinas.

Además, la mayoría de los *hashtags* en este cluster son de naturaleza local, lo que significa que están relacionados con eventos, celebridades o temas específicos de un país o región en particular. Esto indica que este cluster podría ser particularmente relevante para un público geográficamente limitado o interesado en temas específicos de una ubicación en particular.

Cluster 5

Este cluster compuesto por 2,623 *hashtags* y los más cercanos al centroide son *justiceforaudrey*, *reformadeumtrilhao*, *adaletyerinibuldu*, *jepsinobjecciones*, *kpujangancurang*, *pray_for_neasamani*, *tangkapprabowo*, *mamatafreebengal*, *myvoteforcongress*, *national-warmemorial*.

Encontramos menciones a fechas específicas en distintos idiomas, lo que sugiere la presencia de usuarios de diferentes regiones geográficas compartiendo momentos significativos. Además, se observa una serie de *hashtags* relacionados con oraciones de apoyo y solidaridad hacia diversas causas, lo que muestra cómo las redes sociales se utilizan para expresar empatía y preocupación por situaciones y personas que necesitan apoyo.

Dentro de este conjunto de *hashtags*, destaca la mención recurrente de *aldub*, que hace referencia a un fenómeno del entretenimiento filipino, y también se mencionan actores relacionados.

Es importante notar que la mayoría de estos *hashtags* tienen un rango de tiempo relativamente corto y no muestran una alta frecuencia de uso, lo que sugiere que son parte de conversaciones específicas en momentos determinados. Además, se notan variaciones de *hashtags* que hacen referencia al mismo tema. En este contexto, también se destaca la mención recurrente de *BTS*, que representa al grupo musical surcoreano muy popular. Esto logra incluir lo que es llamado Twitter fandom, en donde personas con gustos similares generan una comunidad para comentar sobre sus artistas favoritos, en este caso.

Cluster 6

Este cluster de *hashtags*, que consta de un total de 3189 elementos, es altamente diverso y refleja una amplia variedad de temas y estilos de vida comunes en las redes sociales. Los *hashtags* cercanos al centroide incluyen expresiones como *ilovemyjob*, *savethedate*, *unicorn*, *livingthedream*, *lovewhatyoudo*, *dreamteam*, *surprise*, *eyewear*, *volleyball*, *bus* lo que refleja una actitud positiva y motivada hacia la vida y el trabajo.

Uno de los aspectos más notables de este cluster es la presencia de *hashtags* relacionados con diferentes estilos de vida, como *writerslife*, *barberlife*, *cheflife*, *citylife*, y *djlifestyle*. Estos *hashtags* revelan intereses y profesiones específicas, lo que indica que los usuarios que los utilizan están compartiendo sus experiencias y perspectivas relacionadas con sus respectivas ocupaciones y el entorno urbano en el que viven.

El tiempo libre y la fotografía también son temas presentes en este cluster, con *hashtags* relacionados con viajes, vacaciones y actividades al aire libre, así como *surf*, *swim*, *volleyball*. Los usuarios utilizan estos *hashtags* para compartir sus momentos de ocio y aventuras, a menudo resaltando la belleza de lugares escénicos. Además, este cluster contiene *hashtags* que hacen referencia a lugares y destinos diversos, lo que sugiere que los usuarios etiquetan sus fotos para destacar lugares que visitan, lo que podría inspirar a otros a explorar esos mismos lugares.

Los días de la semana en español, así como los meses del año en inglés, también se hacen presentes en este cluster, lo que demuestra una intención de marcar el tiempo y crear un contexto temporal en las publicaciones, posiblemente relacionado con eventos o actividades específicas.

Otro elemento interesante es la inclusión de términos relacionados con memes, tendencias virales y codificación (*coding*). Estos *hashtags* pueden ser utilizados para identificar contenido popular, tendencias en las redes sociales y temas relacionados con la tecnología, lo que refleja la intersección de la cultura en línea y la tecnología en la vida de los usuarios.

Asimismo, se mencionan marcas reconocidas de la industria como *Adidas*, *Audi*, *Nike*, *Nintendo* y *Nissan*, lo que sugiere que los usuarios utilizan estos *hashtags* para mostrar su apoyo a ciertas marcas o productos, o para compartir contenido relacionado con estas marcas.

En resumen, este cluster de *hashtags* abarca una amplia gama de temas y estilos de vida, desde profesiones y viajes hasta comida, marcas y tendencias. Refleja la diversidad de intereses y actividades que los usuarios comparten en las redes sociales para expresar sus estados de ánimo y experiencias cotidianas.

También es importante mencionar que este cluster incluye *hashtags* relacionados con eventos anuales significativos como los oscars, los grammys y cannes. Aunque estos eventos ocurren una vez al año, los *hashtags* relacionados con ellos son genéricos y se utilizan durante todo el año para contextualizar contenido, señalar acontecimientos relevantes, publicitar fechas y participar en conversaciones relacionadas con la industria del cine y la música.

Cluster 7

Este cluster se caracteriza por tener solo dos elementos, y el atributo más relevante en este caso es la frecuencia, seguido de la cantidad de *hashtags* con los que co-ocurre y la cantidad de co ocurrencias. Los dos *hashtags* son:

- **covid19:** Este es el término más mencionado en todo el mundo en el año 2020, ya que se refiere a la pandemia de COVID-19 que afectó a nivel global. Su alta frecuencia indica la importancia e impacto que tuvo este tema en Twitter.
- **coronavirus:** Similar a *covid19* (véase tabla 1), este término es otra forma de referirse a la pandemia

La presencia de solo estos dos elementos en el cluster, ambos relacionados con la pandemia, es coherente con las características de alto uso y frecuencia de estos términos durante el año 2020. Además, la similitud entre *covid19* y *coronavirus* en la vectorización de palabras refuerza la conexión entre estos dos términos y su aparición en el mismo cluster.

Cluster 8

En este cluster de *hashtags*, que comprende un total de 55 elementos, se observan varias características notables que definen su naturaleza. Las métricas más destacadas son la cantidad de *hashtags* que co-ocurren en publicaciones junto con la frecuencia de uso, lo que sugiere que estos *hashtags* suelen utilizarse en combinación con otros con alta regularidad. Además, la media mensual de uso también es relevante, ya que indica cuán populares son estos *hashtags* a lo largo del tiempo. Entre los elementos más cercanos al centroide se encuentran *repost*, *newprofilepic*, *sales*, *photo*, *life*, *tbt*, *picoftoday*, *mondaymotivation*, *london*, *photooftoday*.

Uno de los aspectos interesantes es la presencia de *hashtags* relacionados con eventos y tendencias significativas. Por ejemplo, *bbb20* se refiere a *Big Brother Brasil 2020*, un programa de televisión que generó un gran interés y conversación a nivel mundial durante ese año. También se encuentran *hashtags* como *endsars* y *blacklivesmatter*, que representan movimientos sociales en diferentes países que abordan la brutalidad policial y la justicia racial, con el segundo ganando relevancia global tras los trágicos eventos relacionados con George Floyd y Breonna Taylor.

Además, se mencionan *hashtags* relacionados con eventos políticos importantes, como *auspol* (elecciones en Australia) y *brexit* (proceso político relacionado con el Reino Unido), lo que demuestra la capacidad de las redes sociales para reflejar y participar en conversaciones globales sobre asuntos políticos. Asimismo, se incluyen *hashtags* relacionados con grupos de música como *BTS* y *Got7*, lo que refleja la popularidad y creación de comunidades en torno a un gusto en común.

Por último, se mencionan *hashtags* relacionados con las medidas de confinamiento por la pandemia de COVID-19, como *stayhome* y *quedateencasa*, que fueron ampliamente utilizados durante el período de cuarentena global.

Los *hashtags* aquí presentes son tanto específicos de eventos y tendencias particulares como genéricos y atemporales que se utilizan en diversas ocasiones. Esto refleja la versatilidad de

las redes sociales como plataforma para compartir ideas, conversaciones y tendencias a nivel mundial.

2. Caso UMAP

Para este caso, se realizará una reducción de dimensionalidad para los atributos numéricos y los vectores de palabras, con un valor de 2 componentes para cada uno, lo que se ve reflejado en la figura 33.

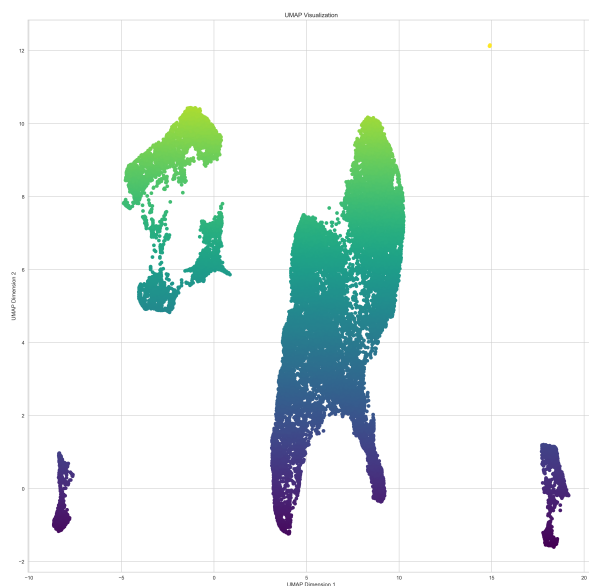


Figura 33: Dimensiones a partir de UMAP
Fuente: Elaboración a partir de código

Una vez reducidas las dimensiones de las características numéricas y la vectorización de palabras, se fusionan en un vector común.

Se consideran resultados a partir de las siguientes métricas para decidir el número de clusters a considerar:

- El método del codo sugiere un número óptimo de 8 clusters

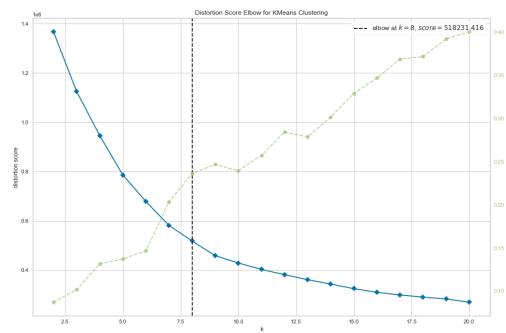


Figura 34: Método del codo
Fuente: Elaboración a partir de código

- el índice de Davies-Bouldin sugiere 7 clusters

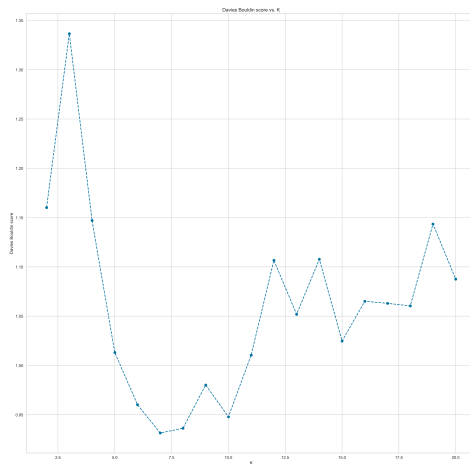


Figura 35: Índice de Davies-Bouldin
Fuente: Elaboración a partir de código

- el índice de Calinski-Harabasz sugiere 2 clusters

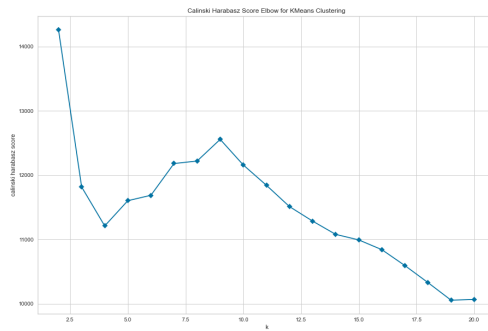


Figura 36: Índice de Calinski-Harabasz
Fuente: Elaboración a partir de código

- el estadístico de gap sugiere 7 clusters

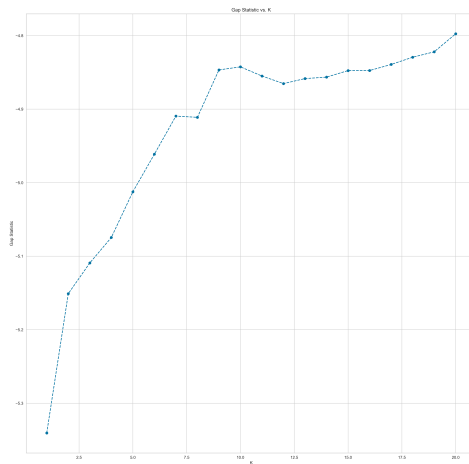


Figura 37: Estadístico de gap
Fuente: Elaboración a partir de código

Para este caso, se selecciona un óptimo de 7 clusters para aplicar el algoritmo K-Means. Esto a partir de lo señalado por la métrica de Davies y del estadístico de GAP. Trás aplicar el algoritmo, los clusters obtenidos se pueden visualizar en la figura 38

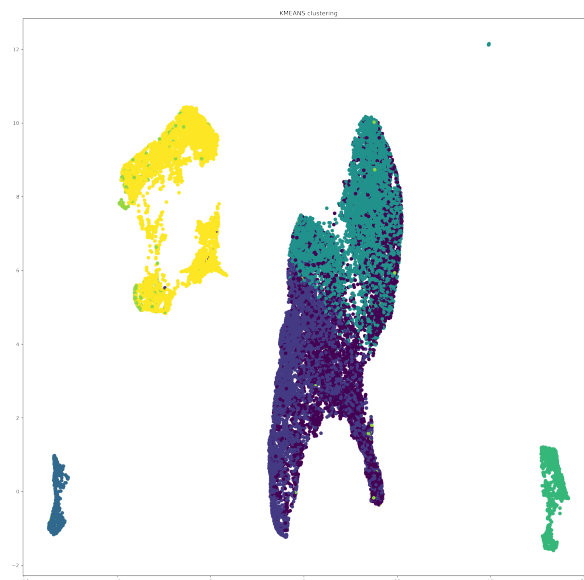


Figura 38: Resultado *k-means* con $k=7$, reducción de dimensiones por UMAP
Fuente: Elaboración a partir de código

Cluster 1

Este diverso cluster de *hashtags*, compuesto por un total de 6,860 elementos, abarca una amplia variedad de temas y ubicaciones, lo que lo hace sumamente heterogéneo. Los elementos cercanos al centroide son representados por *huesca*, *endirecto*, *jaen*, *vinadelmar*, *tigres*, *quilpue*, *larioja*, *superliga*, *cochabamba*, *buenosdiasmundo*.

En este cluster, los atributos con la varianza más alta son la cantidad de *hashtags* que co-ocurren y la frecuencia de uso. Esto significa que estos *hashtags* tienden a aparecer juntos con cierta regularidad en las conversaciones en línea, y también se utilizan con bastante frecuencia en las publicaciones de los usuarios. Esta correlación sugiere que estos *hashtags* están relacionados y que a menudo se utilizan en conjunto para contextualizar o ampliar el alcance de las conversaciones en las redes sociales.

Sin embargo, a pesar de la co-ocurrencia frecuente de ciertos *hashtags*, la cantidad de datos en este cluster es considerable, lo que dificulta encontrar una categoría única que pueda describir la mayoría de estos *hashtags* según su comportamiento. Una de las características notables de este cluster es la diversidad geográfica que abarca. Se mencionan localidades y regiones de todo el mundo, pero se destaca una gran cantidad de lugares en Chile como *vinadelmar*, *quilpue*, *atacama*, *antofagasta*, *santiago*, *coquimbo*, *chiloé*, entre otras. Esta variedad geográfica indica que las conversaciones en este cluster son globales y tocan temas de diversas regiones del mundo.

En cuanto a los temas específicos, este cluster aborda una amplia gama de intereses y ten-

dencias. Por un lado, se mencionan eventos deportivos de relevancia internacional como la *ucl*, siendo este el evento deportivo más comentado en el año 2019, la *CopaAmerica* y la *CopaLibertadores*. El aspecto social también está presente, con *hashtags* relacionados con movimientos sociales significativos como *BLM (BlackLivesMatter)*, *GeorgeFloyd* y *BreonnaTaylor*. Estos *hashtags* reflejan conversaciones en torno a la justicia social y la lucha contra la brutalidad policial, temas que atrajeron una atención significativa en las redes sociales.

La música asiática y las bandas tienen un lugar destacado en este cluster, con menciones a grupos como *BTS*, *Blackpink* y *BNK48*. Esto indica la influencia global de la música asiática y la dedicación de sus seguidores en las redes sociales.

La mayoría de estos *hashtags* podrían considerarse recurrentes, ya que hacen referencia a lugares, bandas o eventos que tienen un período definido, lo que permite saber cuándo ocurrirán. Por otro lado, algunos *hashtags ad hoc*, como *GeorgeFloyd* y *BreonnaTaylor*, se crearon específicamente para comentar, informar y manifestarse respecto a las personas asesinadas con sus nombres.

En conjunto, este cluster es un testimonio de la diversidad y la dinámica de las conversaciones en las redes sociales, donde múltiples temas y tendencias coexisten y se entrelazan, reflejando la complejidad de la comunicación y su capacidad para abordar una amplia gama de intereses y preocupaciones globales.

Cluster 2

Este cluster abarca una amplia variedad de tendencias que fueron destacados en las redes sociales durante un período específico a través de los 6123 elementos que lo componen. Algunos de los *hashtags* cercanos al centroide del cluster son *sponsored*, *squad*, *graduation*, *wrc*, *exploration*, *angels*, *ifb*, *manager*, *hemp*, *quiz*.

Entre los *hashtags* más destacados se encuentran *covid19* y *coronavirus*, que fueron los más utilizados en el año 2020 en relación con la pandemia global. Junto a estos, se mencionan *hashtags* relacionados con la situación de confinamiento, como *lockdown*, *stayhome* y *quarantine*, que reflejan la experiencia vivida de las personas durante la pandemia.

Un tema destacado en este cluster es el *hashtag endsars* y *endsarsnow*, que se relaciona con las protestas en Nigeria contra la brutalidad policial y el movimiento para poner fin a la unidad especial de la policía nigeriana conocida como SARS.

También se mencionan *hashtags* relacionados con programas de televisión y películas populares en ese momento, como *avengersendgame*, que se convirtió en la película más twitteada en la historia y *gameofthrones*, la serie más comentada a nivel mundial. Otros programas como *strangerthings*, *twd*, *thebigbangtheory* y *theboystv* también obtuvieron menciones, surgiendo probablemente debido al estreno de una nueva temporada.

La música también tiene su espacio en este cluster, con *hashtags* que abarcan géneros como *pop*, *heavymetal*, *rap*, *hiphop* y *rock*, así como menciones relacionadas a instrumentos musicales como *drum*, *drummer*, y *guitar*. Además, festivales de música mundialmente conocidos como *lollapalooza*, *coachella* y *tomorrowland* también se hacen presentes.

Aspectos políticos no pasan desapercibidos, con *hashtags* como *brexit*, *auspol* y *cdnpoli* que se relacionan con procesos políticos en el Reino Unido, Australia y Canadá respectivamente.

Este cluster también incluye una serie de *hashtags* recurrentes que se pueden aplicar en una variedad de contextos, desde deportes hasta estilos de vida, tecnología, entretenimiento y actividades cotidianas. Además, hay menciones de eventos conocidos mundialmente como los *Oscars*, la *FIFAWWC* (Copa del Mundo Femenina de la FIFA) y la *rugbyworldcup* (Copa del Mundo de Rugby).

Por último, los *hashtags* cercanos al centroide como *sponsored*, *squad*, *graduation*, *wrc*, *exploration*, *angels*, *ifb*, *manager*, *hemp* y *quiz*, agregan una capa adicional de diversidad y representan temas variados que se relacionan con una amplia gama de conversaciones en las redes sociales.

Todo el conjunto de *hashtags* presentes en el cluster agrega un valor importante a la diversidad de las conversaciones de actualidad generadas en la plataforma dentro del periodo establecido.

Cluster 3

Este cluster compuesto por 1,189 *hashtags* presenta una característica común: todos los elementos incorporan al menos un dígito numérico. La presencia de números en estos *hashtags* les otorga una característica esencial para poder entender su significado. Los *hashtags* cercanos al centroide incluyen *bedtet2011*, *108mp*, *69000teachervacancy*, *sadak2*, *covid19lk*, *savecutoff90_97*, *covid19sl*, *navratri2020*, *sars_cov_2*, *justice4ssr*

Los *hashtags* presentes en este cluster pueden servir para una gran cantidad de propósitos:

- **Ediciones de productos y versiones de software:** Algunos de estos *hashtags* hacen referencia a ediciones específicas de productos como *samsungGearS3*, *iphone11pro*, *windows10* y *office365*.
- **Cantidades:** Los números en estos *hashtags* a menudo indican cantidades, como *10k*, *4200gr*.
- **Referencias Históricas:** Los números también pueden evocar épocas históricas o décadas específicas, como *50s*, *1960s*
- **Eventos y Celebraciones:** Se pueden utilizar para señalar años en eventos o celebraciones específicas como por ejemplo *halloween2020*, *navidad2020*.

- **Campañas:** Algunos *hashtags* con números pueden formar parte de campañas de elecciones como es el caso de *joebidenforPresident2020*, *voteTrump2020*.

La interpretación precisa de estos *hashtags* depende del contexto en el que se utilizan y de la intención del usuario.

Cluster 4

Este cluster, conformado por 7,208 *hashtags*, presenta una característica interesante: la mayoría de ellos tienen una frecuencia promedio de alrededor de 98, lo que sugiere que no son ampliamente utilizados o populares en comparación con otros *hashtags* más frecuentes en Twitter. Entre los *hashtags* que se encuentran al centroide están *affirmation*, *tuesdaymood*, *instafitness*, *periscope*, *wordoftheday*, *encouragement*, *beachbody*, *youcandoit*, *goodpeople*, *brilliant*.

Este cluster se encuentran palabras y expresiones relacionadas con la motivación y el bienestar, como *affirmation* *youcandoit*, y *encouragement*. Esto sugiere que los usuarios que utilizan estos *hashtags* pueden estar compartiendo mensajes positivos y motivadores en sus publicaciones.

Además, se mencionan diversos lugares, como *zambia*, por ejemplo, lo que podría indicar que algunos usuarios en este cluster están discutiendo temas relacionados con esta ubicación geográfica o compartiendo contenido relacionado con Zambia.

Es importante destacar que aproximadamente 75 *hashtags* en este cluster contienen más de 45 caracteres, lo que los hace difíciles de entender o interpretar debido a la longitud y la falta de contexto. Esto podría sugerir que algunos usuarios crean *hashtags* personalizados o únicos que pueden ser específicos de ciertas comunidades o eventos.

También se observa que este cluster contiene *hashtags* relacionados con los días de la semana, lo que indica la presencia de publicaciones que hacen referencia a diferentes días y estados de ánimo durante la semana. Además, algunos *hashtags* están relacionados con conceptos deportivos, aunque no son los más destacados en este cluster.

Cluster 5

Este cluster de *hashtags*, con un total de 1,498 elementos, se caracteriza por la presencia de *hashtags* que contienen números en su estructura, de forma similar al Cluster 3. Estos *hashtags* suelen seguir un patrón específico relacionado con fechas, eventos, años, temporadas y ediciones de programas de televisión, aniversarios y celebraciones. Estas etiquetas numéricas proporcionan información sobre el contexto o el motivo de la publicación en la que se utilizan.

Entre los *hashtags* cercanos al centroide se incluyen ejemplos como *onam2019*, *eidalfitr2019*, *sona2019*, *iwd19*, *bjpsankalppatr2019*, *3musicawards19*, *happyfathersday2019*, *eleksyon2019*, *hblpsl2019*, *mlk2019*.

Se mostraran algunos ejemplos:

- **Temporadas y ediciones de programas:** Nuevas temporadas surgen y también, para los eventos anuales estos tienen un número de edición tal es el caso de *strangerThings3*, *SuperBowl53*, *emmys2019*, *tiff2019*.
- **Días del año:** Para entender esto, se requiere de un contexto más amplio ya que solo se menciona un día y mes como *28ene*, *21nov*, *1mayo*.
- **Conmemoraciones:** Se utiliza para demarcar una fecha recordada de forma anual como *15thdeathanniversary*, *bts6thanniversary*, *10yearsof1d*.

Los *hashtags* en este cluster son útiles para contextualizar las publicaciones y mostrar su relación con eventos o momentos específicos en el tiempo. Otras formas de uso, se mostraron en el Cluster 3, que tenía características similares. Estas etiquetas numéricas permiten a los usuarios compartir recuerdos, conmemorar aniversarios, promocionar eventos o discutir temas relacionados con momentos históricos.

Cluster 6

Este cluster se caracteriza por su diversidad y la presencia de *hashtags* variados que abordan diferentes temas locales. A pesar de su tamaño relativamente pequeño, contiene elementos de relevancia y temas específicos que merecen atención. Los elementos cercanos al centroide lo componen *aldubxadnparakayloloatlola*, *historicrepentancerevival*, *clipetvzvingança*, *aldubxadnmakeadifference*, *aldubxadnunityforall*, *aldubxadnadvocatepeace*, *aldubxadnloveforkeeps*, *aldubxadnsteadyhearts*, *aldubxadnlovinghearts*, *aldubxadnloveembracing*.

Un aspecto destacado de este cluster es la presencia de *hashtags* relacionados con Aldub, como varios representados cercanos al centroide. Aproximadamente el 6 % de los *hashtags* en este cluster hacen referencia a Aldub, lo que sugiere que este tema es relevante y se discute con cierta frecuencia dentro del conjunto de datos analizados.

En cuanto a los atributos, la media muestra valores destacados en el rango y la última fecha de aparición, lo que sugiere que estos *hashtags* tienen una longevidad y continúan siendo utilizados con el tiempo. Por otro lado, en términos de varianza, la cantidad de *hashtags* que co-ocurren muestra una variabilidad significativa, lo que indica que algunos de estos *hashtags* tienden a aparecer juntos en conversaciones relacionadas.

Además de los *hashtags* relacionados con Aldub, se observan otros temas en este cluster, como movimientos sociales y conflictos regionales locales. Por ejemplo, *prayforganja* y *stoparmenianterrorism* se utilizan en conjunto para visibilizar un movimiento social o expresar preocupaciones sobre un conflicto regional específico.

La combinación de temas específicos y locales refleja la diversidad de discusiones en Twitter.

Cluster 7

Compuesto por 5757 elementos, aunque no contiene dígitos, se destacan varios atributos en términos de sus valores medios y variabilidad. Los *hashtags* cercanos al centroide de este cluster, como *bolsonaroimbativel*, *oustdutertenow*, *winnercrosstourinbkk*, *talnation*, *calderonsisabia*, *butecoemcasa*, *livejem*, *defendaosus*, *forabolsonaroemourao*, y *save_the_iraqi_people*, contribuyen a la diversidad temática de este conjunto.

En cuanto a los atributos, es interesante observar que los valores medios más altos corresponden a la última fecha de aparición y la primera fecha, lo que implica que estos *hashtags* tienen una longevidad en su uso y han estado presentes durante un período considerable de tiempo. Esto sugiere que los temas discutidos en estos *hashtags* son relevantes y continúan siendo parte de las conversaciones en línea.

Uno de los temas más prominentes en este cluster es la presencia de *hashtags* relacionados con BTS, una popular banda de K-pop. Estos *hashtags* abordan diversos aspectos de la banda, como nuevas canciones, eventos a los que asisten y celebraciones relacionadas con los cumpleaños de los miembros de BTS.

Además de los temas de K-pop, una parte significativa de este cluster se dedica a cuestiones políticas. Se mencionan candidatos presidenciales como *donaldTrump* y *JoeBiden* en el contexto de las elecciones estadounidenses del 2020, así como el presidente brasileño *JairBolsonaro*.

También se observa la presencia de *hashtags* que contienen errores tipográficos, como *avengersEndame*, *EndSSARS* y *blacklivesmattter*, que parecen ser variantes con errores de *hashtags* populares como *AvengersEndgame* y *BlackLivesMatter*.

Además, algunos *hashtags* hacen referencia al coronavirus y a temas relacionados, como vacunas, países afectados, actualizaciones e incluso, teorías de conspiración. La diversidad de temas y la falta de una categoría general que los abarque demuestran la complejidad de las discusiones generadas por los usuarios.

3. Caso PCA+UMAP

Antes de aplicar el algoritmo UMAP al conjunto ya reducido, se debe seleccionar con cuántas componentes se trabajará. Para esto, de igual forma que en la muestra 1, se analizará el aporte de cada una de las características en los componentes principales. A partir de los resultados obtenidos en 3.4.2 se explica un 96,45 % de la varianza total con 9 componentes.

El primer componente describe el 36,96 % de la variabilidad, siendo los atributos más destacados, en primer lugar, la cantidad de *hashtags* que co-ocurre y le sigue, estos en el mismo nivel, la cantidad de co-ocurrencias, el coeficiente de variación diario y mensual, y el rango; seguido se encuentra la frecuencia, la cantidad de países y la media mensual y diaria. En este caso, el atributo que menor impacto causa es la pendiente mensual. Se puede observar esto en la figura 39.



Figura 39: Aportes de cada atributo en el componente 1

Fuente: Elaboración a partir de código

El componente número 2 logra explicar un 24,54 % de la variación y en conjunto con el anterior, se llega a un valor de 61,5 %. Como atributos, destacan la frecuencia, la media diaria y mensual. En comparación al componente 1, en el caso de la pendiente mensual aumenta aproximadamente 10 veces su valor, mientras que para el caso de la cantidad de países, decae en 3 veces el valor anterior. La fecha máxima se mantiene en un valor bajo. Se puede apreciar en la figura 40.

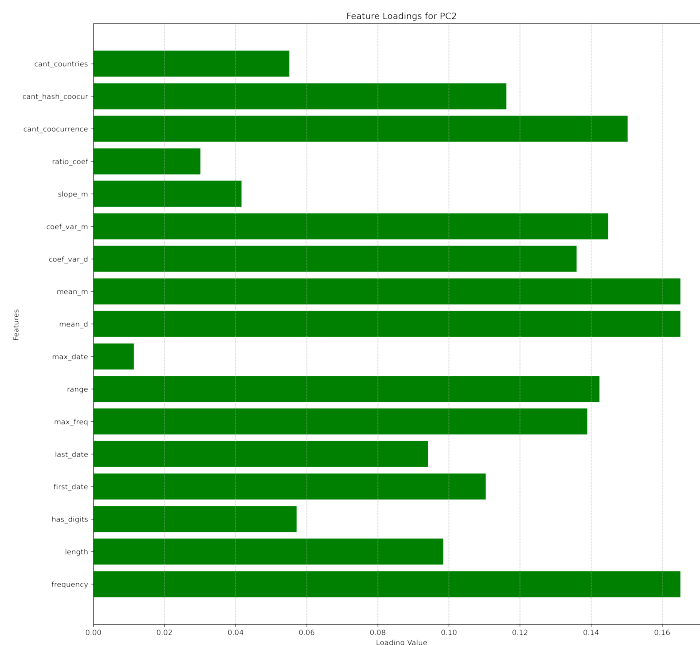


Figura 40: Aportes de cada atributo en el componente 2
Fuente: Elaboración a partir de código

El tercer componente retiene un total de 10.32 % de la varianza. Como se puede observar en la figura 41, los atributos que menos fuerza tenían en los componentes anteriores, son los que destacan para este caso como la pendiente mensual y la fecha máxima. Lo mismo pasa al revés como la cantidad de co-ocurrencias, el rango, la frecuencia, los coeficientes.

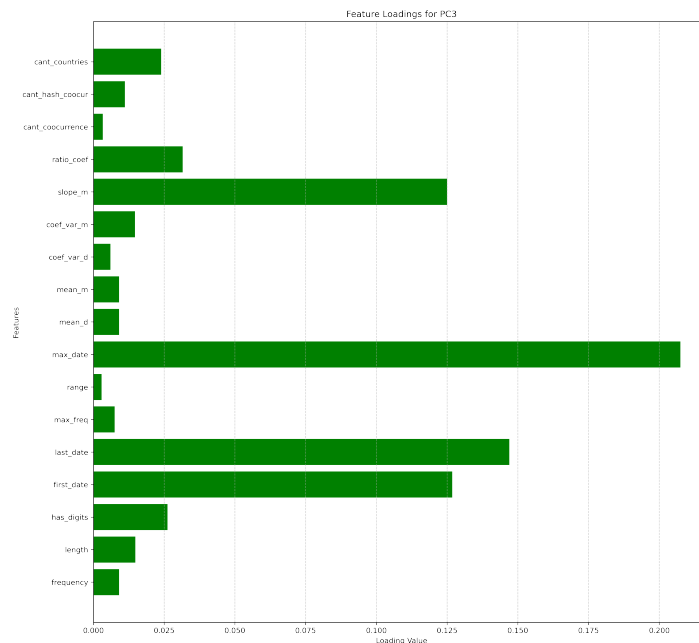


Figura 41: Aportes de cada atributo en el componente 3

Fuente: Elaboración a partir de código

Con este análisis se decide trabajar con solo 2 componentes para los atributos numéricos, ya que de esa forma tiene coherencia los atributos con mayor aporte. Los gráficos que muestran los aportes de los componentes restantes se encuentran en anexo 5.2.

El vector de palabras explica un 95 % de la varianza con 33 componentes. Para este caso, se utilizará los componentes que explican más del 50 % del total, lo que resulta en la selección de 5 componentes.

Una vez realizado, se aplica el algoritmo UMAP a cada vector resultante, para atributos numéricos y de palabras. En la figura 42 se logra observar la representación de las dos componentes obtenidas a partir de aplicar PCA y UMAP al conjunto de datos.

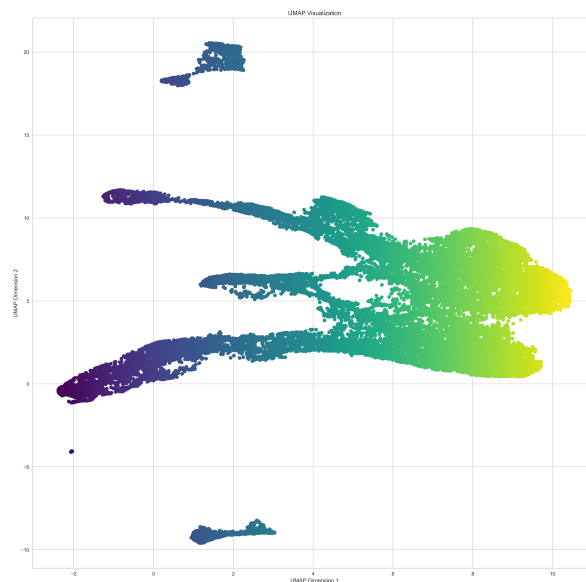


Figura 42: Resultado *k-means* con $k=7$, reducción de dimensiones por UMAP
Fuente: Elaboración a partir de código

Antes de comenzar el clustering con el método K-means, se requiere el número de clusters deseados. Para realizar esto, se emplearán distintas heurísticas:

- el método del codo que sugiere una cantidad de 6 clusters

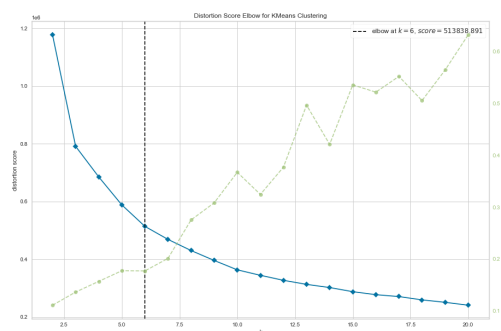


Figura 43: Método del codo
Fuente: Elaboración a partir de código

- el índice de Davies-Bouldien que indica un óptimo de 4 clusters

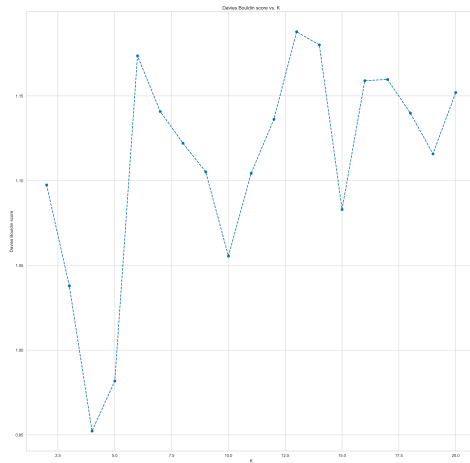


Figura 44: Índice de Davies-Bouldien
Fuente: Elaboración a partir de código

- el índice de Calinski-Harabasz sugiere 3

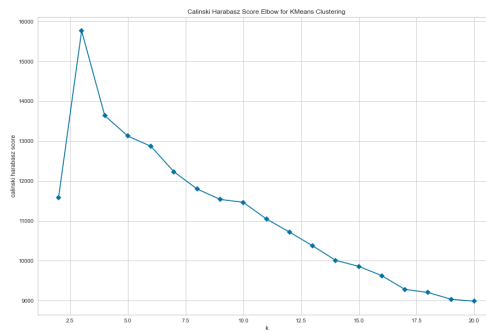


Figura 45: Índice de Calinski-Harabasz
Fuente: Elaboración a partir de código

- El estadístico de GAP propone una cantidad óptima de 20 clusters.

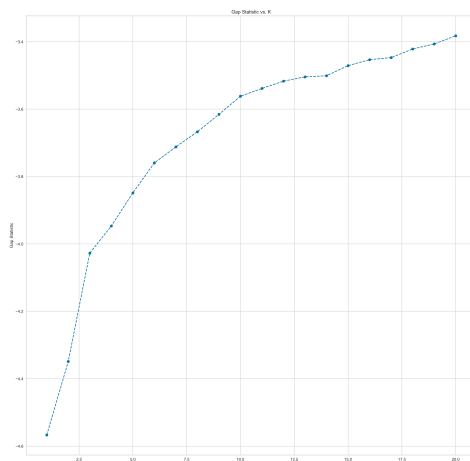


Figura 46: Estadístico de GAP
Fuente: Elaboración a partir de código

Para este experimento se utilizará como número óptimo de cluster el resultado arrojado por el método del codo, lo que equivale a un k con valor igual a 6, resultado reflejado en la figura 47.

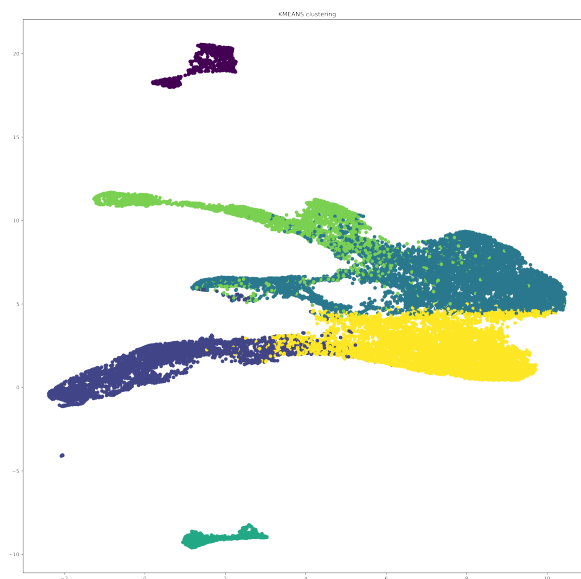


Figura 47: Resultado k -means con $k=6$, reducción de dimensiones por PCA y UMAP
Fuente: Elaboración a partir de código

Cluster 1

Este cluster, compuesto por 1,526 *hashtags*, se caracteriza por incluir *hashtags* que contienen al menos un dígito en su composición. Los *hashtags* cercanos al centroide son cercanos al

centroide *10fight10*, *8martdunyakadinlargunu*, *dia30vemprarua*, *1mayis*, *yks19*, *1dwins10s*, *dia30ultimo*, *aldub188thweeksary*, *fightformonstax7*, *02wintheelection*.

Estos *hashtags* tienden a marcar fechas y períodos específicos durante el rango de tiempo definido. Los atributos con mayor media en este cluster corresponden a la última fecha de aparición, la primera fecha y, como resultado, el rango. Esto sugiere que estos *hashtags* suelen utilizarse para señalar eventos, celebraciones o fechas precisas durante un período específico.

Entre los temas que abarcan estos *hashtags* se encuentran celebraciones y ediciones de eventos, ya que muchos de ellos hacen referencia a fechas exactas o conmemoraciones. También se mencionan temporadas de series y novelas, lo que indica que algunos de estos *hashtags* se utilizan para marcar el inicio o el período de emisión de ciertos programas de televisión.

Además de las celebraciones de aniversarios y cumpleaños, algunos de estos *hashtags* utilizan números como letras o caracteres especiales, lo que les permite demarcar ciertos momentos o eventos de manera única. Esto sugiere que estos *hashtags* pueden ser utilizados para representar de manera creativa fechas o períodos específicos.

Es importante señalar que muchos de los *hashtags* en este cluster no se utilizan durante todo el rango de tiempo definido. En su lugar, muchos de ellos tienen un inicio tardío en relación con el período especificado, lo que indica que su uso se limita a ciertas fechas o eventos específicos.

Cluster 2

El cluster compuesto por 4787 puntos de datos contiene *hashtags* como *ahmadiyyatthetrueislam*, *laxmii*, *potres*, *binod*, *boycottjosim*, *hbdvijaydeverakonda*, *letfmgserve*, *coronovirus*, *hyderabadfloods*, y *obeinrajasthan* los cuales son los más cercanos al centroide.

Los atributos de los *hashtags* en este cluster muestran una alta variación en la frecuencia de uso, la cantidad de co-ocurrencias y la cantidad de *hashtags* que co-ocurren.

En primer lugar, los *hashtags* relacionados con la pandemia de COVID-19 son prominentes en este cluster. *Coronavirus*, *covid*, y *covid_19* son algunos de los *hashtags* utilizados para discutir la pandemia global. Además, se utilizan variantes de estos *hashtags* en los cuales se utiliza como prefijo la palabra coronavirus y se agrega alguna cualidad, localidad, entre otros.

Se observan *hashtags* políticos relacionados tanto con el expresidente Donald Trump como con el entonces candidato Joe Biden. Estos *hashtags* pueden expresar apoyo, críticas o comentarios sobre la política estadounidense y las elecciones presidenciales.

El cluster también incluye *hashtags* relacionados con movimientos sociales, como los *end-*

sarsprotests en Nigeria y los *hashtags* *georgefloydprotests* y *justiceforbreonnataylor* en Estados Unidos. Para ambos casos, estos no corresponden a los *hashtags* más utilizados, sin embargo, estos son utilizados en conjunto para poder dar más contexto y alcance.

Existen *hashtags* en idiomas no fácilmente reconocibles o conocidos, lo que indica la diversidad lingüística de los *hashtags* de este cluster.

Por último, algunos *hashtags* presentan errores tipográficos, lo que puede deberse a la rapidez con la que se crean y comparten publicaciones en las redes sociales.

Dada la diversidad de temas y la gran cantidad de elementos en este cluster, es difícil encontrar una categoría general que los englobe. Los *hashtags* abarcan desde temas de salud pública hasta política, movimientos sociales y tendencias culturales.

Cluster 3

El cluster que engloba *hashtags* cercanos a su centroide como *jleague*, *malvernhillshour*, *pyöräily*, *milestone*, *niigata*, *id*, *bemagenta*, *winitwednesday*, *walls*, y *mt*, esta conformado por 9745 elementos.

En primer lugar, se observa una serie de *hashtags* genéricos que son de los más frecuentes en las redes sociales. Estos incluyen términos como *love*, *travel*, *art*, *music*, *nature*, *free*, *retail*, *sales*, *photooftheday*, y otros similares. Estos *hashtags* suelen utilizarse para etiquetar fotos y publicaciones que se relacionan con actividades cotidianas, intereses personales y expresiones creativas.

También se encuentra el *hashtag* *blacklivesmatter*, que es un movimiento social y un llamado a la igualdad racial. Este *hashtag* se ha utilizado ampliamente para concienciar sobre la injusticia racial y promover el cambio social en Estados Unidos, el cual recobró fuerza luego de los sucesos del año 2020.

Se mencionan eventos y temas relacionados con el entretenimiento, como los Oscars, que son los premios más prestigiosos de la industria cinematográfica. En el ámbito tecnológico, se incluyen *hashtags* relacionados con la innovación y la tecnología, como *ai*, *innovation*, *bigdata*, *machinelearning*, y *blockchain*. Estos *hashtags* indican discusiones y comunidades interesadas sobre avances tecnológicos, inteligencia artificial y temas relacionados con la industria.

En cuanto a la comida y bebida, se mencionan *hashtags* como *coffee*, *pizza*, *cocacola* y *cocktail*, así como también se mencionan las comidas típicas diarias como *breakfast*, *lunch* y *dinner*, los cuales pueden co-ocurrir. Los festivales de música también son un tema destacado en este cluster, con menciones de *coachella*, *lollapalooza*, y *tomorrowland*, siendo festivales muy populares y conocidos en todo el mundo.

Este cluster comparte similitudes con otro cluster previamente identificado, ya que ambos incluyen *hashtags* genéricos y temas de entretenimiento, tecnología, comida y bebida, así como eventos anuales y cuestiones políticas. Sin embargo, a diferencia del otro cluster, este no hace referencia al COVID-19 y se centra más en la cultura pop y la tecnología actual.

Cluster 4

Este cluster está compuesto por 1,201 *hashtags* que comparten la característica de contener al menos un dígito en su composición. Los *hashtags* cercanos al centroide son *bb22*, *oh14*, *midfm761*, *nurses2020*, *wyboryprezydenckie2020*, *thechallenge35*, *freshon947*, *the1975*, *theriver1magic*, *unblocks386*.

Al igual que el cluster número 1, muchos de estos *hashtags* se utilizan para marcar fechas específicas en el tiempo, así como temporadas, ediciones de eventos y programas. También se observa una tendencia a utilizar versiones numeradas en referencia a software, videojuegos y productos tecnológicos.

Además de marcar fechas y eventos, algunos *hashtags* en este cluster utilizan *covid19* como prefijo y luego agregan un país o localidad específica. Esto sugiere que estos *hashtags* se utilizan para discutir temas relacionados con la pandemia de COVID-19 en contextos regionales o locales.

Este cluster se caracteriza por *hashtags* que incluyen dígitos y se utilizan principalmente para marcar fechas, temporadas, ediciones y eventos en el tiempo. También se encuentran *hashtags* relacionados con la pandemia de COVID-19, que hacen referencia a ubicaciones específicas.

Cluster 5

Este cluster se compone de 4,213 *hashtags* que abarcan una amplia variedad de temas y tendencias. Sus *hashtags* representantes cercanos al centroide son *parkjhooninbangkok*, *plltheperfectionists*, *kocaeliyiseviyorum*, *portivalledupar*, *lulalivresabadosdv*, *persiangoddessz*, *lifestylecafe*, *quedateconmigo*, *palantepresidente*, *adelleasofea*.

Entre los *hashtags* más notables se encuentran los relacionados con programas de televisión y eventos deportivos televisados, como *bbb20* y *afazenda12*, que fueron ampliamente comentados en el período. Además, se mencionan eventos deportivos de gran audiencia, como la *ucl* (*UEFA Champions League*) y el *sbliii* (*Super Bowl LIII*), así como la *copaamérica* (Copa América).

También se hacen eco de movimientos sociales y temas políticos relevantes en *hashtags* como *endsars*, que se refiere a un movimiento en Nigeria y *venezuela*, que se relaciona con

los eventos ocurridos en ese país en 2019. La preocupación por el cambio climático se refleja en el *hashtag climatechange*, que también fue ampliamente discutido en relación con las noticias.

Dentro de este cluster, destaca la presencia de más de 100 *hashtags* relacionados con aldub, lo que sugiere un interés considerable en este tema. Además, el *hashtag bts*, que se relaciona con la popular banda surcoreana BTS, se encuentra entre los más frecuentes y está vinculado a diversos aspectos, como celebraciones, premiaciones, lanzamientos de canciones y videos nuevos. Los *hashtags* en este grupo parecen estar relacionados con eventos y tendencias específicas que surgieron durante el período en cuestión, lo que dificulta la definición de una categoría general para todos ellos.

Cluster 6

Este cluster se compone de 7,791 *hashtags* que, en su mayoría, son genéricos y no contienen dígitos. Representan una amplia variedad de temas y palabras clave que abarcan desde palabras en español hasta palabras en inglés. A pesar de la diversidad, hay algunas tendencias notables en este cluster. Los *hashtags* cercanos al centroide corresponden a *retreat*, *opening*, *hotrod*, *crush*, *instahealth*, *bailing*, *dune*, *awake*, *shit*, *sage*

En primer lugar, se mencionan palabras relacionadas con el bienestar y la salud, como *instahealth* y como *retreat*, que sugiere temas de salud y estilos de vida. También se observa una serie de palabras relacionadas con perros, como *maltese*, *cockerspaniel*, *doglife*, *dogoftheday* e *ilovemydog*. Esto sugiere una comunidad de amantes de los perros que comparten contenido relacionado con sus mascotas en las redes sociales.

Además, se encuentran *hashtags* relacionados con equipos de fútbol, como *fcbarcelona*, *fc-bayern* y *sevillafc*, lo que indica que los usuarios están discutiendo y compartiendo contenido sobre estos equipos y sus partidos. Por último, se observan algunos *hashtags* con errores tipográficos, como *photograghy*, *photografy* y *photograpy*, lo que sugiere que los usuarios pueden haber cometido errores al escribir estos *hashtags* en sus publicaciones.

Los *hashtags* en este grupo parecen reflejar intereses y pasatiempos diversos entre los usuarios de las redes sociales.

CAPÍTULO 4

ANÁLISIS DE RESULTADOS Y DISCUSIÓN

En el caso de la primera muestra, se opta por seleccionar la mayor cantidad de clústeres posible debido a su tamaño reducido. Esto nos brinda una ventaja significativa, ya que nos permite analizar en profundidad el comportamiento de los hashtags. Además, esta estrategia nos permite acercarnos a las categorías previamente mencionadas por los autores en 1.3.

En cuanto al análisis de componentes principales (PCA), se observa que tanto para la muestra 1 como para la muestra 2 se obtiene el mismo número de componentes principales para trabajar. Incluso cuando se utiliza UMAP después de reducir el conjunto de datos con PCA, los mismos atributos destacan, lo que conduce a la selección de la misma cantidad de componentes principales en ambos casos.

Sin embargo, al determinar el valor óptimo para k , el número de clusters, se encuentra una diferencia notable en una de las heurísticas, específicamente en el método de Davies-Bouldin. Para la muestra 1, este método sugiere la existencia de 17 clústeres como el valor óptimo, lo cual fue seleccionado. En contraste, para la muestra 2, el método de Davies-Bouldin arroja solo 3 clusters como el valor óptimo.

Al realizar un análisis exclusivo con UMAP en ambas muestras, se selecciona la misma cantidad de clusters. En ambos casos, tanto el estadístico de gap como el índice de Davies-Bouldin indican que el valor óptimo para k es de 7 clusters, y esta cantidad es la que se selecciona, ya que dos métodos coinciden en esta elección.

Esperar resultados distintos al utilizar dos técnicas de reducción de dimensiones, una lineal y otra no lineal, es completamente razonable, dado que estas técnicas tienen enfoques y capacidades diferentes. No obstante, lo que resulta sorprendente es que los resultados no estén muy alejados uno del otro. Además, es evidente que representan estructuras sustancialmente diferentes.

La cantidad de clusters seleccionada para cada muestra se puede observar de forma resumida en la tabla 6

Tabla 6: Resultados cantidad de clusters

Cantidad de clusters		
	Muestra 1	Muestra 2
PCA	17	9
UMAP	7	7
PCA+UMAP	9	6

En cada una de las muestras y experimentos realizados, se obtuvieron valores específicos

para las medias y varianzas de cada atributo en los clusters resultantes. La media representa el valor promedio de un atributo dentro de un cluster, donde un valor alto indica que los puntos dentro de ese cluster tienden a tener un valor más alto en ese atributo. Por otro lado, la varianza mide la dispersión de los datos con respecto a la media. Cuando la varianza es baja, significa que los valores de la característica están agrupados cerca de la media, lo que indica que el atributo es consistente dentro del cluster. En contraste, una alta varianza sugiere que los valores están dispersos, lo que puede indicar diferencias significativas entre ellos dentro del cluster. Si un cluster presenta valores de media extremadamente altos o bajos para un cierto atributo, esto podría señalar que esa característica es distintiva para ese cluster en particular.

En términos generales, los atributos que tienden a destacar en cuanto a la media son la frecuencia de uso, la cantidad de co-ocurrencias, la cantidad de *hashtags* que co-ocurren y la duración temporal en la que se mantienen en tendencia. Estos atributos indican que los clusters que muestran valores elevados en estas características tienden a ser más activos y atraer una mayor atención en las redes sociales. Sin embargo, los valores de longitud del *hashtag* tienden a mantenerse similares en la mayoría de los casos, ya que, como se mencionó previamente, los *hashtags* más largos suelen ser menos populares y, por lo tanto, tienden a tener valores de media más bajos.

En cuanto a la varianza, los mismos atributos mencionados anteriormente también tienden a presentar valores más altos. Esto sugiere que estos atributos pueden variar significativamente dentro de los clusters, lo que indica que algunos *hashtags* pueden tener una presencia más errática o variable en las redes sociales. Por otro lado, se observan valores de varianza muy bajos en atributos como la fecha máxima de uso, la frecuencia máxima y el coeficiente de variación. Esto sugiere que estos atributos tienden a mantenerse constantes o uniformes dentro de los clusters.

Es importante destacar que en algunos casos, cuando se encontraron una gran cantidad de temas locales en los clusters, se observó un índice de varianza bajo en la cantidad de países. Esto indica que, en esos casos, los *hashtags* locales suelen ser consistentes en cuanto a su alcance geográfico.

En el mundo de los *hashtags*, aquellos que muestran una alta co-ocurrencia, ya sea en términos de cuántos *hashtags* se asocian o con qué frecuencia lo hacen, suelen ser los más frecuentes. Esto se debe a que al estar constantemente en tendencia, las personas tienden a usarlos repetidamente para aumentar su visibilidad, independientemente de si están relacionados con el tema en discusión. Además, a menudo se utilizan con fines de spam, ya sea para promocionar productos o servicios o simplemente para inundar las conversaciones con mensajes automatizados que aprovechan todas las tendencias del momento.

Mediante el análisis de las características y la vectorización de palabras, podemos comprender el comportamiento de los *hashtags*, centrándonos principalmente en su frecuencia y las co-ocurrencias. A pesar de no contar con el contexto completo de los tweets durante la vectorización de palabras, observamos que, independientemente de esto, logramos identificar

categorías coherentes y similitudes semánticas. Esto sugiere que la co-ocurrencia desempeña un papel significativo en el comportamiento de los *hashtags*. Nuestro algoritmo aprendió del comportamiento de los *hashtags* sin necesidad de considerar las palabras, el contexto u otros factores. Este fenómeno es especialmente evidente en nuestra primera muestra, que al ser pequeña, nos permitió analizar con mayor detalle cada uno de los clusters.

Esto nos lleva a concluir que, en base a nuestros resultados empíricos, la co-ocurrencia de los *hashtags* es una característica relevante para el estudio. Esto podría deberse a que se utilizan para aumentar la visibilidad en las redes sociales o porque tienen una similitud contextual. La vectorización de palabras también desempeñó un papel importante, ya que nos permitió abordar la falta de contexto completo, especialmente en una red social donde se emplean modismos conocidos solo por la comunidad. Esto se reflejó en uno de los clusters, donde los *hashtags* parecían ser comprensibles solo para un grupo específico, al igual que si fueran palabras conocidas por todos. Siguiendo la misma línea, como menciona [Dela Rosa *et al.*, 2011], los *hashtags* pueden proporcionar pistas sobre el tema completo del tweet. A partir del análisis del *hashtag* y su comportamiento, es posible identificar el contexto en el que se basa el tweet

Durante el período que abarca desde 2019 hasta 2020, se produjo un evento sin precedentes: la pandemia de COVID-19, que afectó a todo el mundo y generó incertidumbre, medidas de salud pública, restricciones de viaje y otros impactos en la sociedad y la economía. En las redes sociales, la pandemia generó una amplia discusión, y se crearon *hashtags* específicos para fomentar estas conversaciones. Estos *hashtags* se utilizaron a nivel global, y se observó que surgieron diversas variantes que incluían nombres de lugares, características específicas y emociones relacionadas, así como *hashtags* relacionados con medidas de seguridad.

Este *hashtag* logró posicionarse como el más tuiteado durante ese período, lo que ilustra claramente el impacto de los eventos en las redes sociales y la capacidad de los *hashtags* para reflejar el momento presente. En cuanto a los clusters, en dos de los experimentos, se logró identificar clusters que contenían dos o tres elementos, como “covid19” y “coronavirus”, lo que sugiere que, aunque estos *hashtags* podrían compartir características similares con otros, están más estrechamente relacionados, ya sea por la fecha en que surgieron, su popularidad máxima o su frecuencia de uso.

Dado que la primera muestra de datos era relativamente pequeña, fue posible identificar con mayor claridad las categorías. En el experimento que aplicó PCA, se identificaron diversas categorías que, gracias al análisis de palabras clave, también revelaron similitudes semánticas o de uso conjunto, lo que facilitó la distinción entre categorías como viajes, propiedades, comida, entretenimiento y otros *hashtags* relacionados con eventos actuales.

Al analizar más allá de lo observado en otros experimentos, se llegó a la conclusión de que se pueden considerar dos categorías principales: los *hashtags* recurrentes y los *hashtags* de tendencia o ad hoc. Dentro de los *hashtags* recurrentes, se identificaron patrones temporales, como eventos anuales fijos (por ejemplo, el Super Bowl, los Oscars y los Grammys) y *hashtags* relacionados con fechas que se utilizan periódicamente. Los *hashtags* anuales mos-

traron un patrón de aproximadamente un año de duración, comenzando cerca de la fecha indicada y terminando un año después. Otros *hashtags* recurrentes, más genéricos, se relacionaron con actividades y estados de ánimo y se mantuvieron vigentes durante todo el período analizado debido a su atemporalidad y versatilidad.

Por otro lado, los *hashtags* generados espontáneamente, como los relacionados con el COVID-19, no necesariamente abarcan todo el período de estudio, ya que su aparición puede ser tardía en comparación con los recurrentes. Los programas de televisión y otros eventos puntuales generan *hashtags* específicos en el momento en que se transmiten las ediciones correspondientes. Los movimientos sociales también generan *hashtags* de tendencia, ya que surgen de manera impredecible en respuesta a eventos actuales, al igual que los desastres naturales. Estos *hashtags* pueden aparecer y desaparecer de forma repentina. Estas dos categorías principales se alinean con investigaciones previas citadas.

La elección de valores elevados para el número óptimo de clusters surge debido a la falta de un valor “real” que permita categorizar los *hashtags* de manera precisa. A partir de las observaciones, se puede notar la existencia constante de dos categorías principales en los *hashtags*: los recurrentes, ampliamente reconocidos por la comunidad, y los ad hoc, que surgen en respuesta a eventos impredecibles, generando conversaciones espontáneas. Estas dos categorías más amplias pueden subdividirse en diversas áreas temáticas que se reflejan en los distintos clusters identificados. Curiosamente, estos clusters tienden a converger hacia las mismas dos categorías principales, que abarcan temas como deportes, entretenimiento, noticias, movimientos sociales y políticos, viajes, tiempo libre, estilos de vida, entre otros.

Los resultados y la elección del número de clusters para cada muestra y experimento proporcionan una variedad de valores en las métricas de calidad de los clusters. En general, estos valores sugieren que muchos de los clusters se encuentran próximos o se superponen en cierta medida, lo que podría indicar la posibilidad de combinar algunos de ellos para formar clusters más amplios y distintos de los demás. Este enfoque podría beneficiarse de un refinamiento de los hiperparámetros, en particular, la cantidad óptima de clusters para el algoritmo de K-Means. Este proceso de ajuste podría ayudar a lograr una partición de clusters más clara y efectiva, lo que podría tener un impacto positivo en la interpretación y aplicación de los resultados.

A continuación se mostrará dos tablas comparativas con los resultados de distintas métricas de validación de calidad de clusters, para la muestra 1 (Tabla 7) y muestra 2 (Tabla 8), como el método de la silueta, el índice de Davies-Bouldin y el de Calinski-Harabasz, siendo estos dos últimos utilizados como métricas para la determinación del número óptimo de clusters.

Tabla 7: Resultados para muestra 1

Muestra 1			
Métrica	PCA	PCA+UMAP	UMAP
Silueta	0.12	0.32	0.26
Davies-Bouldin	1.93	1.05	1.19
Calinski-Harabasz	44.43	671.17	458.64

Tabla 8: Resultados para muestra 2

Muestra 2			
Métrica	PCA	PCA+UMAP	UMAP
Silueta	0.14	0.27	0.33
Davies-Bouldin	1.78	1.17	0.93
Calinski-Harabasz	2836.54	12871.55	12879.98

A modo de contexto de resultados, se debe saber que el método de la silueta considera un rango que va desde -1 a 1, donde un valor cercano a 1 indica que los clusters se encuentran bien separados y para el caso contrario, valores cercanos a -1, indican mala calidad de clusters; cuando se encuentra cercano a 0, significa que estos se superponen. Para el caso de Davies-Bouldin un valor más bajo indica una mejor calidad de clusters y por último, para Calinski-Harabasz, indica una mejor agrupación mientras más alto el valor de índice.

En la primera muestra, se observa que la mezcla de métodos de reducción de dimensionalidad PCA+UMAP muestra resultados notables en términos de métricas de calidad de clustering. Específicamente, presenta el valor más alto en la métrica de la silueta, el valor más bajo en Davies-Bouldin y el valor más alto en Calinski. Esto sugiere que la combinación de un método lineal (PCA) seguido de uno no lineal (UMAP) puede resultar en una mejor calidad de clustering cuando se carece de la verdad absoluta en la evaluación.

En la segunda muestra, el método UMAP muestra un desempeño destacado en términos de métricas de calidad de clustering. En particular, obtiene el valor más alto en Silhouette, el valor más bajo en Davies-Bouldin y el valor más alto en Calinski. Estos resultados indican que, en esta muestra, otros métodos de reducción de dimensionalidad podrían beneficiarse de la consideración de un número de clusters diferente, ya que el número seleccionado podría conducir a superposición entre los clusters, lo que sugiere una falta de distancia significativa entre ellos. Estas observaciones destacan la importancia de elegir sabiamente el método de reducción de dimensionalidad y el número de clusters, ya que pueden tener un impacto significativo en la calidad del clustering, especialmente en ausencia de etiquetas verdaderas.

En el análisis, encontramos que los clusters revelan características comunes en los *hashtags*. Estos patrones y temas emergentes proporcionan una visión más clara de cómo los *hashtags* se relacionan entre sí. Nuestros hallazgos reflejan la diversidad y relevancia de los *hashtags*.

Los valores atípicos identificados desempeñan un papel fundamental en la tipología, ya que representan *hashtags* relevantes. Es importante reconocer que existen limitaciones en nuestro análisis, como las suposiciones realizadas y las restricciones de datos. Se podría decir que este análisis nos acerca a nuestro objetivo de elaborar una tipología de *hashtags* basada en características, proporcionando una visión más profunda de cómo se agrupan y relacionan los *hashtags*.

CAPÍTULO 5

CONCLUSIONES

En el marco de esta investigación, se ha llevado a cabo un estudio centrado en el análisis de hashtags, específicamente a partir de sus características. Para llevar a cabo este análisis, se eligió Twitter como fuente de datos, y se estableció un período de extracción que abarca desde 2019 hasta 2020, inclusive.

En una primera evaluación, se ha observado que, aunque el período de datos seleccionado es considerable, podría ser beneficioso ampliar este rango temporal para poder identificar patrones más sólidos, especialmente aquellos de naturaleza temporal.

En cuanto al proceso de procesamiento de datos, se utilizó el enfoque ETL (extracción, transformación y carga) en la plataforma Talend. Este enfoque demostró ser altamente efectivo, ya que se manejaban múltiples archivos de diversas fuentes, cada uno con su propia estructura, lo que presentaba un desafío en términos de la limpieza y consolidación de datos. Talend permitió obtener datos limpios, estandarizados y centralizados en una única fuente, lo que generó un nuevo conjunto de datos a partir de los millones de tweets obtenidos, considerando sólo los relevantes para el estudio que debía cumplir con la geolocalización y contener hashtags, no era excluyente.

Posteriormente, se extrajeron una serie de características relacionadas con los hashtags, algunas directamente y otras derivadas de las primeras. Estas características se consideraron apropiadas en el contexto del análisis de hashtags. Sin embargo, se reconoce la necesidad de profundizar en la identificación de nuevas características específicas de los hashtags, que puedan enriquecer aún más nuestro análisis.

La vectorización de palabras emergió como una característica esencial en este estudio, desempeñando un papel fundamental en la revelación de relaciones semánticas que permitieron la distinción de subcategorías dentro de los hashtags. Al transformar las palabras en vectores numéricos, esta técnica proporcionó una manera efectiva de representar la información semántica contenida en los hashtags, lo que a su vez posibilitó una segmentación más detallada y precisa. Esta representación vectorial permitió no solo agrupar los hashtags por similitud semántica, sino también identificar conexiones y patrones subyacentes en los datos que, de otra manera, podrían haber pasado desapercibidos.

Debido a la alta dispersión presente en el conjunto de datos, se tomó la decisión de trabajar con los valores atípicos. Estos valores atípicos desempeñaron un papel crucial en la investigación y el análisis, ya que representan hashtags excepcionales y relevantes para el estudio. Esto se debe a la amplia variabilidad en la frecuencia de uso de estos hashtags, especialmente bajo el filtro aplicado previamente.

Una vez identificados los valores atípicos, se generaron dos subconjuntos a partir del conjun-

to de datos principal. Uno de estos subconjuntos incluye exclusivamente los valores atípicos, mientras que el otro, de mayor tamaño, contiene los valores atípicos mencionados en el primer subconjunto, así como una variedad de valores cercanos al conjunto de datos completo, aunque siguen siendo distintivos y destacados en términos de frecuencia de uso.

A partir de esta segmentación, se aplicaron dos métodos de reducción de dimensiones, uno lineal (PCA) y otro no lineal (UMAP). Ambos métodos resultaron altamente útiles tanto por separado como en conjunto, ya que proporcionaron resultados consistentes y similares. Posteriormente, se aplicó el algoritmo de k-means para crear clusters y analizarlos en detalle. El objetivo era descubrir similitudes y diferencias dentro de estos clústeres para avanzar hacia el objetivo general de esta investigación, que es desarrollar una tipología de hashtags basada en sus características. La combinación de la identificación de valores atípicos, la reducción de dimensiones y el análisis de clusters resultó esencial para acercarse al objetivo. A través de un análisis minucioso de las medias y las varianzas de atributos clave presentes en los clusters resultantes, se ha logrado identificar patrones y tendencias intrigantes en la manera en que los hashtags se agrupan y se entrelazan.

Este estudio arroja luz sobre dos categorías fundamentales de hashtags: los recurrentes, que gozan de un amplio reconocimiento y se mantienen en constante circulación a lo largo del tiempo; y los de tendencia o ad hoc, que emergen de manera impredecible en respuesta a acontecimientos actuales y efímeros. Estas dos categorías se reflejan de manera sorprendente en los clusters identificados, y hemos observado que estos conglomerados tienden a converger hacia estas dos categorías primordiales, que abarcan una amplia gama de temas, que van desde deportes y entretenimiento hasta noticias de actualidad, movimientos sociales y mucho más.

A través de este estudio, comprendimos que, bajo ciertos comportamientos y atributos, los hashtags pueden categorizarse en las dos grandes categorías mencionadas anteriormente. Sin embargo, la libertad de creación de hashtags introduce complejidades adicionales, ya que estos pueden ser neologismos, palabras de otros idiomas o términos oscuros, entre otras posibilidades. También encontramos desafíos en cuanto a la diversidad de idiomas presentes en los hashtags, lo que dificultó la categorización, especialmente cuando se buscaba no limitar el análisis a un conjunto de idiomas específico.

Al analizar en profundidad, independientemente del tamaño del conjunto de datos, se pueden identificar dos categorías principales relacionadas con la longevidad y la espontaneidad de los hashtags. Estas dos categorías principales se refieren a los hashtags recurrentes, ampliamente reconocidos y consistentes con el tiempo, y a los hashtags ad hoc, que emergen de manera espontánea en respuesta a eventos o situaciones actuales. Esto implica que existen hashtags que perduran en el tiempo y son conocidos por las comunidades en las redes sociales, mientras que otros son generados de manera instantánea según el contexto y pueden desvanecerse rápidamente en popularidad o persistir si logran captar la atención de la comunidad.

Los hashtags, en esencia, son una representación que abarca la diversidad de temas que

se discuten en las redes sociales. Estos pueden englobar una amplia gama de asuntos, desde cuestiones de salud y entretenimiento hasta programas de televisión, manifestaciones, movimientos sociales, tecnología y mucho más. Dada su naturaleza de libre creación, cada usuario puede diseñar hashtags según su preferencia, incluso generando nuevas tendencias en la comunidad de Twitter y marcando el rumbo de las conversaciones.

Además, se ha subrayado la trascendental importancia de la selección de métodos de reducción de dimensionalidad y la cantidad de clusters en la calidad de los resultados obtenidos. Los hallazgos indican que enfoques diversos pueden derivar en resultados notoriamente distintos, lo que acentúa la necesidad de realizar una elección meticulosa de estos parámetros en futuros análisis.

Es importante reconocer que existen limitaciones en nuestro análisis, como las suposiciones realizadas y las restricciones de datos. Este análisis nos acerca a nuestro objetivo principal de elaborar una tipología de hashtags basada en características, proporcionando una visión más profunda de cómo se agrupan y relacionan los hashtags

En el futuro, sería valioso considerar agregar una característica adicional que refleje la cantidad de usuarios que utilizan un hashtag. Esto proporcionaría información sobre la formación de comunidades en torno a temas específicos y ayudaría a comprender si un hashtag es utilizado de manera exclusiva por un usuario individual o si logra expandirse a un público más amplio. Además, la segmentación de hashtags podría ser un enfoque interesante, aunque sería un desafío considerable debido a la diversidad de formas y estilos de los hashtags. Adicionalmente, este estudio podría complementarse con un análisis de las redes de co-ocurrencia entre hashtags y usuarios para obtener una comprensión más completa del comportamiento dentro de la plataforma de Twitter.

Otra dirección para futuras investigaciones podría ser la ampliación del período de tiempo analizado, lo que permitiría un análisis más exhaustivo de la evolución de los hashtags efímeros. También se podría considerar la posibilidad de aplicar filtros adicionales, como la clasificación de idiomas, especialmente en lenguas romances e inglés, que son más comunes en hashtags populares.

Siguiendo con la perspectiva para investigaciones futuras, se podría explorar la generación de predicciones de categorías a partir de las características de los hashtags mediante algoritmos de aprendizaje automático supervisado. También sería valioso investigar diferentes algoritmos de clustering y ajustar los hiperparámetros para evaluar su impacto en los resultados. Esto podría incluir la variación en el número de iteraciones, la aplicación de algoritmos de clustering jerárquicos y basados en densidad, entre otros enfoques.

En un mundo cada vez más digital y conectado, la comprensión de cómo los hashtags se agrupan y evolucionan en las redes sociales adquiere una relevancia fundamental para descifrar la dinámica de la comunicación en línea y las tendencias que rigen nuestra sociedad.

Este trabajo final representa un primer alcance respecto a la forma en que los hashtags se

agrupan y conectan en las redes sociales, y proporciona perspectivas que podrían ser útiles para futuras investigaciones.

ANEXOS

5.1. Características que aportan al PCA, muestra 1

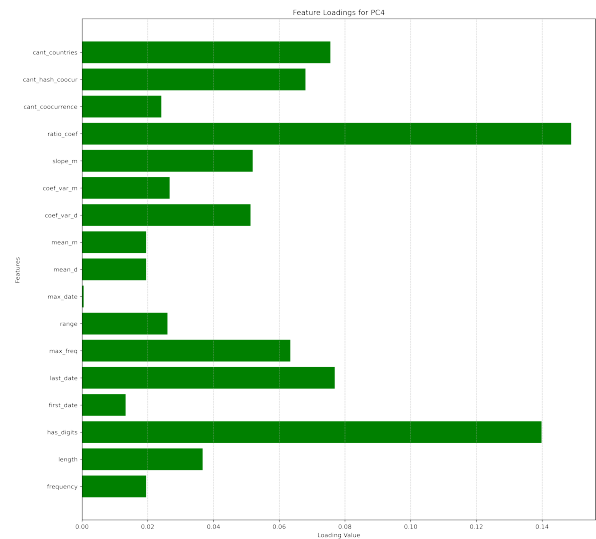


Figura 48: Aportes de cada atributo en el componente 4
Fuente: Elaboración a partir de código

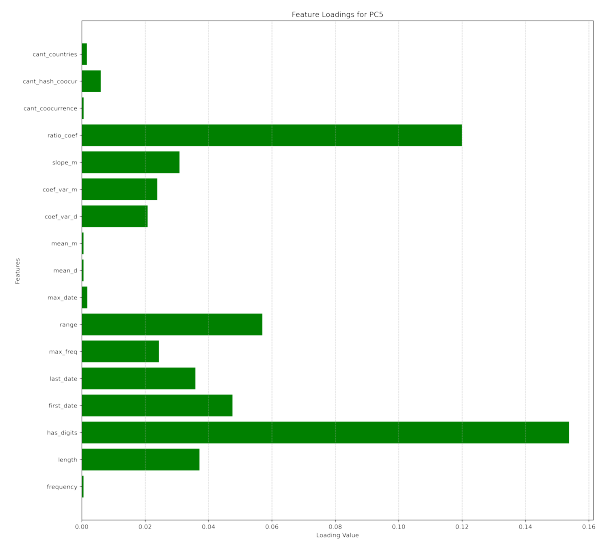


Figura 49: Aportes de cada atributo en el componente 5
Fuente: Elaboración a partir de código

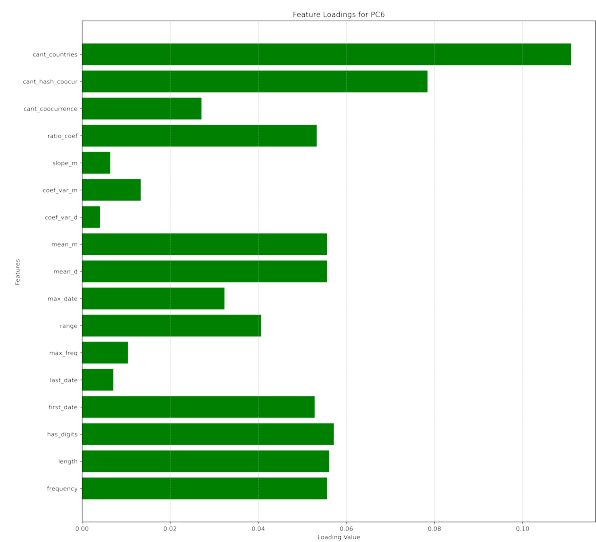


Figura 50: Aportes de cada atributo en el componente 6
Fuente: Elaboración a partir de código

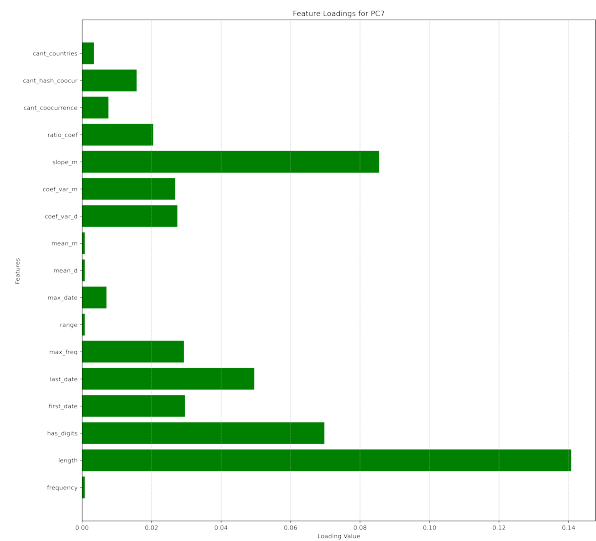


Figura 51: Aportes de cada atributo en el componente 7
Fuente: Elaboración a partir de código

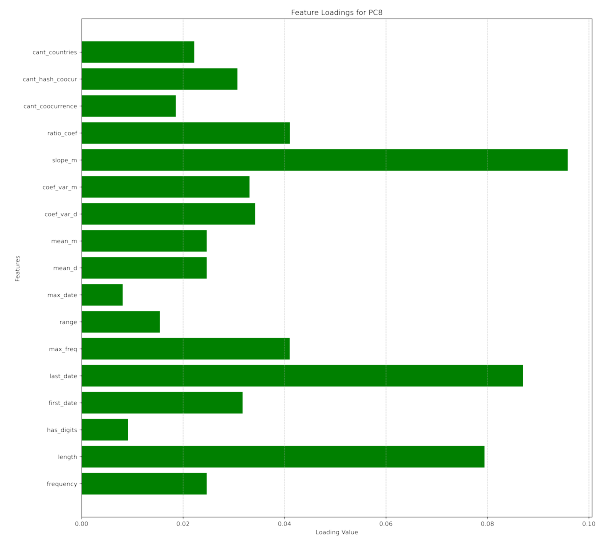


Figura 52: Aportes de cada atributo en el componente 8
Fuente: Elaboración a partir de código

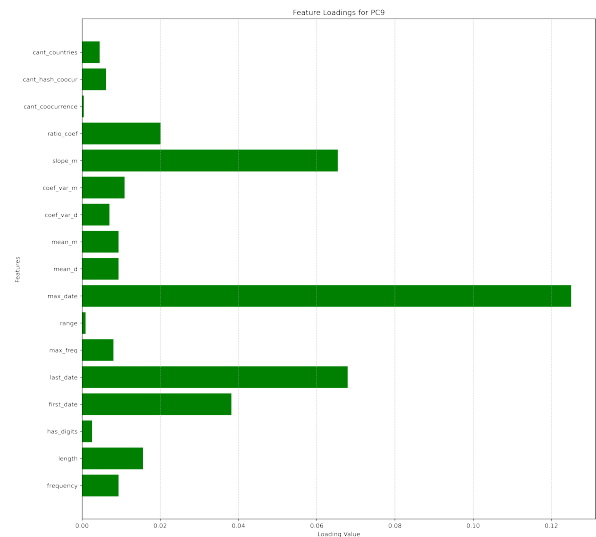


Figura 53: Aportes de cada atributo en el componente 9
Fuente: Elaboración a partir de código

5.2. Características que aportan al PCA, muestra 2

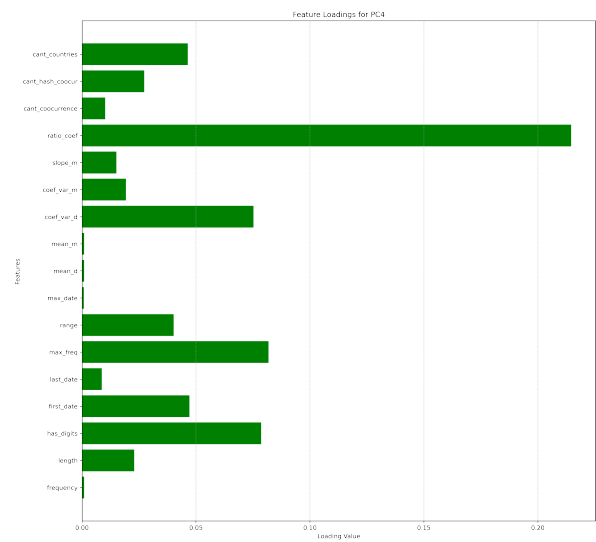


Figura 54: Aportes de cada atributo en el componente 4
Fuente: Elaboración a partir de código

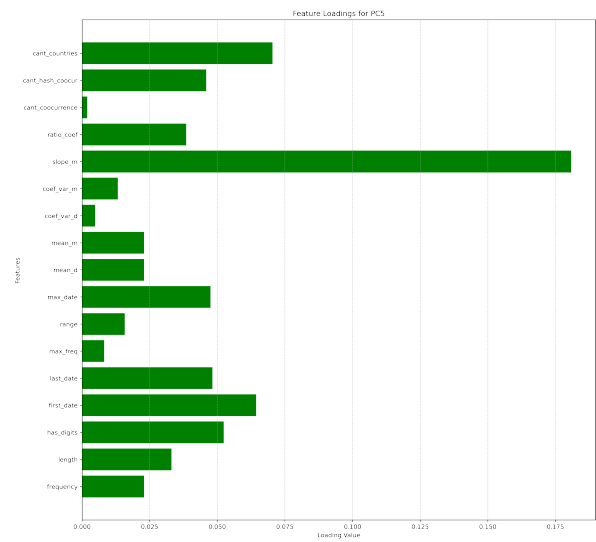


Figura 55: Aportes de cada atributo en el componente 5
Fuente: Elaboración a partir de código

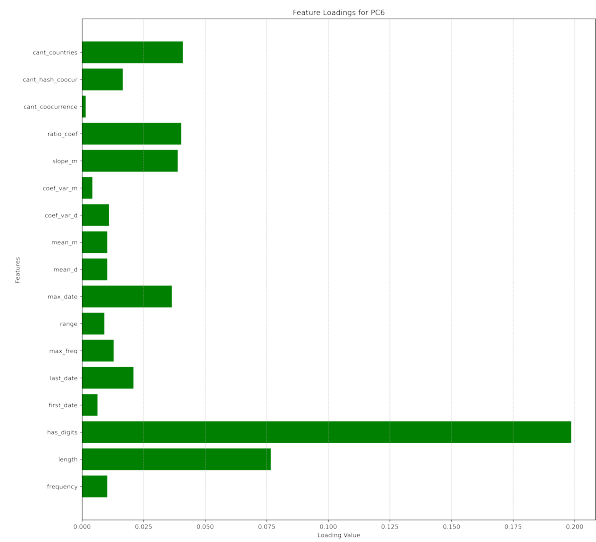


Figura 56: Aportes de cada atributo en el componente 6
Fuente: Elaboración a partir de código

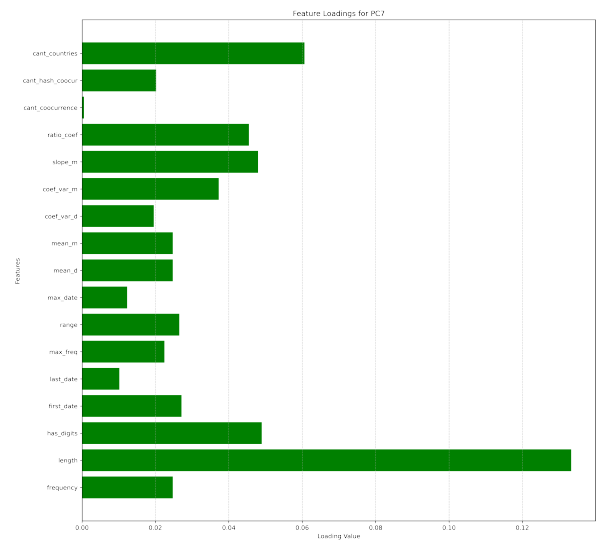


Figura 57: Aportes de cada atributo en el componente 7
Fuente: Elaboración a partir de código

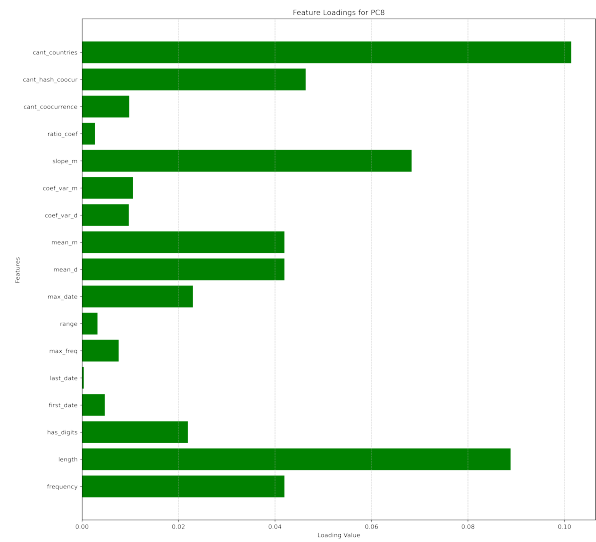


Figura 58: Aportes de cada atributo en el componente 8
Fuente: Elaboración a partir de código

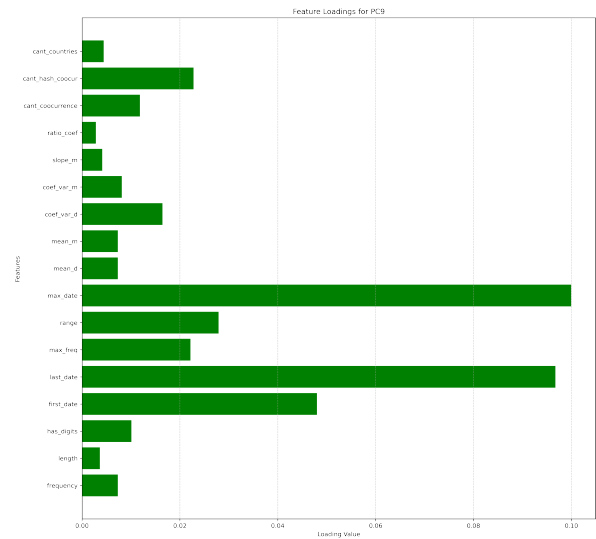


Figura 59: Aportes de cada atributo en el componente 9
Fuente: Elaboración a partir de código

REFERENCIAS BIBLIOGRÁFICAS

- [arc, 2011] (2011). Internet Archive: Digital Library of Free & Borrowable Books, Movies, Music & Wayback Machine — archive.org. <https://archive.org/details/twitterstream>. Accessed: 2023-10-23.
- [Bastos et al., 2021] Bastos, M., Travitzki, R., y Puschmann, C. (2021). What sticks with whom? Twitter follower-followee networks and news classification. *Proceedings of the International AAAI Conference on Web and Social Media*, 6:6–13.
- [Bruns y Burgess, 2011] Bruns, A. y Burgess, J. (2011). The use of twitter hashtags in the formation of ad hoc publics. En *Proceedings of the 6th European Consortium for Political Research (ECPR) General Conference 2011*, pp. 1–9.
- [Bruns et al., 2016] Bruns, A., Moon, B., Paul, A., y Münch, F. (2016). Towards a typology of hashtag publics: a large-scale comparative study of user engagement across trending topics. *Communication Research and Practice*, 2:20–46.
- [Bruns y Stieglitz, 2013] Bruns, A. y Stieglitz, S. (2013). Towards more systematic twitter analysis: Metrics for tweeting activities. *International Journal of Social Research Methodology*, 16:91–108.
- [Dela Rosa et al., 2011] Dela Rosa, K., Shah, R., Lin, B., Gershman, A., y Frederking, R. (2011). Topical clustering of tweets. En *3rd Workshop on Social Web Search and Mining, SWSM 10*.
- [DeMasi et al., 2016] DeMasi, O., Mason, D. J., y Ma, J. (2016). Understanding communities via hashtag engagement: A clustering based approach. En *Proceedings of the Tenth International AAAI Conference on Web and Social Media*, pp. 102–111.
- [Espin-Riofrio y Zumba, 2021] Espin-Riofrio, C. y Zumba, J. (2021). Métodos de extracción de comentarios de la red social twitter para uso en procesamiento de lenguaje natural. *Polo del Conocimiento*, 6:104–123.
- [Han et al., 2011] Han, J., Kamber, M., y Pei, J. (2011). *Data Mining: Concepts and Techniques (3rd Ed.)*. Morgan Kauffman, Massachusetts.
- [La Rocca, 2020] La Rocca, G. (2020). La fuerza de un signo. perspectivas teóricas para el análisis de los hashtags #. *Barataria. Revista Castellano-Manchega de Ciencias Sociales*, (27):46–61.
- [La Rocca y Boccia Artieri, 2023] La Rocca, G. y Boccia Artieri, G. (2023). Interpreting the changeable meaning of hashtags: Toward the theorization of a model. *Frontiers in Sociology*, 7.
- [Laniado y Mika, 2010] Laniado, D. y Mika, P. (2010). Making sense of twitter. En *The Semantic Web – ISWC 2010*, pp. 470–485, Berlin, Heidelberg. Springer.

- [Ma *et al.*, 2013] Ma, Z., Sun, A., y Cong, G. (2013). On predicting the popularity of newly emerging hashtags in twitter. *Journal of the American Society for Information Science and Technology*, 64:1399–1410.
- [Mahesh, 2020] Mahesh, B. (2020). Machine learning algorithms - a review. *International Journal of Science and Research*, pp. 381–386.
- [Maćkiewicz y Ratajczak, 1993] Maćkiewicz, A. y Ratajczak, W. (1993). Principal components analysis (PCA). *Computers & Geosciences*, 19(3):303–342.
- [McInnes *et al.*, 2018] McInnes, L., Healy, J., y Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints*.
- [McInnes *et al.*, 2018] McInnes, L., Healy, J., Saul, N., y Grossberger, L. (2018). Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861.
- [Messina, 2016] Messina, A. C. (2016). Groups for twitter; or a proposal for twitter tag channels. Blog: <https://factoryjoe.com/2007/08/25/groups-for-twitter-or-a-proposal-for-twitter-tag-channels/>. Accessed: 2023-10-10.
- [Pöschko, 2011] Pöschko, J. (2011). Exploring twitter hashtags. *arXiv preprint arXiv:1111.6553*.
- [Romero *et al.*, 2011] Romero, D., Meeder, B., y Kleinberg, J. (2011). Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter. En *WWW '11: Proceedings of the 20th international conference on World wide web*, pp. 695–704.
- [Sedhai y Sun, 2017] Sedhai, S. y Sun, A. (2017). An analysis of 14 million tweets on hashtag-oriented spamming. *Journal of the Association for Information Science and Technology*, 68:1638–1651.
- [Stilo y Velardi, 2016] Stilo, G. y Velardi, P. (2016). Hashtag sense clustering based on temporal similarity. *Computational Linguistics*, 43:1–32.
- [Wang *et al.*, 2016] Wang, R., Liu, W., y Gao, S. (2016). Hashtags and information virality in networked social movement: Examining hashtag co-occurrence patterns. *Online Information Review*, 40:17.
- [Wang y Zheng, 2014] Wang, Y. y Zheng, B. (2014). On macro and micro exploration of hashtag diffusion in twitter. En *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, pp. 285–288.
- [Weng y Menczer, 2015] Weng, L. y Menczer, F. (2015). Topicality and impact in social media: Diverse messages, focused messengers. *PloS One*, 10:e0118410.

[Zak, 2013] Zak, E. (2013). How Twitter's Hashtag Came to Be — wsj.com. <https://www.wsj.com/articles/BL-DGB-29742#>. [Accessed 10-05-2023].