UNIVERSIDAD DE SALAMANCA

Departamento de Estadística

Máster en Análisis Avanzado de Datos Multivariantes y Big Data

Trabajo Fin de Máster

El HJ Biplot con Clustering K-Means como técnica de análisis estadístico textual:

Exploración de una muestra de Literatura Científica en Psicología

Mariya Ilieva Palikarska

Tutores: Daniel Caballero Juliá y José Luis Vicente Villardón

2023

El HJ Biplot con Clustering K-Means como técnica de análisis estadístico textual:

Exploración de una muestra de Literatura Científica en Psicología

Mariya Ilieva Palikarska

Tutorizado por Daniel Caballero Juliá y José Luis Vicente Villardón

Universidad de Salamanca

2023

Dpto. de Estadística
Universidad de Salamanca

Don José Luis Vicente Villardón

Profesor Contratado Doctor del Departamento de Estadística de la Universidad de Salamanca

CERTIFICA que **D./D.ª Mariya Ilieva Palikarska** ha realizado en la Universidad de Salamanca, bajo su dirección, el trabajo que para optar título de **Máster en Análisis Avanzado de Datos Multivariantes y Big Data** presenta con el título ***El HJ Biplot con Clustering K-Means como técnica de análisis estadístico textual: exploración de una muestra de Literatura Científica en Psicología***, autorizando expresamente su lectura y defensa.
Y para que conste, firma el presente certificado en Salamanca a 18 de septiembre de 2023.

Nombre/Firma

VICENTE VILLARDON JOSE LUIS - 07839395Y

Firmado digitalmente por VICENTE VILLARDON JOSE LUIS - 07839395Y
Fecha: 2023.09.19 12:58:51 +02'00'

AGRADECIMIENTOS

A mis padres por su apoyo constante.

A los amigos que estuvieron animándome para completar este proyecto.

A mis tutores y profesores, por su orientación y paciencia.

ÍNDICE

o Resumen:

Este trabajo presenta una propuesta metodológica para el análisis estadístico de textos académicos. Utilizando el método Reinert como punto de partida, se desarrolla una alternativa metodológica más rica basada en la combinación del HJ biplot con técnicas de clasificación.

La muestra de estudio consiste en 180 abstracts de artículos académicos seleccionados a través de una búsqueda bibliográfica exhaustiva en el campo de la psicología, específicamente centrada en el impacto de las redes sociales en la atención y la ansiedad. Estos abstracts se transformaron en tablas léxicas y se sometieron, tras una ponderación previa, a análisis multivariante usando el HJ Biplot para reducir la dimensionalidad de los datos. Además, se llevó a cabo una exploración de diferentes técnicas de clustering, concretamente DBSCAN, K-Means y clustering jerárquico con método de ward. Se exploraron las ventajas e inconvenientes de las diferentes técnicas. Los resultados más destacados y relevantes se obtuvieron a través del método de clustering K-Means.

Palabras clave: HJ Biplot, clúster K-means, topic modeling

o Abstract:

This work presents a methodological proposal for the statistical analysis of academic texts. Using the Reinert method as a starting point, a richer methodological alternative is developed, based on the combination of HJ biplot with classification techniques.

The study sample consists of 180 abstracts from academic articles selected through an exhaustive bibliographic search in the field of psychology, specifically focused on the impact of social networks on attention and anxiety. These abstracts were transformed into lexical tables and underwent multivariate analysis using the HJ biplot after prior weighting to reduce the dimensionality of the data. In addition, an exploration of different clustering techniques, specifically DBSCAN, K-Means, and Hierarchical clustering with ward linkage, was conducted. The advantages and disadvantages of the different techniques were explored. The most outstanding and relevant results were obtained through the K-Means clustering method.

Keywords: HJ Biplot, K-means clustering, topic modeling

1. INTRODUCCIÓN

En el contexto actual, la disponibilidad de datos textuales está aumentando exponencialmente, generándose masas de datos tan vastas que a menudo es difícil-si no imposible-extraer estructuras de sentido fácilmente comprensibles para un ser humano. Por tanto, emplearlos para desarrollar y aumentar nuestros conocimientos requiere de herramientas para su procesamiento automatizado.

El ámbito académico, si bien se caracteriza por la publicación y manejo de textos altamente especializados, rigurosos en su contenido y estructurados, no es una excepción. A menudo, tanto el estudiante como el académico experimentado tienen que hacer una detallada exploración del mayor número de textos posibles publicados en su área de estudio a través de la búsqueda bibliográfica para asegurar la calidad de su propio trabajo, la comprensión adecuada y la relevancia de sus ideas. Exploración que frecuentemente es tan productiva que requiere de una gran cantidad de tiempo y recursos cognitivos.

Por ello se hace patente la necesidad de algoritmos y sistemas de procesamiento de datos textuales que faciliten y automaticen estas tareas.

El presente trabajo pretende explorar el uso combinado del HJ Biplot, una técnica de visualización basada en la reducción de la dimensionalidad de los datos, con una posterior aplicación de técnicas de clústering para extraer patrones temáticos y estructuras de significado a partir de datos textuales provenientes de abstracts de artículos académicos.

El caso usado para ilustrar la aplicación de estas técnicas es una revisión bibliográfica sobre el impacto del uso de las redes sociales en la atención y la ansiedad contextualizado desde el campo de la psicología de la salud.

1.1 El método ALCESTE

Este trabajo parte de y se inspira en el método de análisis textual ALCESTE (Del francés: Analyse des Lexèmes Coocurrents dans les Enoncés Simples d'un Texte)- Reinert (1990). Se trata de un método informatizado para el análisis de textos que está bien establecido y tiene un amplio uso en el campo de las ciencias sociales. Fue creado por Max Reinert en el marco de la investigación sobre el desarrollo de métodos de análisis de datos lingüísticos, iniciada por

Jean-Paul Benzécri. en Francia. Reinert tomó el análisis de Correspondencias desarrollado por Benzécri (1973) y lo aplicó a los datos textuales, añadiendo el paso a mayores de un análisis de clúster jerárquico.

Según De Alba (2004, p20) el proceso de aplicación del método desarrollado por Reinert se despliega a través de varias fases:

Durante la primera fase, el software descompone el texto en segmentos contextuales básicos, distingue entre palabras esenciales y palabras relacionales. Las palabras relacionales son las que sirven a la construcción sintáctica de la frase -artículos, conjunciones, preposiciones, pronombres. Luego elimina los sufijos de las primeras, posibilitando así la generación de una tabla de datos binarios con los segmentos contextuales en filas y las unidades léxicas en columnas. Este se convierte en el procedimiento fundamental para llevar a cabo los análisis estadísticos consecutivos.

En la segunda fase, se efectúa una categorización de los segmentos contextuales basada en la similitud o disimilitud de sus vocablos. Un enfoque doble de categorización se emplea para verificar la constancia de las categorías: se efectúa una primera categorización con segmentos contextuales de un tamaño específico (por ejemplo, 10 unidades léxicas) y luego otra con un tamaño superior (12 unidades léxicas). La categorización es considerada estable si esta variación en el tamaño del vocabulario por segmento contextual elemental no altera la estructura ni la distribución de las categorías ni su contenido.

La tercera fase detalla las categorías obtenidas mediante diversas técnicas estadísticas adicionales: selección del vocabulario particular de cada categoría, elección de los segmentos contextuales representativos de cada categoría, cálculo de repeticiones de segmentos por categoría, de formas abreviadas y sus sufijos, realización de un análisis factorial de correspondencias basado en la primera categorización, y la ejecución de un clúster jerárquico para cada categoría, entre otros.

En conjunto, podemos extraer que, estadísticamente, ALCESTE tiene como piedras angulares, sin obviar todo el preprocesamiento léxico de los datos que se ha descrito, el análisis de correspondencias (múltiples) como técnica de reducción de la dimensionalidad y una posterior agrupación aglomerativa (clúster jerárquico ascendente) como técnica de clasificación. Esto le permite extraer temáticas o “mundos lexicales” como los nombra el propio Reinert (1993), en la práctica generando la posibilidad de extraer observaciones sobre el significado de los segmentos textuales. Sin embargo, es un método no exento de limitaciones.

Por ejemplo, Dalud-Vincent (2011) presenta resultados de ALCESTE que no necesariamente coinciden con la visión del sociólogo sobre la estructura léxica real que hay detrás de los datos y es susceptible a sobreinterpretación. También, ALCESTE presenta pocas opciones para el investigador más allá de quedarse con la única agrupación que le ha presentado, que además está basada en frecuencias y siendo el caso, como ha explorado Caballero (2011) de que hay muchas palabras con una frecuencia total muy alta que no tienen por qué aportar ningún insight sobre los temas reales detrás del corpus.

La pregunta que está en la raíz de este trabajo es ¿se podría aplicar la misma metodología de análisis de textos con una reducción de la dimensionalidad y un posterior proceso de clasificación para obtener resultados similares o mejores, usando otras técnicas alternativas de reducción de la dimensionalidad y clasificación? ¿Qué ventajas tendría esto y cómo se compararía con el análisis planteado por Reinert?

1.2 Métodos BIPLOT en el contexto del análisis de datos textuales

Gower y Hand (1995) se refieren al Biplot como “el análogo multivariante del gráfico de dispersión”. La ventaja más saliente de los gráficos biplot es que permiten la exploración conjunta de variables e individuos proyectándolos sobre los ejes de mayor variabilidad.

El término "bi" en "biplot" se refiere a esta capacidad de mostrar simultáneamente información sobre las filas y las columnas en un solo gráfico. Esta característica distingue a los biplots de otros tipos de representaciones gráficas que se centran exclusivamente en visualizar filas o columnas por separado. Los biplots son una herramienta valiosa en el análisis de datos multivariantes y pueden proporcionar una comprensión muy intuitiva de las relaciones y patrones presentes en la matriz de datos original.

Para explicar por qué las técnicas biplot son una alternativa al análisis de Correspondencias, debemos mencionar la Descomposición en Valores Singulares (DVS). Es una técnica fundamental y básica en álgebra lineal que consiste en una factorización que descompone una matriz en tres componentes esenciales: una matriz unitaria izquierda, una matriz diagonal de los llamados valores “propios” o singulares y una matriz unitaria derecha. (Osuna, 2006). Todas las matrices tienen una descomposición en valores singulares, incluso si no son cuadradas. (Trefethen y Bau,1997). Esto permite la reducción de la dimensionalidad de tablas de datos de

gran tamaño. La creciente complejidad de los datos con los que es habitual tratar actualmente ha llevado a la aplicación frecuente de técnicas fundamentadas en la DVS. Una es la ya mencionada de Benzécri en su trabajo en el campo del Análisis de Correspondencias.

Otra de las técnicas basadas en la DVS es el Biplot clásico, introducido inicialmente por Gabriel (1971), que él propone como un método sumamente útil para explorar y representar la estructura de conjuntos de datos amplios. Mediante una adaptación de la Estadística Descriptiva y la Geometría Elemental, esta técnica se ajusta a la naturaleza de los datos multivariantes. Como ya dijimos acerca de los biplots en general, sus pilares de interpretación se basan en gráficos que posibilitan la representación simultánea de "individuos" (para el caso presente, documentos) caracterizados por "variables" (para el caso presente, palabras) en el estudio. Esta aproximación visual de las relaciones que subyacen en los conjuntos de entidades exploradas se lleva a cabo representando gráficamente una matriz rectangular de datos, abordando los elementos a través de marcadores ligados a las filas y a las columnas, en un espacio de dimensiones inferiores al rango de la matriz. No se limita únicamente a situaciones donde las filas representan individuos y las columnas reflejan variables, sino que se extiende a otros contextos, como tablas de doble entrada, matrices de covarianza o tablas léxicas.

Parafraseando a Osuna, (2006) la interpretación de los Biplots se cimienta en los siguientes conceptos geométricos sencillos:

-La similitud entre individuos se refleja en la inversa de su distancia.

-La longitud y los ángulos de los vectores que representan variables se interpretan en términos de variabilidad y covariabilidad, respectivamente.

-Las relaciones entre filas y columnas se desentrañan mediante proyecciones

Una de las mayores ventajas de un Biplot como técnica de representación es la posibilidad de visualizar intuitivamente las variables, los individuos y las relaciones entre los mismos, estimando con la dirección y el ángulo la covarianza y con la cercanía y la proyección su similitud. Un Biplot resulta fácil de entender gráficamente, aunque esté dirigido a un público no familiarizado con la estadística multivariante o la geometría.

En 1986, Galindo Villardón propone el HJ Biplot como una alternativa simétrica con calidad de representación igual para filas y columnas (variables e individuos, o en nuestro caso, palabras y segmentos textuales) en contraposición a los Biplot propuestos anteriormente por Gabriel, que

proporcionaban una mayor calidad de representación a las columnas (GH Biplot) o a las filas (JK Biplot).

Esta es la técnica que propone Osuna (2006) para el análisis textual, en su tesis doctoral. Esta tesis es otra de las principales inspiraciones para esta línea de trabajo e investigación en datos textuales. Explora el HJ Biplot y el Biplot robusto como alternativas al análisis de correspondencias en el análisis textual de datos.

En 2011, Caballero, usando también el HJ biplot pero en el contexto del análisis de grupos de discusión, hace una aportación a esta línea de investigación de Osuna desde un enfoque cualitativo, introduciendo la codificación de sentido y referencia en el análisis de textos y el índice de caracterización, una ponderación que relativiza la frecuencia con la que aparecen las palabras dentro de cada segmento a la palabra más utilizada en el mismo en vez de la frecuencia total. Su aportación es una propuesta para dar solución a las limitaciones cualitativas de emplear frecuencias en la exploración de significados cualitativos en textos. El análisis textual presenta mucho ruido porque, a pesar de algunas de las técnicas de preprocesamiento que hemos mencionado como la lematización o la eliminación de palabras que no aportan significado como preposiciones, conjunciones, etcétera, las palabras más frecuentes siguen sin ser necesariamente las que recogen mejor el significado de un determinado discurso o segmento textual.

Posteriormente, el equipo de la Universidad de Salamanca ha seguido explorando las técnicas Biplot en el análisis textual empleando el meta-Biplot y el MANOVA Biplot en estudios que tienen como objetivo precisamente el análisis de textos bibliográficos (Caballero, Galindo y García, 2017). Estas últimas publicaciones son el estado actual del arte del que parte directamente el trabajo presente.

1.3 Técnicas de clasificación

Como establecimos en el apartado correspondiente al método de análisis textual de Reinert, la estructura estadística en la que se inspira este trabajo es el uso de una técnica de reducción de dimensionalidad posteriormente refinada con un análisis de clúster. Por tanto, cabe introducir brevemente (se explorará y justificará más en la sección de metodología) las técnicas de clúster a las que se hará referencia a lo largo de este trabajo.

En el trabajo original de Reinert, la técnica de clúster elegida es el clúster jerárquico. Es un método que busca agrupar los datos en una estructura jerárquica, donde los clústeres más pequeños se combinan en clústeres más grandes de manera sucesiva; los dos grupos con la menor distancia entre grupos se fusionan sucesivamente hasta que todos los objetos han sido fusionados en un único grupo (Ao et al., 2005). Este enfoque permite visualizar tanto clústeres individuales como agrupaciones de clústeres en diferentes niveles de resolución. El método jerárquico se basa en la similitud o distancia entre los puntos de datos y ofrece una comprensión intuitiva de la estructura de los datos. Diversas estrategias, como el enlace completo, el enlace único y el enlace promedio, determinan cómo se calcula la similitud entre clústeres y cómo se fusionan. En el trabajo presente, se emplea el enlace de Ward, que busca minimizar la varianza dentro de los grupos al combinar elementos, lo que tiende a generar grupos compactos y bien definidos en función de la distancia euclídea (Ward,1963). El clúster jerárquico es ampliamente utilizado en diversas disciplinas, como la biología, la ciencia de datos y la minería de texto, para explorar relaciones complejas en datos heterogéneos.

Otra de las técnicas exploradas es el algoritmo DBSCAN. En cuanto al algoritmo DBSCAN (Density-Based Spatial Clustering of Applications with Noise), fue propuesto por Martin Ester, Hans-Peter Kriegel, Jörg Sander y Xiaowei Xu en 1996. Es una técnica que se destaca por su capacidad para identificar clústeres basados en la densidad espacial. A diferencia de los métodos tradicionales que requieren una especificación previa del número de clústeres, DBSCAN determina automáticamente el número de clústeres y detecta áreas de alta densidad rodeadas por regiones de baja densidad, por lo cual resulta especialmente interesante para este trabajo. Este enfoque permite identificar clústeres de formas y tamaños irregulares en datos ruidosos y con variaciones de densidad, lo cual se ajusta al análisis de datos textuales, que suele presentar mucho ruido.

Por último, y tal vez con los resultados más interesantes, se ha empleado el algoritmo de clústering de K-medias. El algoritmo de k-medias es una técnica que busca encontrar k centroides óptimos para minimizar la suma de las distancias cuadradas entre cada punto de datos y el centroide más cercano. El procedimiento de k-medias es fácilmente programable y es económicamente eficiente desde el punto de vista computacional, por lo que es factible procesar muestras muy grandes en un ordenador (MacQueen, 1967).

El algoritmo converge de manera eficiente hacia una solución siempre (todos los individuos pertenecerán a un clúster), pero su resultado depende de la elección inicial de los centroides y asume una distribución esférica de los clústeres.

1.4 Caso de estudio: efectos de la tecnología en la atención y la ansiedad

Los efectos de la tecnología en la atención y la ansiedad han sido objeto de un creciente interés desde una perspectiva psicológica. La posesión universal de dispositivos electrónicos y las constantes comunicaciones digitales inmediatas (habitualmente hablamos de cientos al día) están planteando cuestiones importantes en lo que respecta al funcionamiento cognitivo y emocional de las personas.

En términos de atención, la revolución digital ha introducido distracciones constantes que pueden dificultar la capacidad de concentración. La notificación constante de mensajes y el acceso a múltiples fuentes de información pueden fragmentar la atención y hacer que las personas tengan dificultades para mantener el enfoque en una tarea específica. Esto puede llevar a un fenómeno conocido como "atención dividida" (multitasking), donde la capacidad para procesar información de manera profunda y sostenida se ve comprometida.

El paradigma actual en Psicología propone que cuando intentamos prestar atención “simultánea” a varias tareas, lo que ocurre realmente es que hay una rápida redirección de la atención que tiene un coste para el sistema ejecutivo, particularmente cuando las tareas son complejas (Lam, Vartanian & Hollands, 2022). De esta manera, el uso frecuente de redes sociales se asocia a síntomas de TDAH (Trastorno por déficit de atención e hiperactividad) (Fisher, Hopp, Chen & Weber, 2023).

En cuanto a la ansiedad, la tecnología puede desencadenar y exacerbar trastornos como la ansiedad social y la ansiedad por la comparación constante. Las redes sociales, en particular, pueden generar ansiedad al fomentar la comparación con los demás y promover una versión idealizada de la vida de las personas particularmente entre adolescentes y gente joven. (Mann & Blumberg, 2022). Además, la presión para mantenerse "conectado" en todo momento (FOMO, Fear of Missing Out) puede contribuir a la ansiedad y al agotamiento digital (Eitan & Gazit, 2023).

Hay que reconocer que, aunque la tecnología puede tener efectos negativos, también ofrece oportunidades para la intervención y el autocuidado. Las aplicaciones de seguimiento del bienestar emocional y la terapia en línea se han masificado desde la pandemia de COVID 19 y

son ejemplos de cómo la tecnología puede utilizarse para abordar y gestionar los problemas de atención y ansiedad.

Para comprender completamente estos efectos y desarrollar estrategias de intervención efectivas, se están llevando a cabo investigaciones continuas en este campo. Estas investigaciones van a ser nuestra muestra objetivo a la hora de conducir el presente análisis.

2. OBJETIVOS

Se busca desarrollar un enfoque que permita a los investigadores y académicos analizar de manera eficiente grandes cantidades de información textual en su área de estudio, reduciendo así la carga de trabajo.

Objetivos generales:

1. **-Identificación de Temas y Patrones**: identificar temas o patrones en los datos textuales sin requerir un conocimiento previo específico sobre el contenido de los textos
2. **-Usabilidad y aplicación:** aunque el presente trabajo tiene un enfoque metodológico y estadístico, una de las consideraciones que valoraremos a la hora de decantarnos por una técnica u otra será el que sea fácil de aplicar para profesionales académicos de todos los campos a la hora de estudiar resultados de búsquedas bibliográficas.
3. **-Comparación de técnicas**: explorar las ventajas y limitaciones de diferentes enfoques de análisis y encontrar alternativas al ya mencionado método Alceste para extraer núcleos temáticos de grandes grupos de textos.

Objetivos específicos de la aplicación práctica:

4. -Explorar las tendencias en la literatura dentro de una muestra de artículos de investigación de la última década acerca del impacto de las redes sociales en la ansiedad y la atención.
5. -Ilustrar las técnicas usadas adecuadamente.

3. MATERIALES Y MÉTODOS

3.1 Revisión de literatura y selección de artículos

La presente investigación se ha llevado a cabo con el propósito de explorar y analizar exhaustivamente el estado del arte en relación con los efectos del uso de la tecnología en la ansiedad y la atención. El procedimiento de búsqueda bibliográfica se llevó a cabo en la base de datos de ScienceDirect, considerada una fuente confiable de literatura científica, utilizando una serie de operadores booleanos con el fin de refinar y enfocar la búsqueda de manera precisa.

La estrategia de búsqueda empleada consistió en la combinación de los siguientes términos clave: "anxiety" (ansiedad), "attention deficit" (déficit de atención) y "social media" (redes sociales). Estos operadores booleanos (con AND o Y) se utilizaron para identificar estudios que abordaran simultáneamente estos tres elementos en su tema de investigación. Esta primera fase de búsqueda arrojó un conjunto inicial de **1,121 artículos** relacionados con los términos de interés, todos ellos publicados **entre 2013 y 2023**.

Para garantizar la relevancia y la coherencia con los objetivos de la investigación, se procedió a refinar aún más la selección de artículos. Primero, se restringió la búsqueda al campo de la Psicología, con el fin de centrarse en la comprensión de los efectos y las implicaciones de la tecnología en la salud mental y cognitiva. Además, se filtraron los resultados según los tipos de publicación, priorizando los artículos de investigación y las revisiones sistemáticas como las categorías más pertinentes para obtener información sintetizada.

Una vez aplicados estos filtros nos quedamos con 305 artículos. Se revisaron estos nuevos resultados de búsqueda de manera manual, con el objetivo de eliminar aquellos artículos que no se alinearan con el objeto de estudio. Se excluyeron investigaciones que se enfocaban exclusivamente en la ansiedad o el Trastorno por Déficit de Atención e Hiperactividad (TDAH) sin tener en cuenta la dimensión de las redes sociales. También se excluyeron revisiones generales de trastornos mentales no relacionados (como el autismo o la misofonía, entre otros) y aquellos artículos que habían surgido en la búsqueda debido a la coincidencia de una sola palabra (por ejemplo, "social") sin abordar adecuadamente la temática propuesta.

Finalmente, cuatro artículos fueron excluidos debido a la falta de resúmenes (abstracts) válidos que proporcionan la información esencial para hacer nuestro análisis de texto. Como resultado de este proceso de búsqueda y selección, se ha conformado **una muestra final**

compuesta por n= 179 artículos para representar una base sólida para el análisis del estado del arte en este área de estudio.

3.2 Preprocesamiento

Por motivos de formato y manejabilidad, los datos bibliográficos de los artículos seleccionados han sido exportados en formato Excel (.xlsx). Después se ha creado un archivo con tres columnas: el año de publicación, un identificador para cada artículo y el abstract correspondiente. Se ha empleado Microsoft Word para dar el formato adecuado que se necesita para que cada abstract sea identificado por separado. Ya que el software usado posteriormente para el análisis de frecuencias presenta incompatibilidades con Microsoft Office, se han guardado los datos en formato UTF-8.

3.3 Tablas léxicas con el software IraMuTeQ y aplicación del método Reinert

IraMuTeQ (Interfaz de R para el Análisis Multidimensional de los Textos y Cuestionarios) es un software libre basado en los lenguajes R y Python desarrollado por Pierre Ratinaud en 2009 para el análisis de datos cualitativos utilizado principalmente en investigaciones textuales y análisis de contenido.

Es la herramienta empleada en este trabajo para la creación de tablas léxicas, que muestran la relación entre las palabras y su frecuencia de aparición en un corpus de texto o conjunto de documentos. En nuestro caso, los corpus serían cada uno de los abstracts.

Cuando Iramuteq crea tablas léxicas, realiza un proceso de análisis para extraer información relevante sobre las palabras y términos presentes en un conjunto de documentos.

Iramuteq cuenta primero la frecuencia de cada palabra en el conjunto de documentos. En nuestro caso se ha establecido un número límite de 10 o más palabras para proceder a incluírlas en la tabla léxica.

Además de contar las palabras, Iramuteq calcula las frecuencias específicas de cada palabra en diferentes documentos, en este caso en cada abstract.

Funciona por diccionarios; en nuestro caso, el de inglés.

3.3.1 Palabras activas y suplementarias

Iramuteq puede eliminar palabras “vacías”, (las preposiciones, conjunciones, pronombres, demostrativos, números y abreviaturas) para centrarse en las palabras más significativas para el análisis léxico como verbos y sustantivos. En este caso hemos activado esta opción y sólo nos hemos quedado con las palabras “activas”.

3.3.2 Lematización

Otro proceso que aplica este software es poner todas las palabras en masculino singular. Esto se conoce como lematizar (de esta forma, las palabras “sueñas” y “soñaría” contarían para la misma variable, “soñar”).

Finalmente, nos hemos quedado con **una tabla de 255 palabras** como filas y los 179 artículos como columnas con la frecuencia absoluta de aparición de cada palabra en cada abstract de artículo.

Tras la generación de esta matriz, se procede a la aplicación del Método Reinert; como se ha explicado anteriormente en la sección de introducción, este método consiste en un análisis factorial de correspondencias múltiples y de un clúster jerárquico.

Hemos obtenido como resultado de este análisis 5 Clústeres**.**

3.4 Valor de caracterización

A la tabla con frecuencias absolutas se ha aplicado el valor de caracterización propuesto por Caballero (2011).

Este autor parte de un problema específico y muy característico del tipo análisis y de datos que vamos a aplicar a la tabla léxica.

Para empezar, aunque estamos aplicando un análisis cuantitativo, la frecuencia no puede ser siempre una medida absoluta fiable de la relevancia de una palabra si lo que estamos intentando obtener es técnicamente, un resultado cualitativo. Como veremos más adelante, en la sección de resultados, la palabra más frecuentemente repetida en nuestro ejemplo es “estudio”, algo que nos aporta poco o nada de significado a la hora de entender cuál es el tema o el sentido del conjunto de textos con el que estamos tratando.

Para continuar, el tipo de datos que estamos tratando presenta una gran cantidad de ceros, puesto que una palabra que se emplea con frecuencia en uno de nuestros textos no tiene siquiera por qué existir en otro. Por ejemplo, un estudio que tiene en cuenta el impacto del uso del móvil en la calidad del sueño va a tener esta última palabra con una frecuencia mucho mayor que otro que esté hablando del ciberbullying. Por tanto, es necesario “relativizar” matemáticamente cada palabra a su contexto.

Para abordar esta necesidad, Caballero propone ignorar los ceros en cada columna y considerar todas las demás puntuaciones como candidatas. Luego, compara las filas en cada una de las columnas y se designa como "característica" a la puntuación más alta.

Parafraseando el trabajo original de Caballero (2011), cada coordenada en la tabla léxica se pondera no en función de su margen, sino en función de la puntuación máxima. Y esta transformación de los datos tiene varias propiedades importantes:

- Los valores resultantes se escalan entre 0 y 1, donde 0 representa la ausencia de la palabra en un código específico y 1 representa que la palabra es la más característica de ese código.
- Cuando la puntuación original es máxima tanto en fila como en columna, la puntuación transformada es igual a 1.
- Cuando la puntuación original es máxima en la columna, pero no en la fila, se relativiza el valor original utilizando el máximo de la fila correspondiente.

Esta es la fórmula del valor de caracterización:

$$f'_{np} = \frac{f_{np}}{\sqrt{max_i}\sqrt{max_j}}$$

Siendo que f_{np} es la puntuación y max_i es el máximo de la fila y max_j es el máximo de la columna.

Para aplicar esta ponderación a los datos de este trabajo se ha usado el paquete base de R. Para explorar los datos se han generado varias variantes de la tabla léxica general creada con IraMuTeQ; con las palabras como filas y los artículos como columnas y viceversa, con las frecuencias absolutas y ponderadas con el valor de caracterización y con y sin el año de

publicación como variable nominal. En todo caso, los resultados finales se han hecho en base a la tabla ponderada por el valor de caracterización.

3.5 Análisis HJ Biplot

Como ya se ha expuesto en la introducción, el HJ-Biplot (Galindo, 1986), es una técnica de exploración y representación gráfica de datos multivariantes que permite una calidad de representación máxima y simétrica para filas y columnas.

En este caso se ha aplicado la técnica mediante el paquete MultBiplotR para el lenguaje R, desarrollado por Villardón. Ya que en algunas aplicaciones el software considera automáticamente a las filas como individuos y las columnas como variables, se ha aplicado el HJ biplot a dos variantes de nuestros datos:

- Una matriz de frecuencias ponderadas por el valor de caracterización con las palabras como filas (individuos) y los artículos como columnas (variables) que a partir de ahora llamaremos abreviadamente Matriz P
- Una matriz de frecuencias ponderadas por el valor de caracterización con los artículos como filas (individuos) y las palabras como columnas (variables) que a partir de ahora llamaremos abreviadamente Matriz A.

Posteriormente, se han extraído las coordenadas fila (que representarían los individuos) de cada uno de los dos análisis HJ biplot realizados. Se ha aplicado un proceso de normalización a las puntuaciones y se ha creado una matriz de distancias correspondiente a cada uno de los dos HJ Biplot para poder proceder a los posteriores análisis de clúster. Estas matrices de distancias a su vez denominaremos Matriz de distancias P y Matriz de distancias A respectivamente.

3.6 Estimaciones preliminares al análisis de clúster

Creadas las matrices de distancias con las coordenadas de los individuos, se ha procedido a aplicar varios análisis exploratorios para determinar estadísticamente, cuál sería el número apropiado de clústeres para cada caso.

No es relevante para el objeto de estudio entrar en detalle acerca de cada una de estas técnicas, puesto que se pueden aplicar fácilmente de manera automatizada, pero sí es de vital

importancia citarlas para explicar los criterios según los cuales hemos decidido aplicar un tipo u otro de clúster.

Los clústeres jerárquicos y clústeres k-medias requieren un número predeterminado por el usuario de clústeres antes de poder aplicar los algoritmos, se ha aplicado una función del paquete de R NBclust. Este paquete proporciona casi 30 estimaciones para determinar el número óptimo de grupos para el conjunto de datos correspondiente y ofrece al usuario esquemas de agrupamiento a partir de los diferentes resultados.

Aquí está la enumeración completa:

Regla del codo, Índice de Kapila-Love (kl), Índice de Calinski-Harabasz (ch), Índice de Hartigan, Índice C de Harrell, Índice de Scott, Índice de Marriott, Índice de Tracew, Índice de Friedman, Índice de Rubin, Índice de Validación Dunn, Índice de Hubert, Índice de desviación estándar, Índice D-Index, Índice de validez de la desviación estándar entre clústers, Medida de Silueta, Índice de Duda, Índice de Pseudo-T2, Índice de Beale, Índice de Ratkowsky, Índice de Ball, Índice Biserial-puntual, Índice de Gap, Índice de Frey, Índice de McClain, Índice de Gamma, Índice de Gamma plus e Índice de Tau.

Una de las funciones del paquete devuelve un resumen de todas estas técnicas indicando según cuántas de ellas hace falta un número de clúster x. Por ejemplo, 3 técnicas indican 2 clúster, 4 indican 3 clúster, 13 indican 4 clúster, 7 indican 5 y así sucesivamente. Se selecciona la moda, particularmente si es muy elevada respecto a las otras puntuaciones como ha sido en nuestro caso.

En este tratamiento de datos hemos aplicado este proceso para:

- La matriz de distancias P respecto a un clúster jerárquico
- La matriz de distancias A respecto a un clúster jerárquico
- La matriz de distancias P respecto a un clúster k-medias
- La matriz de distancias A respecto a un clúster k-medias

Para el análisis DBSCAN, como se trata de un modelo basado en densidades, los cálculos previos no requieren establecer un número predeterminado de clústers sino un número de vecinos para establecer un clúster y un índice de distancia para establecer que dos individuos pertenecen al mismo clúster. Dicho índice se denomina épsilon.

Se ha usado un número de vecinos estándar (k=5) y se ha calculado un gráfico de distancia k (el proceso de decisión es similar a la regla del codo) para elegir épsilon

3.7 Análisis de clúster

En este trabajo se han empleado tres técnicas de clúster; DBSCAN, Jerárquico y K-medias.

Las consideraciones han sido las siguientes; por una parte los datos textuales basados en frecuencias suelen tener mucho ruido. Además, no partimos de un número determinado de temas o códigos de significado que hay que aplicar, sino que la intención es desarrollar una técnica exploratoria.

DBSCAN es útil cuando se desconoce el número de clusters y se desean clusters de diferentes formas. El clúster jerárquico normalmente no requiere un número previo, pero debido al software de aplicación (tenemos que hacerlo sobre las coordenadas del biplot) en nuestro caso sí. También es relevante porque es el elegido en el método Reinert. K-medias es eficiente y adecuado cuando se conoce el número de clústeres y los clústeres son globulares.

Se han aplicado los tres métodos para compararlos y comprobar con cuál hay resultados interesantes para los objetivos propuestos.

3.7.1 Clúster de K medias

Se ha aplicado el algoritmo de k-medias para llevar a cabo el proceso de agrupación de temas. En la fase de inicialización, se seleccionó un número “adecuado” de clústeres (k) basado en el análisis de la sección previa.

El algoritmo k-medias procede a la inicialización de centroides aleatorios como puntos de partida para la agrupación en función del número de clústeres que le hemos propuesto. El siguiente paso consistie en asignar cada punto de datos a uno de los clústeres en función de la distancia euclídea entre ellos y los centroides. Este proceso de asignación es iterativo, y en cada iteración, se recalculan los centroides tomando la media de los puntos asignados a cada clúster. Continúa iterándose el proceso de asignación y recálculo hasta que los centroides dejan de cambiar significativamente o hasta que se alcanza un número máximo de iteraciones definido previamente. Esto asegura que el algoritmo converga en una solución estable y final

(todos los individuos van a pertenecer a un clúster). En nuestro caso, el número máximo de iteraciones seleccionado por defecto es de 100.

Dado que los resultados no fueron satisfactorios para los objetivos de la investigación, el proceso se ha repetido seleccionando un número de clústeres mayor a criterio del usuario y no siguiendo las pruebas estadísticas de la sección anterior.

Este proceso se ha conducido de manera automática en la interfaz de MultbiplotR, usando como punto de partida la matriz de coordenadas/distancias de los dos HJ biplot, tanto para las palabras como variables como para los artículos como variables.

3.7. 2 Clúster Jerárquico

En el algoritmo de clúster jerárquico, inicialmente, cada dato se considera un clúster individual, y en cada iteración se fusionan los dos clusters más cercanos (Nielsen, 2016) según la métrica de la suma de sus cuadrados en nuestro caso, por el enlace elegido (criterio de Ward). Este proceso continúa hasta que todos los datos están agrupados en un único cluster. La convergencia se logra cuando se ha construido completamente la jerarquía de clusters y se obtiene un dendrograma que representa las relaciones de agrupación a diferentes niveles de similitud. El método de Ward minimiza la varianza dentro de los clusters fusionados en cada paso, lo que conduce a una estructura jerárquica donde los clusters están bien definidos y compactos en forma esférica o elipsoidal.

Se ha empleado el método de Ward (Ward, 1986) con la esperanza de minimizar la sensibilidad al ruido y a los outliers que presenta esta técnica (problema que va a encontrarse a menudo en muestras de datos parecidas a la usada).

De nuevo nos encontramos con el mismo problema: los resultados previos no cumplían con los objetivos de la investigación y se procedió a una nueva iteración del proceso, eligiendo un número de clústeres mayor a discreción del investigador en lugar de seguir los enfoques estadísticos.

Esta técnica de clúster también fue aplicada mediante la interfaz de MultbiplotR. Se utilizó otra vez como punto de partida las matrices de coordenadas/distancias de los dos HJ biplot, tanto para las palabras como variables, como para los artículos como variables.

3. 3. 7. 3 Clúster DBSCAN

El algoritmo DBSCAN (Ester, Kriegel, Sander y Xu,1996) normalmente se aplica a datos muy diferentes a los de esta muestra, como por ejemplo, datos astronómicos. Pero debido a su bajísima sensibilidad al ruido y al hecho de que no necesita de un número previo de clúster era una prueba interesante.

El algoritmo detecta automáticamente regiones densas de datos y asigna puntos a clústeres en función de su proximidad y densidad (que hemos establecido mediante los cálculos en el paso anterior). Comienza seleccionando un punto de datos aleatorio y examina su entorno para identificar otros puntos cercanos. Si encuentra suficientes puntos dentro de un radio específico (según el parámetro de distancia epsilon ε), forma un clúster y se expande iterativamente para incluir más puntos densos. Los puntos aislados se consideran ruido o atípicos. Puede iterar durante su proceso, pero el número de iteraciones no se fija de antemano como en algunos otros algoritmos de clustering.

Debido a que los resultados no fueron satisfactorios, se ha probado manualmente a repetir el proceso con valores más bajos de épsilon, hasta dar con un valor que generase varios clústeres.

Este clúster, a diferencia de los dos anteriores, se llevó a cabo en R Studio mediante el paquete dbscan.

4. RESULTADOS

4.1 Desarrollo:

4.1.1 ALCESTE

La aplicación del método de Reinert en IraMuTeQ ha dado como resultado 5 agrupaciones en el clúster jerárquico.

En la primera división principal del dendrograma tenemos los clústers 1 y 2. El clúster 1 (gris en la figura 1) reúne los términos dedicados a la metodología y se podría atribuir a que revisamos artículos científicos; “revisión”, “investigación”, “publicación”, “ciencia”, etc.

El clúster 2 (rojo en la figura 1) agrupa un núcleo temático que podría darnos una pista acerca del hecho de que estamos investigando el área de la salud: las dos primeras palabras son “salud” y “mental” (aunque sea en orden inverso). También agrupa las palabras referentes a la medicina y a la pandemia de COVID 19 (que es obviamente un tema relevante teniendo en cuenta la franja temporal de las publicaciones entre 2013 y 2023.

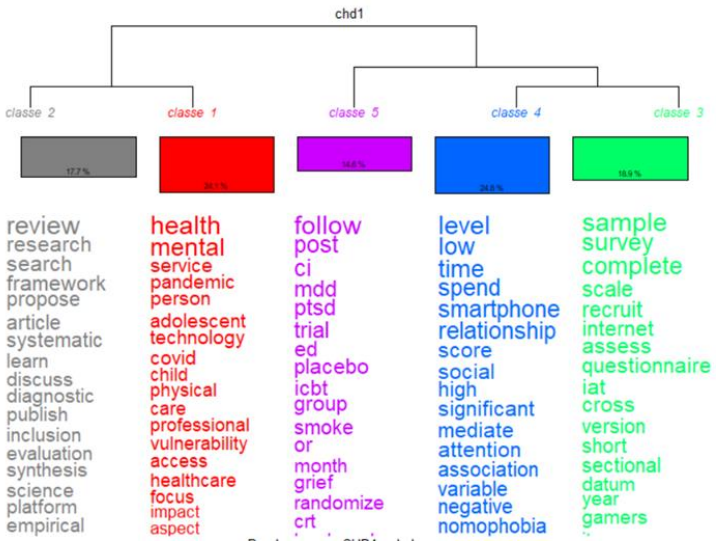

Figura 1. Clúster jerárquico producido por el método Reinert. En este trabajo se han comentado de izquierda a derecha, no por el número de la clase

En la segunda ramificación del dendrograma tenemos dos subdivisiones con el clúster tres (lila en la figura uno) por libre y los clústeres 4 y 5 juntos.

El clúster 3 agrupa varias abreviaturas diagnósticas del área de la psicología (mdd, ed ptsd), pero más allá de eso no presenta una temática tan clara como los clústeres anteriores. Los clústeres 4 y 5 (azul y verde en la figura 1) contienen algunas de las palabras clave que nos podrían orientar hacia el hecho de que la búsqueda se ha realizado acerca del impacto de la tecnología y las redes sociales en la atención, pero también mezclan esas palabras con términos metodológicos y con poco significado (“variable”, “encuesta”, “puntuación”, etcétera). Además, tampoco presentan una clara distinción entre ambos ni está claro de qué habla cada uno a diferencia del otro.

En resumen, a nivel cualitativo, los resultados ofrecidos por el método de Reinert, en esta muestra nos ofrecen:

1. Una indicación acerca del hecho de que estamos hablando de una revisión de literatura científica.
2. Un núcleo temático que nos sitúa en el área de la salud, y posiblemente de la psicología si tomamos en cuenta también el clúster 3.
3. Una cierta separación de los términos metodológicos respecto a los términos que realmente se refieren a los temas de investigación
4. Algunas pistas no del todo bien diferenciadas acerca de los temas de investigación en sí que motivaron la búsqueda.

4.1.2 HJ BIPLOT

Sin entrar aún en el análisis detallado de los temas y de la información que vamos a obtener acerca de los artículos mediante técnicas de clasificación, conviene hacer un breve resumen de las características a priori de los dos biplots que van a servir de base para nuestro análisis.

NOTA: El biplot separa las variables en función de los dos primeros ejes principales de variabilidad. A partir de este momento se hará referencia a los sectores de los biplots como primer cuadrante (altos valores de eje 2, bajos valores de eje 1), segundo cuadrante (altos valores de eje 2, altos valores de eje 1), tercer cuadrante (bajos valores de eje 2, bajos valores de eje 1) y cuarto cuadrante (bajos valores de eje 2, altos valores de eje 1).

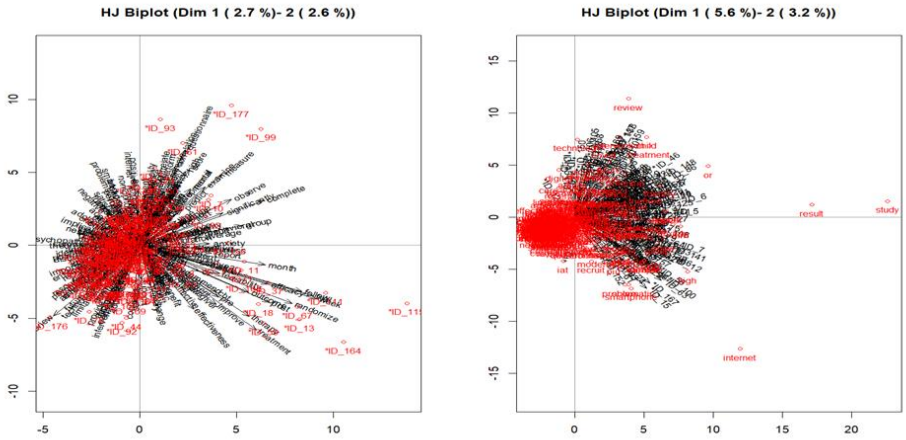

Figura 2. Izquierda: Biplot de la matriz de palabras como variables (columnas). Derecha: Biplot de la matriz de artículos como variables (columnas).

En el caso del análisis de componentes principales de la **matriz de variables = artículos**, los tres primeros ejes explican el 11.633% de la varianza; el eje uno el 5.642%, el eje 2 el 3.215% y el eje 3 el 2.776%. A partir del eje tres, los porcentajes de varianza explicada decaen muy progresivamente y la varianza acumulada de los primeros 10 ejes es de sólo 25.846%, por tanto, cuesta establecer cuántos ejes deberíamos retener exactamente en un análisis de PCA pero por motivos de representatividad en este análisis nos centraremos en los ejes 1 y 2. Esto probablemente se debe al gran número de variables.

La contribución más alta al eje 1 es de lejos de la palabra “estudio” con una contribución de 593 (cuando las próximas contribuciones más altas a este eje son en torno a 150 aproximadamente). La contribución más alta al eje 2 la tiene similarmente la palabra “resultado” (418). Esta es una ilustración de las limitaciones de hacer análisis basado en frecuencias numéricas de repetición de palabras, incluso si ponderamos por el valor de caracterización. Las palabras con contribuciones más altas son las referidas a la metodología científica (las que en el análisis Alceste pertenecían al clúster 1). Estas palabras, como estudio y resultado, también tienen contribuciones muy altas a otros ejes alternos. Mucho menos prominentes, pero aún significativas, tenemos contribuciones de palabras que reflejan los temas que han salido en la búsqueda bibliográfica (“internet”, contribución al eje 2=196, suicidio, contribución al eje 2= 111)

En conjunto, tanto en cuanto a varianza explicada, como en cuanto a contribuciones y calidad de representación, se verá que ambos análisis de componentes principales en sus resultados

numéricos resultan más bien poco interesantes, porque salvo identificar ciertos outliers como los que se ha mencionado, la cantidad de variables es tan inmensa que es muy difícil extraer conclusiones en cuanto a elementos definitorios de los ejes.

Y el objetivo del presente trabajo es extraer núcleos temáticos desde la perspectiva de personas no necesariamente familiarizadas con la estadística multivariante. Por tanto, se va a centrar la atención en la exploración visual de la representación del biplot.

Si examinamos la **Figura 3** Lo primero que llama la atención de este biplot es que en los cuadrantes 1 y 3 (a la izquierda, con los valores negativos) hay una nube de palabras que no están cerca de ninguna variable-artículo, mientras que casi todas las variables se agrupan a la derecha en los cuadrantes 2 y 4, formando una especie de “abanico” cuyos extremos son paralelos al eje 2. Los vectores más largos están en el cuadrante 2, formando un ángulo aproximado de 45 grados respecto al eje 2.

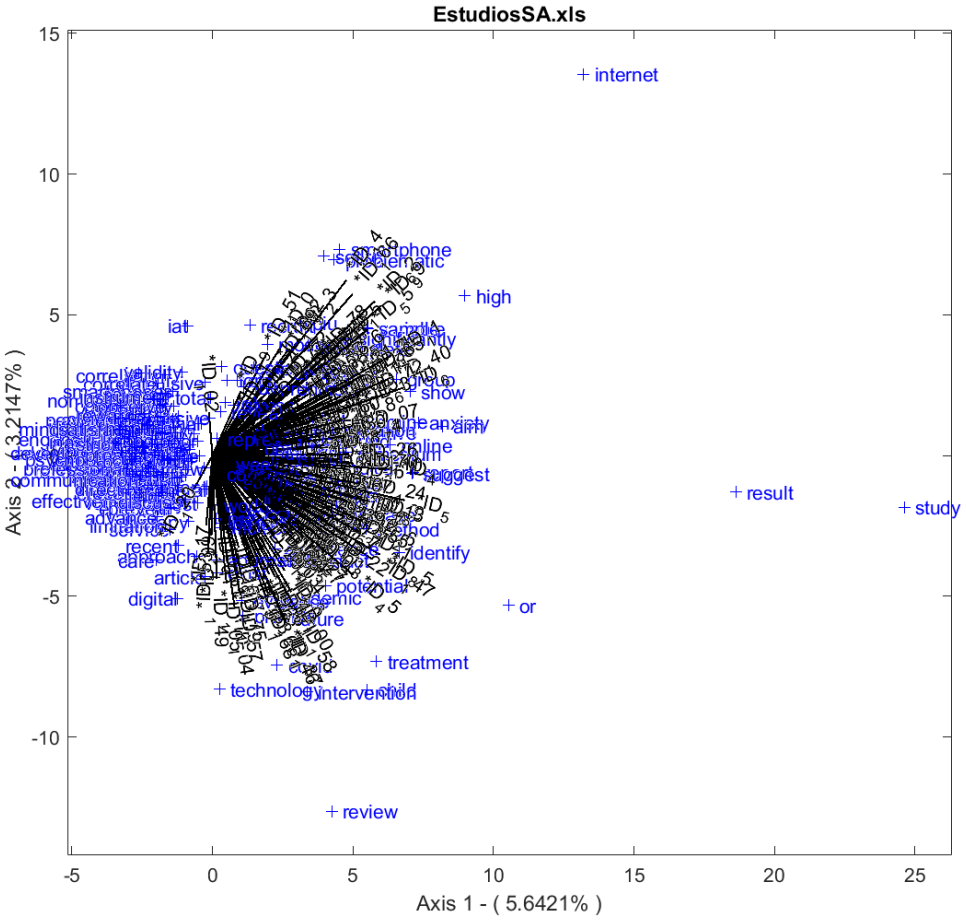

Figura 3. Biplot de variables artículos. Los vectores se agrupan en los valores positivos del eje1

Si examinamos las palabras en torno a estos vectores, encontramos las palabras “smartphone”, “problematic” y como outlier, en los valores más altos del eje 2, “internet”. Dada la gran cantidad de variables e individuos, no es pertinente hacer aseveraciones demasiado osadas, pero tras inspeccionar manualmente los artículos cuyos vectores apuntan en la dirección de esas palabras, vemos que en su mayoría tienden a hacer referencia al uso problemático de internet y de las redes sociales.

La distribución de los vectores hacia los cuadrantes 2 y 4 y la nube de lapabras a su izquierda dan a entender que hay un grupo de palabras de frecuencias totales muy bajas (probablemente específicas a artículos concretos) y otras que caracterizan más al total de la muestra. Como indicación de esto tenemos a los outliers “estudio” y “resultado” paralelos al eje 1, con valores mucho más elevados en este eje que cualquier otro individuo.

También hay algunas parejas de palabras que a nivel intuitivo parecen tener significado por ser parejas que sabiendo el contenido de la muestra de antemano en el lenguaje natural aparecerían juntas. Es muy probable que su cercanía sea fortuita, pero no obstante resultan interesantes: el ya mencionado caso de “estudio” y “resultado”; “covid”, “pandemia” y “tratamiento”; “variable” y “muestra”; “smartphone” y “problemático”.

Si nos fijamos en el extremo opuesto del “abanico”, con valores bajos en el eje 2, vemos que los vectores están agrupados de manera prácticamente paralela a este eje en torno a las palabras “pandemia”, “covid”, “tratamiento”, “intervención” y como outliers, “niño” y “tecnología”. De nuevo hay resultados mixtos y no tenemos pruebas claras del porcentaje de “clasificación correcta” de la técnica (porque no hay un criterio previo de clasificación) pero muchos vectores de este sector pertenecen a los artículos publicados a partir y después de 2020 y referentes al impacto de la pandemia en la salud mental y el uso de redes sociales.

En el caso del análisis de componentes principales de la **matriz de variables = palabras**, los tres primeros ejes explican el 7.672% de la varianza; el eje uno el 2.748%, el eje 2 el 2.582% y el eje 3 el 2.342%. En este análisis de componentes principales, la dificultad de quedarse con un número de ejes o con elementos característicos de los ejes es aún mayor. Lo cual es absolutamente previsible teniendo en cuenta el hecho de que, con las palabras como variables,

hay mayor número de variables que de individuos (en este caso, casi dos veces superior al número de individuos porque tenemos 255 palabras).

No existe el mismo nivel de asimetría que en el biplot anterior (ya que en este caso apenas hay diferencia entre los ejes 1 y 2) y pero los vectores más largos se sitúan en el cuadrante 4 y corresponden a palabras de metodología que podemos aseverar que encontraremos en la mayoría de artículos, como “randomizar”, “efectividad”, “tratamiento”, “mejorar” o “prueba” y “semana” o “mes” que podrían corresponder al tiempo que se han tardado en aplicar esas pruebas y tratamientos. También están en este cuadrante las palabras “cbt” (terapia cognitivo conductual) y “terapia” en sí, que podrían indicarnos que se trata de una búsqueda bibliográfica en el campo de la psicología. Llama la atención el hecho de que una de las palabras clave en nuestra búsqueda “ansiedad”, si bien no presenta uno de los vectores más largos, está prácticamente alineada con el eje 1 hacia la derecha (aumento de los valores).

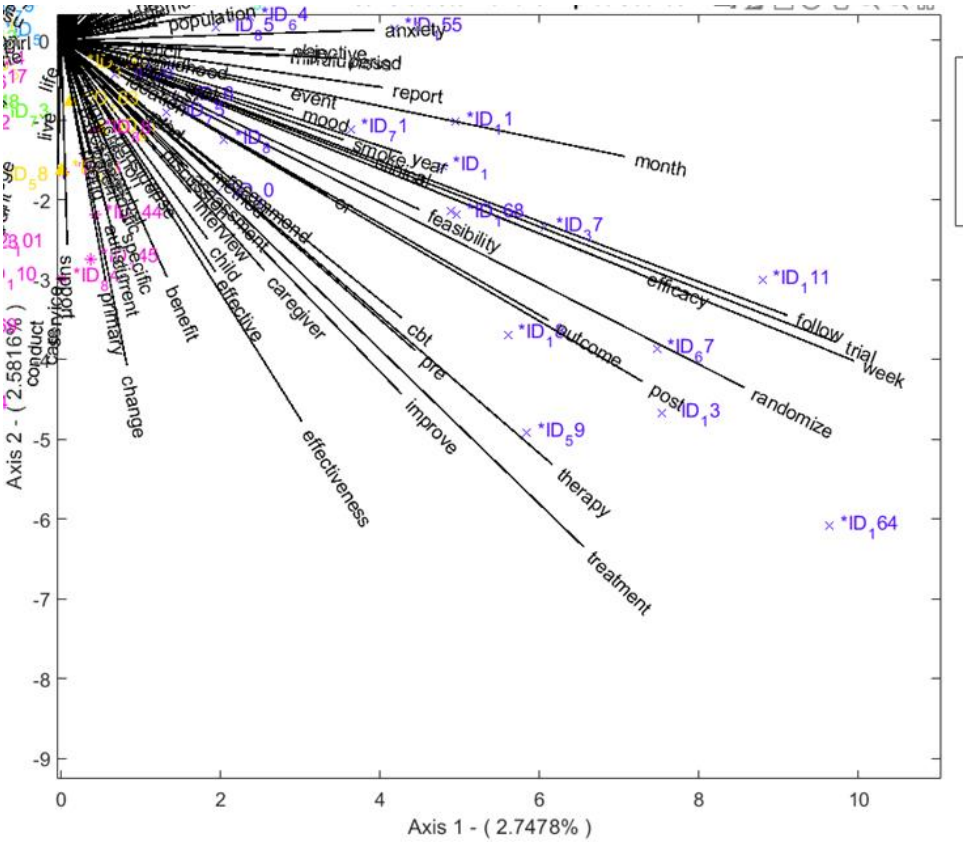

Figura 4. Cuadrante 4 del HJ biplot de las palabras como variables

Si observamos la totalidad de este análisis biplot y los otros cuadrantes, también en ellos los vectores más largos pertenecen generalmente a las palabras referentes a la metodología. En el

cuadrante dos tenemos las palabras referentes a medidas, cuestionarios y evaluaciones. Las palabras con “significado” respecto a los temas que tratan los artículos están distribuidas de manera bastante simétrica, sin muchas relaciones notables con alguno de los ejes a primera vista. Hay de nuevo algunas parejas y tríos de palabras interesantes como “smartphone”, “problemático” y “nomophobia”, “high” y “score” etc, lo cual nos aporta la información de que posiblemente aparezcan juntas con cierta frecuencia.

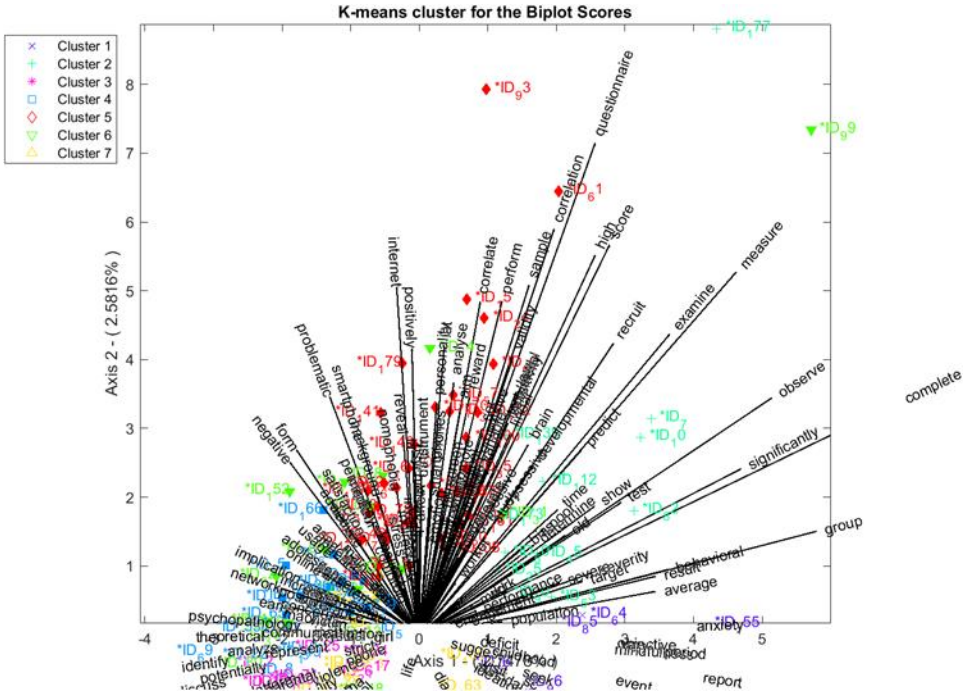

Figura 5: cuadrantes 1 y 2 con los vectores representando a las palabras

De la comparación de ambas representaciones, tanto por la comparación a simple vista como por los índices proporcionados en el análisis de componentes principales, podemos concluir que, por la distribución de la variabilidad y la manera de agrupar los vectores, el biplot que usa las palabras como individuos y los artículos como vectores es más interesante para los objetivos de este trabajo. Sin embargo, se seguirá explorando ambas opciones mediante las técnicas de clasificación.

4.1.3 Estimación de parámetros de clúster

4.1.3.1 DBSCAN

Siendo el análisis DBSCAN en base a la densidad de individuos cercanos y no requiriendo un número de clústeres previo, los parámetros que hay que establecer antes de llevarlo a cabo son un número mínimo de vecinos (número mínimo de individuos necesarios para considerar que una agrupación es un clúster) llamado k y un índice de densidad llamado épsilon, que establece a cuánta distancia se considera que dos individuos pertenecen al mismo grupo.

Para el cálculo de una medida de densidad adecuada, se ha seleccionado el número de vecinos estándar y se han generado en R gráficos de distancia Knn (k nearest neighbours) en función de k=5.

El cálculo del gráfico se hace en base a la matriz de distancias del biplot con puntuaciones normalizadas.

La regla de decisión a la hora de interpretar este tipo de gráfico exploratorio es similar a la regla del codo; se selecciona la distancia en donde hay una curva entre los valores altos y bajos.

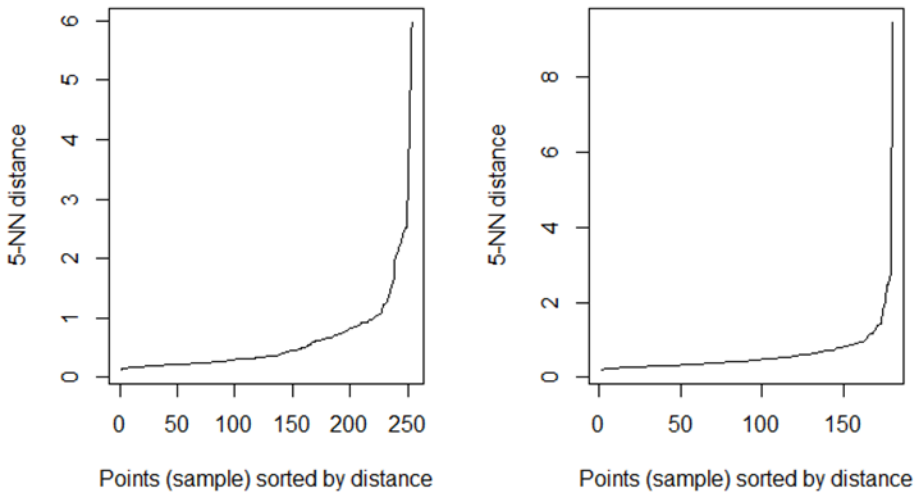

Figura 6: distancia de K vecinos próximos para la matriz de distancias A (izquierda) y matriz de distancias P (derecha)

4.1.3.2 Clúster jerárquico y K medias

En ambas técnicas tenemos que establecer el número de clústeres previo (aunque en el caso del clúster jerárquico no siempre se tiene por qué, pero debido al tipo de software usado y por cuestiones de libertad de exploración de diferentes opciones, se ha hecho este cálculo también).

Con la función de R especificada en la sección de metodología, se ha dado un máximo de clústeres de 15 y se ha seleccionado el número con más técnicas a favor de un número concreto de clústeres.

Se procede a resumir estos resultados esquemáticamente:

- Análisis 1 - Clustering Jerárquico (método de Ward) para matriz de distancias P: La mayoría de los indicadores (siete) respaldan la elección de 3 clústeres como el número más adecuado.
- Análisis 2 - Clustering Jerárquico (método de Ward) para matriz de distancias A: La mayoría de los indicadores (cinco) respaldan la elección de 3 clústeres como el número más adecuado.
- Análisis 3 - K-Means para matriz de distancias P: La mayoría de los indicadores (seis) respaldan la elección de 3 clústeres como el número más adecuado.
- Análisis 4 - K-Means para matriz de distancias A: La mayoría de los indicadores (siete) respaldan la elección de 4 clústeres como el número más adecuado.

Como se establecerá a continuación, el número de clústeres establecido con estas reglas no siempre va a aportar la riqueza necesaria para nuestro análisis.

4.1.4 HJ BIPLOT con cluster basado en densidades

Los dos motivos principales por los que se había propuesto el uso de esta técnica eran, por un lado, 1- la ausencia de necesidad de pensar en un número predeterminado de núcleos temáticos por parte de una persona que no esté familiarizada con los textos o que se esté viendo sobrepasada por su cantidad y tamaño y 2- la poca sensibilidad de este tipo de técnica de clasificación a los outliers y al ruido, que puede suponer y supone problemas a la hora de aplicar otras técnicas, como veremos en los próximos apartados.

Sin embargo, el resultado de usar esta técnica en este ejemplo y sobre los biplots presentados ha sido bastante deficiente.

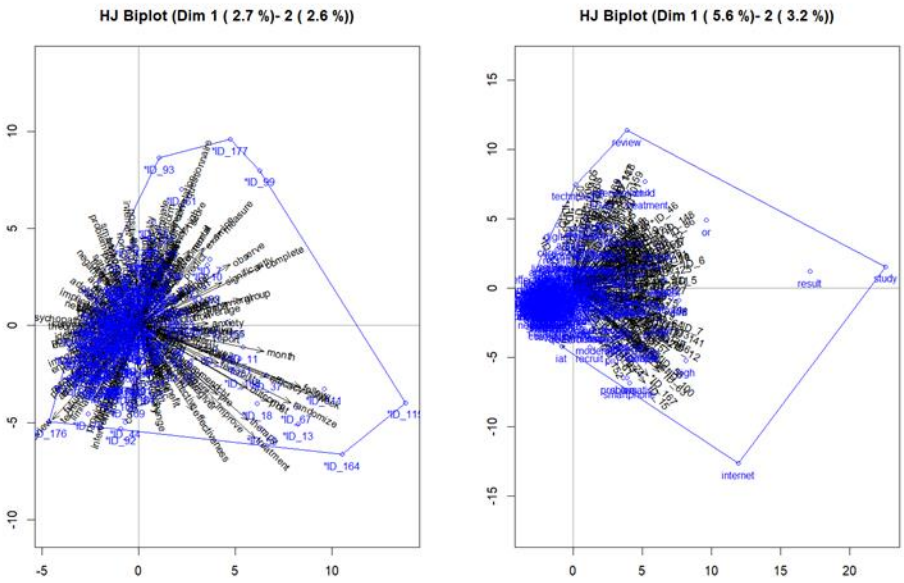

Figura 7: Representación del HJ biplot de la matriz de variables palabras (izquierdas) y variables artículos (derecha) con clúster basado en densidades.

- Matriz de Variables Palabras:

El análisis de clustering en la matriz de variables de palabras utilizando DBSCAN con parámetros eps = 160 y minPts = 5 ha dado como resultado un solo cluster con 180 objetos. No se ha identificado ningún punto de ruido. Esto significa que todos los objetos en esta matriz se consideran parte del mismo grupo.

- Matriz de Variables Artículos:

De manera similar, en la matriz de variables de artículos, el DBSCAN con parámetros eps = 240 y minPts = 5 también ha agrupado todos los objetos en un solo cluster con 254 objetos, sin puntos de ruido identificados.

Los parámetros epsilon (radio de búsqueda) y minPts (número mínimo de puntos en un vecindario para formar un clúster) son críticos en DBSCAN. Si se eligen valores inapropiados para estos parámetros, puede conducir a la formación de un solo clúster. Es posible que los valores seleccionados no fueran adecuados a pesar de la prueba previa, entonces se ha hecho un intento de rebajar el valor de épsilon hasta obtener resultados diferentes.

- Matriz de Variables Palabras:

Tras ajustar los parámetros de DBSCAN en la matriz de variables de palabras, se han obtenido resultados más significativos en términos de clustering. Utilizando un valor de eps = 0.3 y minPts = 5, DBSCAN ha identificado la presencia de 4 clusters diferentes junto con 102 puntos de ruido.

- Cluster 0: 102 individuos
- Cluster 1: 62 individuos
- Cluster 2: 5 individuos
- Cluster 3: 6 individuos
- Cluster 4: 5 individuos

- Matriz de Variables Artículos:

De manera similar, en la matriz de variables de artículos, la modificación de los parámetros de DBSCAN (eps = 0.5, minPts = 5) ha conducido a resultados más informativos. Se han identificado 2 clusters principales junto con 3 puntos de ruido.

- Cluster 0: 75 individuos
- Cluster 1: 176 individuos
- Cluster 2: 3 individuos

Como se puede ver, ahora hay un exceso enorme de ruido, los individuos en cada clúster son muy pocos y la información sigue siendo inutilizable, particularmente en la primera matriz, al margen de que la épsilon ha sido bajada varios cientos de veces respecto a lo que se había establecido según las pruebas.

Con valores de épsilon por encima de los decimales (mayor o igual a uno) sigue encontrándose un único clúster.

Por tanto, queda concluir que esta técnica no es adecuada para esete tipo de datos a pesar de las esperanzas puestas en ella. Es posible que en el caso de la matriz de variables artículos esto se deba al menos en parte al gran número de individuos agrupados como palabras de baja frecuencia en los cuadrantes uno y tres.

4.1.5 HJ BIPLOT con clúster jerárquico

4.1.5.1 Matriz de variables palabras

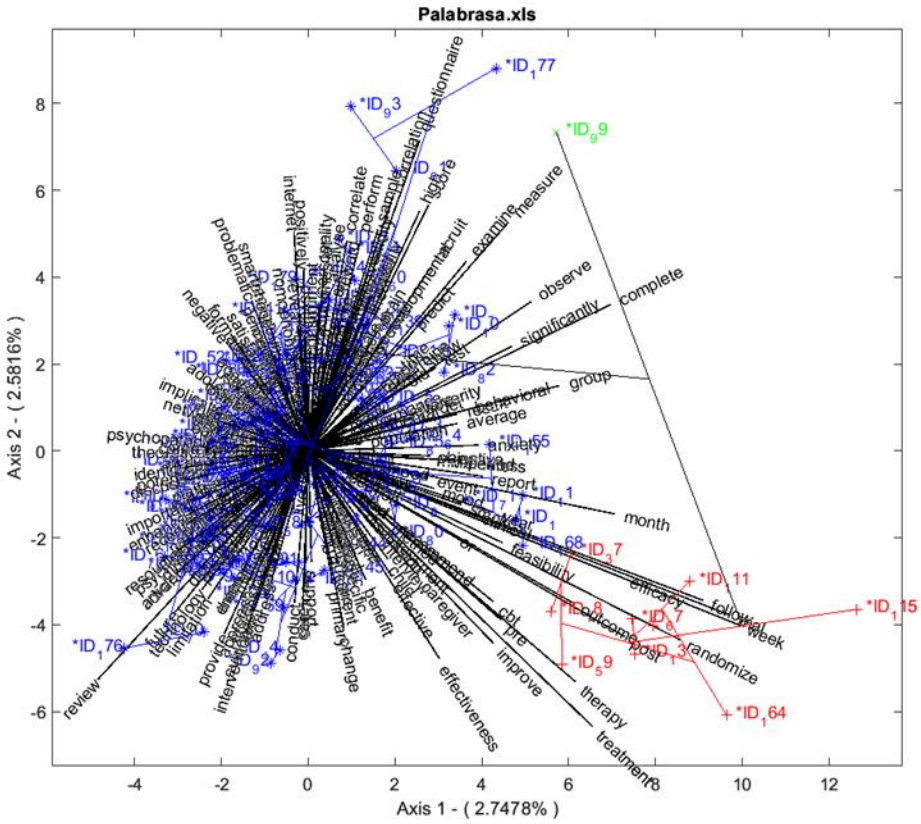

Figura 8: Biplot de la matriz P con clúster jerárquico con dendrograma

Lo primero que podemos notar examinando este clúster a simple vista es que, aplicando el número de clústeres que habíamos establecido en la exploración previa, los outliers forman dos de tres clústeres absorbiendo prácticamente toda la riqueza informativa que nos podrían proporcionar los datos. Esta técnica es excesivamente sensible a outliers y los resultados son pobres.

Dado este caso y como el preprocesamiento de los datos es demasiado costoso a estas alturas del proceso (aunque podría proponerse para futuras aplicaciones), se ha hecho la prueba de pedir un número mayor de clúster al inicio del algoritmo, controlando que hay por lo menos dos grupos de outliers y que nos gustaría identificar al menos cuatro o cinco categorías o núcleos temáticos dentro de los datos. Con estas consideraciones, se ha generado una representación del biplot para 7 clústers.

En la sección de Materiales y métodos comentábamos el motivo por el cual se escogió la unión de Ward comparado con otros métodos para crear los enlaces de agrupamiento. Aún con los resultados poco interesantes que estamos obteniendo hasta ahora, cabe ilustrar qué ocurre cuando usamos otros métodos de enlace **(Figura 9)**. Los clústeres en base a medias y medias ponderadas son aún más sensibles a outliers y agrupan las palabras por frecuencias, aún estableciendo un mayor número de clústeres antes de la iteración del algoritmo.

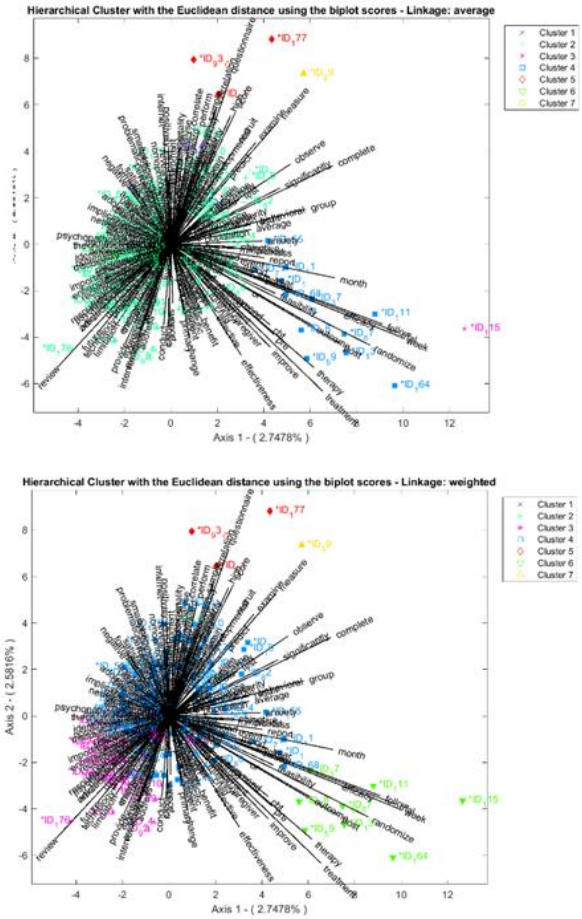

Figura 9: Clúster en base a medias (arriba) y medias ponderadas (abajo).

Se prosigue con el clúster jerárquico de 7 grupos con unión de Ward **(Figura)** Ante la inspección visual, los clústeres 1 y 6 y probablemente el 7 están conformados por outliers asociados a las palabras (principalmente de metodología) de uso más frecuente.

Vamos a explorar más detenidamente los clústeres 2, 3, 4, 5 y 7 para determinar si hay algún tipo de agrupación temática de los artículos que sea interesante.

- Cluster 2 El segundo clúster incluye bastantes artículos que contienen la palara “autismo” y en general, referentes a niños con trastornos (déficit de atención, depresión).
- Cluster 3: En el tercero encontramos agrupados algunos artículos que contienen la palara “social” y por tanto bastantes de los artículos referentes a las redes sociales. Los artículos más relacionados con el uso nocivo de internet se encuentran principalmente entre este clúster y el clúster 5.
- Cluster 4: El cuarto clúster contiene bastantes artículos que contienen la palara “covid”, si bien es de contenido misceláneo
- Cluster 5: En el quinto clúster, encontramos agrupados basteantes de los artículos que contienen la palara “internet”
- Clúster 7: El clúster 7 contiene muy pocos individuos, pero examinando los artículos detenidamente, se puede comentar un fenómeno curioso; se trata de artículos que se desvían ligeramente del tema del resto. Mencionan trastornos poco comunes, aspectos biológicos no muy relacionados o tienen los términos de la búsqueda bibliográfica como algo secundario. Por ejemplo, un artículo sobre los efectos de la nicotina en la atención o uno sobre el síndrome premenstrual.

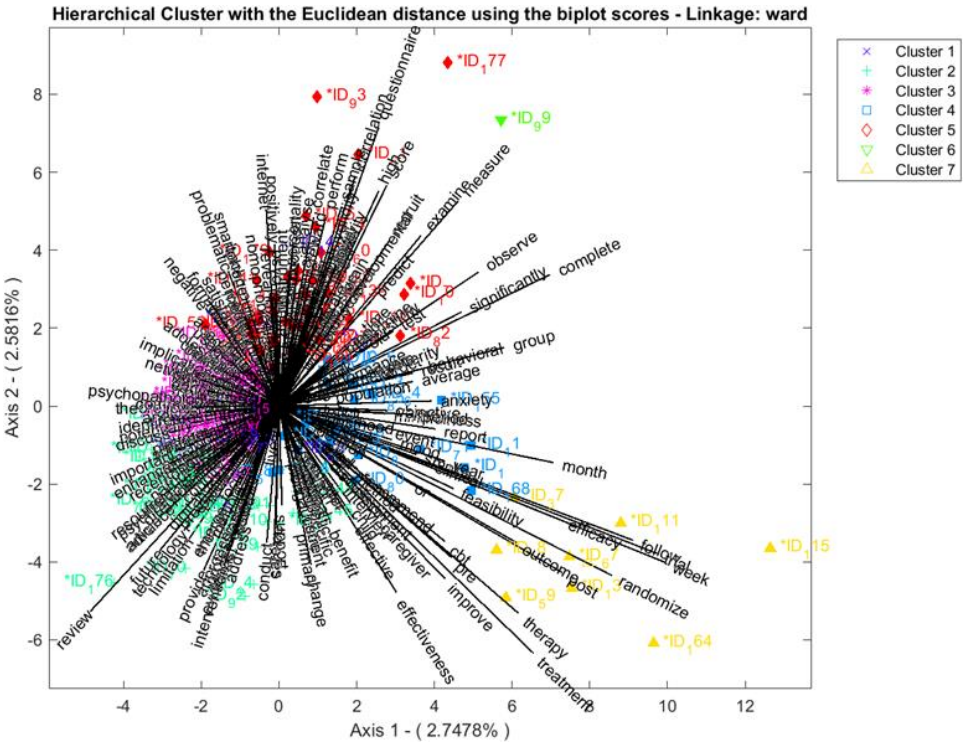

Figura 10: Biplot de la matriz P con clúster jerárquico para 7 grupos

Las tendencias encontradas en estos clústeres son una interpretación y una impresión humana. Requieren de más pruebas mediante una clasificación detallada previa, como asignar un código de significado a cada artículo para poder hacer un examen de los porcentajes de acierto, o una revisión del número total de clúster que contienen cada palabra de las que se ha detectado aquí como moda para poder decir que aportan un significado sólido mediante esta clasificación.

Una gran desventaja de esta visualización es que, con la cantidad de ruido del gráfico inicial es muy difícil establecer que los clústeres se agrupen realmente en torno a los significados que se han extraído de la lectura, uno por uno, de los abstracts que conforman cada clúster. Esto podría solucionarse por el usuario, una vez se ha realizado una primera exploración, eliminando manualmente las palabras más genéricas y dejando aquellas que se perciben como posibles características de cada clúster.

Aparecen menos separados y peor clasificados los temas que transcienden en toda la búsqueda; por ejemplo, los artículos referentes al déficit de atención y al trastorno de ansiedad

no estaban caracterizados concretamente en ninguno de los clústers. Los clusteres de outliers parecen correlacionarse con outliers temáticos (artículos con temas particulares).

4.1.5.2 Matriz de variables artículos

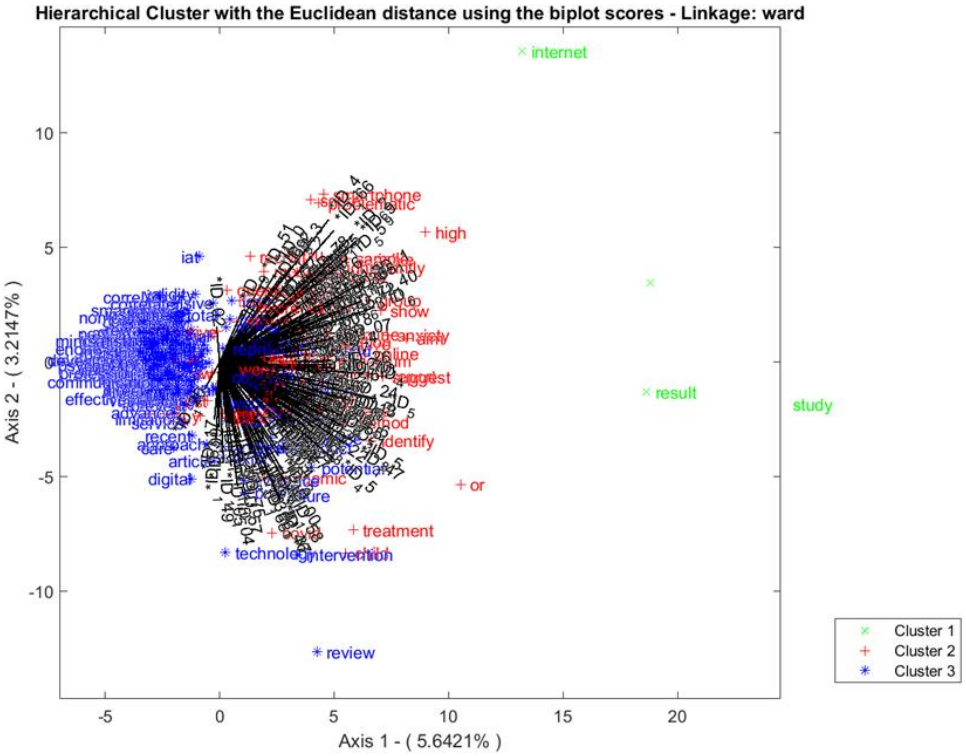

Figura 11: Biplot con clúster jerárquico para tres grupos

Volvemos a sufrir las limitaciones de esta técnica en cuanto a la sensibilidad a outliers; el clúster 1 está conformado exclusivamente por las palabras “internet”, “resultado” y “estudio” en su calidad de outliers. Los clústeres 2 y 3 reflejan la limitación del trabajo con frecuencias; el clúster 2 reúne las palabras más frecuentemente empleadas, en su mayoría de metodología y algunas acerca de los temas generales que componen el marco común de la búsqueda, como “ansiedad” o “smartphone”. El clúster 3 contiene la nube de palabras fuera del “abanico” de variables, limitándose así al a las palabras con poca frecuencia total específicas de ciertos artículos concretos.

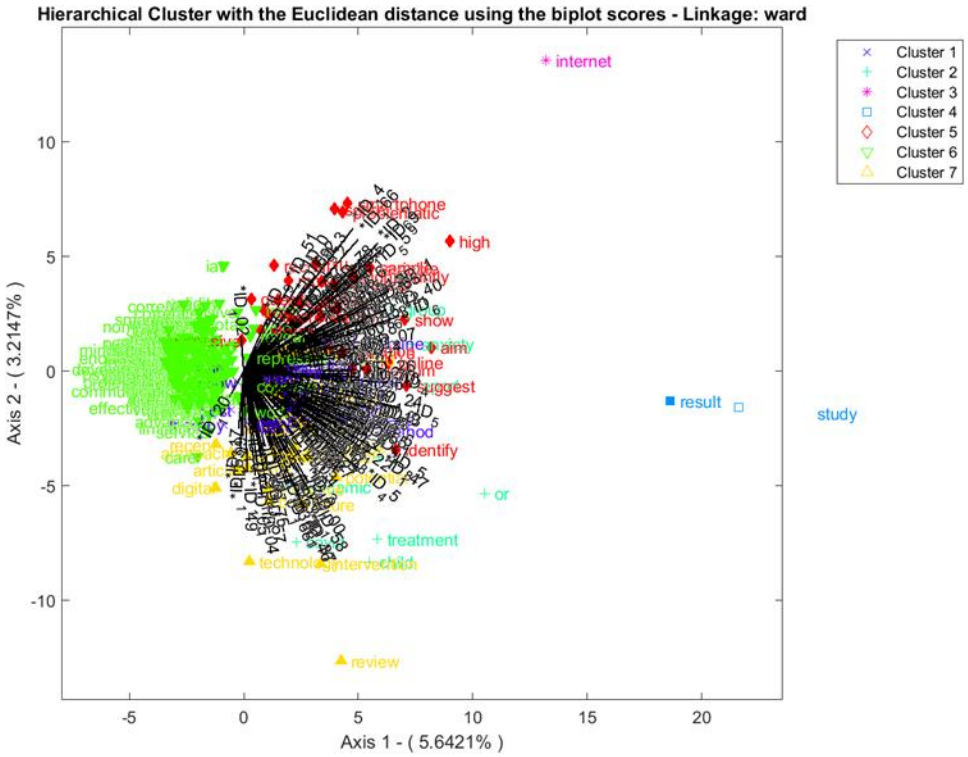

Figura 12: Biplot de la matriz de artículos como variables con 7 clústers jerárquicos con el método de Ward

Se ha intentado aplicar el procedimiento que ha funcionado, hasta cierto punto, para tener agrupaciones más ilustrativas, en la matriz de palabras como variables. Sin embargo, el éxito en cuanto a agrupar y distinguir las palabras que no se proyectan sobre el abanico de variables sigue siendo muy limitado. No nos detendremos a comentar en detalle esta última clasificación salvo para decir que el clúster k-medias produce en esta situación un resultado más rico e interesante; será bajo ese apartado en el que se caracterizarán los clústeres por temáticas.

4.1.6 HJ BIPLOT con cluster K medias

4.1.6.1 Matriz de variables palabras

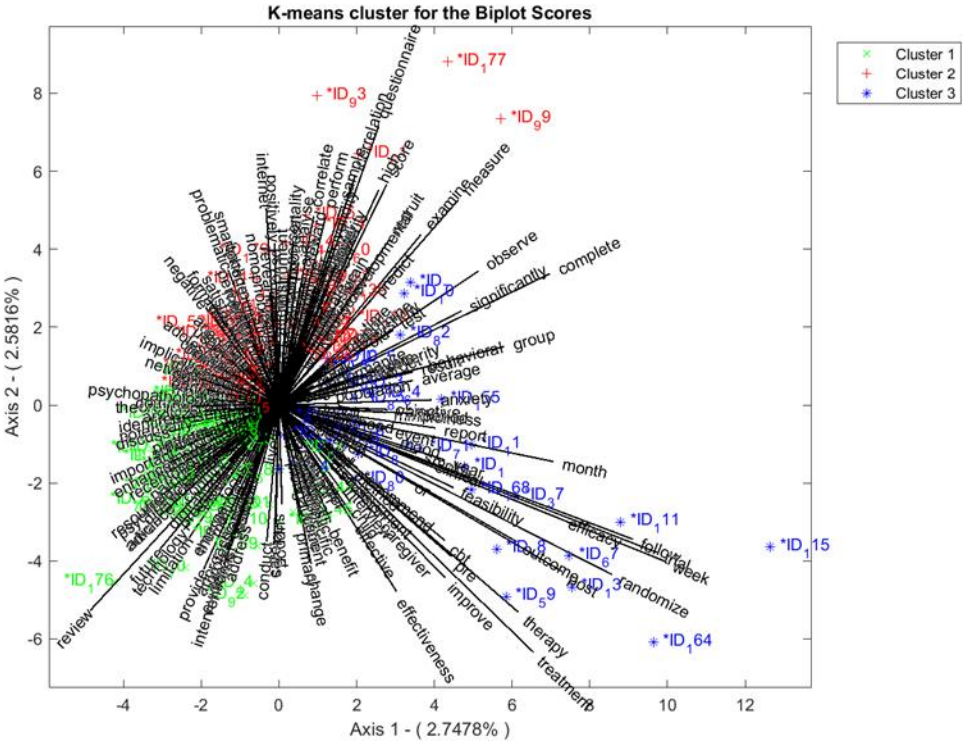

Figura 13: Clúster K medias con 3 grupos sobre el biplot de palabras como variables

Con el clúster de k medias repetimos el criterio aplicado anteriormente; primero probamos con el número de clústers establecido por el análisis previo, es decir, tres.

Hay una sensibilidad a los outliers considerablemente menor que en la anterior técnica. El clúster dos se sitúa en torno a los valores positvos del eje uno y de forma más o menos simétrica en cuanto al eje dos (cuadrantes uno y dos). El clúster uno se sitúa en los valores negativos de ambos ejes (cuadrante 3) mientras que el clúster 3 presenta valores altamente positivos para el eje 1 y mayormente negativos para el eje 2 (cuadrante 4).

En circunstancias normales trataríamos de establecer alguna relación entre estos grupos y sus características de significado mediante sus proyecciones sobre los ejes, pero como establecimos, particularmente en este biplot es casi imposible correlacionar los ejes con alguna temática por la distribución demasiado dispersa de la variabilidad y el exceso de variables.

En cuanto a la clasificación de artículos, es muy difícil explorar esa cantidad de datos dividiendo en solamente tres grupos, teniendo en cuenta además que las variables sobre las que los individuos están proyectados con vectores más largos son palabras generales y de

metodología. Por tanto, como en el caso anterior, vamos a proceder a indicar manualmente en el algoritmo un número mayor de clústers. Cabe comentar que en este caso, al ser una cuestión de visibilidad y exploración, no se tiene la justificación previa del número de outliers en la decisió de cuántos clúster crear. Por tanto, el quedarse con 7 clústeres se está haciendo puramente por comparar los resultados con los de la aplicación del clúster jerárquico.

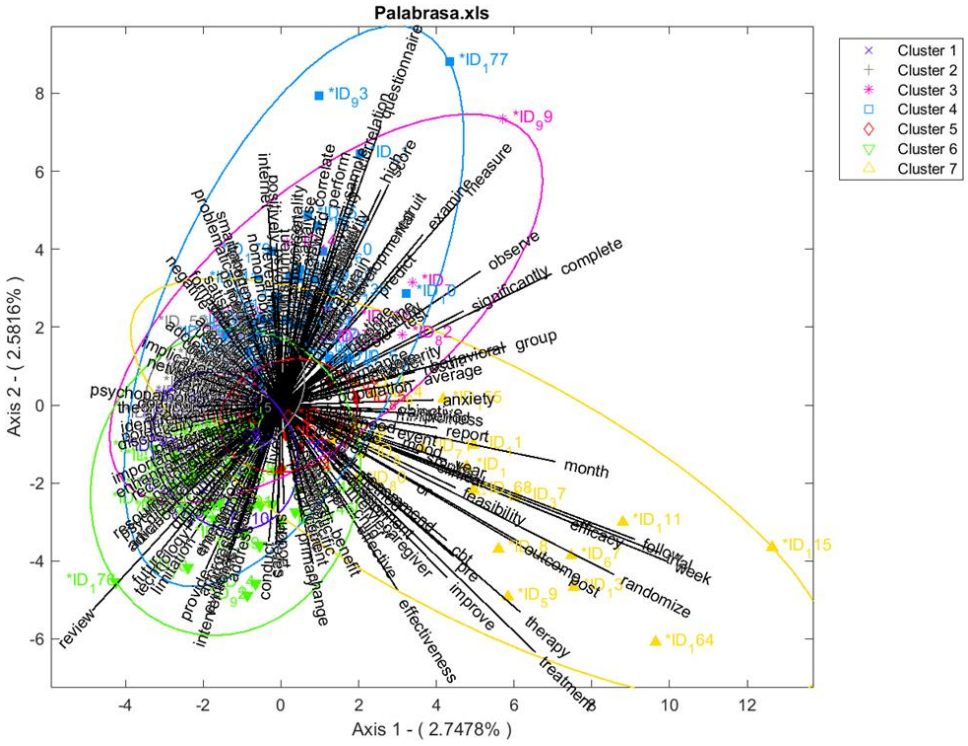

Figura 14: Clúster K-medias con 7 grupos sobre el HJ biplot de palabras como variables

Dibujando las elipsis alrededor de los clústeres, una de las primeras cosas que podemos notar es que los individuos se agrupan de una manera más dispersa y menos claramente proyectada en cierto cuadrante o respecto a cierto eje respecto tanto a la aplicación con el mismo número de grupos, siete, mediante clúster jerárquico como respecto a la aplicación de k-medias con menos grupos. Por tanto, resulta aún más difícil relacionar los grupos vectores concretos.

- Clúster 1

  Se ha identificado que la mayoría de los estudios en este clúster tienen que ver con sujetos jóvenes y su uso de la tecnología, pero también del impacto de esta en la productividad de trabajadores mayores de edad.
- Clúster 2

Este cluster presenta prácticamente la misma temática que el anterior; el impacto de la tecnología en la gente joven. Tiene aún menos excepciones en cuanto a este tema.

Ambos clústeres están focalizados cerca de las coordenadas 0 de la matriz.

- Clúster 3

Es un clúster pequeño en cuanto a número de individuos y amplio en diámetro, por tanto los individuos están relativamente dispersos. Presenta valores relativamente altos del eje 2.

Los temas de los estudios son muy misceláneos sin relación entre sí; hay algunos estudios acerca del uso de la tecnología o la atención, pero todos tienen enfoques inusuales o característicos, por ejemplo, un estudio en el efecto de las drogas que aumentan la concentración entre la gente que practica e-sports.

- Clúster 4

De nuevo, un clúster con un diámetro amplio. No hay ningún patrón de información muy claro, salvo que la mayoría de los individuos tienen como objeto de estudio algún trastorno psicológico concreto.

- Clúster 5

Uno de los clusteres más pequeños en cuanto a individuos y centrado en las coordenadas de origen de la matriz.

Debido al número pequeño de estudios que lo componen, podemos identificar como temas el TDAH (trastorno por déficit de atención e hiperactividad) y la adicción (muchas veces en referencia a la tecnología, como es de esperar con nuestros términos de búsqueda.

- Clúster 6

Este clúster es relativamente voluminoso en número de individuos y se centra en el cuadrante 3 (bajos valores en ambos ejes).

No hay temas destacables. Reúne algunos de los estudios acerca del impacto del COVID, pero éstos también se reparten entre los clústeres 2, 4 y 7.

- Clúster 7

Aparte de artículos de contenido “habitual” (TDAH, Ansiedad, uso problemáticos de redes sociales) contiene algunos de los outliers interesantes del cuadrante 4 (estudios sobre tabaco, menstruación).

Se concluye que los clústeres 1 y 2, más cercanos al origen de los vectores, contienen los artículos más cercanos al tema de investigación mayoritario que fue objetivo de búsqueda. Los clústeres más alejados y pequeños contienen outliers de estudios con un enfoque particular o que se salen del tema. No hay una particular ventaja en esta matriz en aplicar el clúster de k medias en vez del clúster jerárquico. Incluso, este último parece detectar mejor artículos que “se salen” del tema de investigación mediante clústers pequeños de outliers.

4.1.6.2 Matriz de variables artículos

Este es el análisis más rico e interesante en cuanto al objetivo inicial del estudio de extraer núcleos temáticos, por tanto, nos vamos a parar un poco más detalladamente en analizar cada una de las partes. Se van a presentar los grupos de palabras que corresponden a cada clúster porque al tratarse de palabras sueltas es relativamente fácil para el lector sacar sus propias conclusiones.

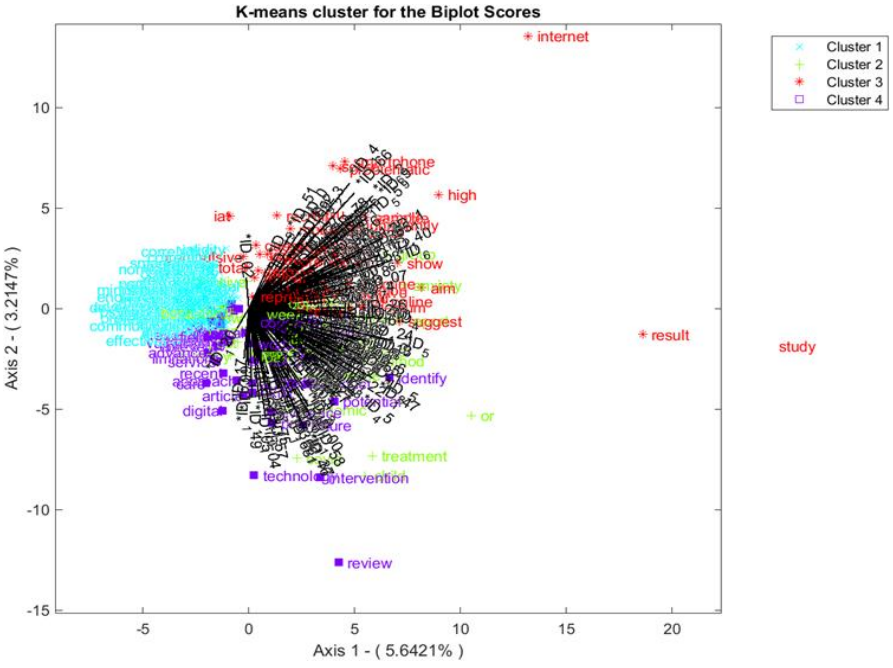

Figura 15: Clúster K medias con 4 grupos sobre el biplot de artículos como variables

Como en el caso de las aplicaciones anteriores, primero vamos a aplicar el número de clústeres reducido que nos han indicado las técnicas de decisión estadísticas.

- Clúster 1:

Junto con el 4, contiene una cantidad de términos mucho mayor que los clústeres 2 y 3. A él pertenecen la mayoría de las palabras que no se asocian directamente a ninguna de las variables/ artículos por baja frecuencia en los valores bajos del eje 1.

Si intentamos hacer un análisis de significados, este clúster incluye palabras relacionadas con la salud mental y el bienestar, como "suicida", "sueño", "afecto", "soledad", " comorbilidad", "emoción", "psicométrico" y "psicopatología". También contiene palabras que sugieren evaluación y diagnóstico, como "percepción" y "evaluación", así como términos relacionados con la atención médica, como "médico" y "cuidador".

- Clúster 2:

Este es el clúster tiene relativamente pocos individuos y es el que más dispersión entre los datos presenta, sin una clara relación con ninguno de los ejes, aunque tiene una cierta extensión hacia el cuadrante 4 (valores bajos de ambos ejes).

En cuanto a significados léxicos, este clúster parece estar relacionado con la investigación y el análisis de datos, con palabras como "estudio", "total", "negativo", "puntuación", "variable" y "prevalencia". También incluye términos relacionados con la salud mental y el marco general de la búsqueda que hemos realizado como "TDAH" (trastorno por déficit de atención con hiperactividad) y "antecedentes".

- Clúster 3:

Es el clúster con menor número de individuos y aúna los términos con valores más altos en ambos ejes, incluyendo los muy visibles outliers “internet”, “estudio” y “resultado”.

Encontramos palabras relacionadas con la terapia y la atención psicológica, como "terapia", "ansiedad", "psiquiátrico" y "tratamiento". También incluye términos relacionados con la comunicación, la tecnología y el comportamiento, como “internet”, "comunicación" y "conductual".

- Clúster 4:

Presenta valores bajos en el eje 2 y simétricos en el eje 1.

Siendo el segundo más grande después del 1, es muy difícil sacar una temática concreta. Podríamos aseverar que está relacionado con la educación y la evaluación, con palabras como "escuela", "mejorar", "conducta", "vida" y "educación". También contiene términos relacionados con la efectividad y el rendimiento, como "eficaz", "eficacia" y "rendimiento". Es el clúster al que pertenecen las palabras “COVID” y “pandemia” y además unas cuantas palabras relacionadas con metodología.

En resumen, conservar 4 clústeres permite solamente entender que se trata 1.una revisión de textos científicos 2. En el campo de la salud mental y la psicología.

Por eso pasamos a realizar el análisis de 7 clústeres. En este caso vamos a incluir todas las palabras que pertenecen a cada uno de los clústeres para que el lector pueda decidir si coincide con la interpretación y porque esta es la técnica más rica e interesante que se ha aplicado.

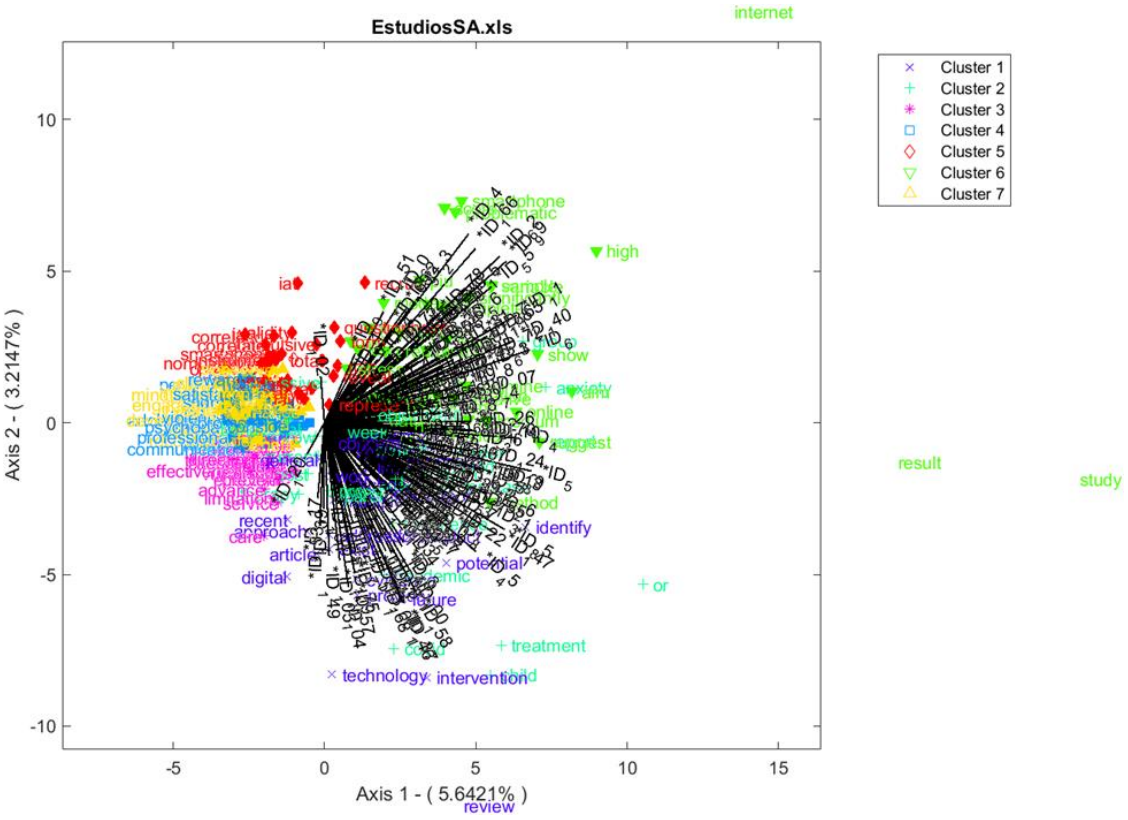

Figura 16: Clúster K medias con 7 grupos sobre el biplot de artículos como variables

- Clúster 1:

El clúster 1 contiene: *escuela, conducta, trabajo, proveer, implicación, enfoque, vida, identificar, específico, apoyo, evaluación, existir, digital, enfoque, dirección, criterio, futuro, potencial, múltiple, preocupación, revisión, artículo, evidencia, tecnología, desarrollar, intervención, general, importante, reciente*

Este clúster parece estar relacionado con la educación y el desarrollo, incluyendo temas como la escuela, la conducta, el trabajo, la evaluación, la tecnología y la intervención. Podríamos decir que aporta la idea de que hay una orientación a la psicología educativa en la muestra. También contiene bastantes palabras genéricas de metodología.

- Clúster 2:

Contiene: *seguir, terapia, ansiedad, TDAH (Trastorno por Déficit de Atención e Hiperactividad), psiquiátrico, informe, depresivo, mejorar, eficacia, niño, experiencia, publicación, clínico, tiempo, actual, ensayo, grupo, año, aumento, pandemia, COVID, resultado, pre, objetivo, padre, semana, tratamiento, o, conductual.*

En este clúster se encuentran palabras relacionadas con la salud mental y el tratamiento, como terapia, ansiedad, diagnóstico y tratamiento médico. También contiene las palabras referentes a la pandemia de 2020. Podemos extraer que nuestra muestra está relacionada con el área de la salud y que hay muchos artículos en torno al impacto de la pandemia en la salud mental.

- Clúster 3

Contiene: *vulnerabilidad, involucrar, efectivo, contenido, servicio, objetivo, socio, efectividad, tema, naturaleza, limitación, prevenir, cambiar, emerger, discusión, política, discutir, diagnóstico, cuidado, avanzar, recurso, parental, dirección, temprano, campo*.

En este clúster, se abordan temas relacionados con la vulnerabilidad, la efectividad de los servicios, la política, el diagnóstico, el cuidado y la intervención temprana. Podríamos decir que en nuestra muestra existe una preocupación por mejorar servicios, intervenciones y condiciones de ciertas poblaciones o afectados por ciertas dolencias. También tenemos palabras que evalúan las limitaciones de los propios estudios.

- Clúster 4

Contiene: *afectar, mejorar, viabilidad, emplear, sustancia, tipo, teléfono, adictivo, máquina, DSM (Manual Diagnóstico y Estadístico de los Trastornos Mentales), percepción, herramienta,*

PSMU (Uso Problemático de las Redes Sociales), autista, analizar, país, trabajador, potencialmente, humano, derecho, físico, psicosocial, saludable, positivo, teórico, SNS (Redes Sociales), SC (Criterio de Sistemas), plataforma, principal, niña, requerir, profesional, estigma, adolescencia, recompensa, satisfacción, uso, violencia, comunicación, psicopatología, promover, considerar.

Este clúster está relacionado con la tecnología y la psicología, incluyendo palabras como adicción a las redes sociales, inteligencia artificial, percepción, redes sociales, plataforma, estigma y comunicación. Podríamos estimar que recoge la parte de la búsqueda referente al impacto de la tecnología en la salud mental y hay un énfasis en la juventud o los niños.

- Clúster 5

Contiene: *obsesivo, total, compulsivo, correlación, red, lineal, positivamente, personalidad, reclutar, IA (Inteligencia Artificial), realizar, nomofobia, forma, emocional, rol, diario, ítem, aparecer, chico, validez, correlacionar, jugadores, IAT (Test de Adicción a Internet), variar, revelar, teléfonos inteligentes, grave, instrumento, cuestionario, representar.*

Aquí, se discuten temas relacionados con la compulsión, la adicción a Internet, la evaluación psicológica, la personalidad, la correlación y el uso problemático de dispositivos móviles. Podría relacionarse también con el estudio de las respuestas emocionales y los patrones de comportamiento. Tanto en este como en el anterior se implica o menciona la adicción, que en contexto sería a la tecnología.

NOTA: Los clústeres 4 y 5 no están claramente diferenciados y aportan información parecida.

Clúster 6 contiene: *riesgo, estudio, resultado, total, negativo, estudio, mediar, vulnerabilidad, informar, naturaleza, muestra, correlación, idea, estrés, grupo, internet, diagnóstico, bajo, completo, examinar, sugerir.*

- Clúster 6:

Contiene: *riesgo, estudio, psicológico, negativo, adulto, PIU (Uso Problemático de Internet), método, resultado, puntuación, objetivo, sugerir, antecedentes, antiguo, medio, analizar, prueba, variable, dato, teléfono inteligente, nivel, problemático, predecir, muestra, estrés, significativamente, internet, población, medida, en línea, mostrar, prevalencia, modelo, déficit, moderado, determinar, bajo, completo, examinar, alto.*

En este clúster, las palabras sugieren un enfoque en análisis estadísticos y metodológicos. Contiene claramente las palabras de metodología acerca de investigaciones psicológicas y estudios, incluyendo el estrés, el riesgo, el diagnóstico psicológico, términos como “resultado”, “población” y “prevalencia”. Por desgracia, mezcla los términos metodológicos con un par de términos muy clave del objeto de búsqueda, como tecnología, internet y PIU.

- Clúster 7

Contiene: *suicida, sueño, soledad, fumar, mediar, recomendar, comórbido, característica, entrevista, emoción, MDD (Trastorno Depresivo Mayor), primario, compromiso, buscar, médico, aleatorizar, percibir, víctima, sensibilidad, inatención, caracterizar, gravedad, inventario, académico, cuidador, PSNSU (Uso Problemático de las Redes Sociales), indirecto, confinamiento, par, PTSD (Trastorno de Estrés Postraumático), diagnosticar, global, suicidio, visita, vivir, cerebro, ideación, infancia, largo, enfermedad, atención plena, rendimiento, promedio, universidad, pantalla, período, observar, caso, grupo, sesgo, preocupación, beneficio, desarrollo, estado de ánimo, estricto, TCC (Terapia Cognitivo-Conductual), evento, pobre, mes.*

Este clúster aborda la salud y el bienestar, incluyendo palabras como ejercicio, nutrición y actividad física pero más concretamente de salud mental, incluyendo: (suicidio, trastorno depresivo mayor, estrés postraumático, terapia cognitivo-conductual, la atención plena y salud emocional). Destacan, por desgracia, “pantalla” y “PSNSU”.

La conclusión que se presenta acerca de este último análisis es que podemos extraer bastante información acerca de la muestra y algunos núcleos temáticos interesantes, pero los núcleos temáticos más frecuentes que caracterizan el total de la muestra aparecen peor separados. Concretamente, si tuviéramos que llamarlos de alguna forma, “Psicología y salud mental” (el área de la búsqueda) y “uso problemático de la tecnología” aparecen a través de varias agrupaciones, aunque en cada una adquieran matices.

4.2 Resultado del análisis temático de la muestra

La muestra de esta búsqueda bibliográfica refleja diversidad de temas pero la relación entre la tecnología y la salud mental es el hilo conductor (como cabe esperar por los términos de búsqueda).

Persiste una preocupación por el uso problemático de la tecnología, las redes sociales y los dispositivos móviles en personas de diferentes características.

Algo que sabemos por la lectura detenida de los abstracts pero que este análisis NO ha podido identificar es que un número limitado de estudios también se centran en la tecnología como plataforma y oportunidad de hacer la salud mental más disponible.

Destacan problemas y trastornos como la ansiedad, el trastorno por déficit de atención e hiperactividad (TDAH), el autismo, la depresión y el estrés postraumático, lo que sugiere una predominancia del enfoque en el ámbito de la psicología clínica. También tiene influencia el enfoque educativo y del neurodesarrollo.

Además, la influencia de la pandemia de COVID-19 en la salud mental se encuentra entre los temas más abordados.

La muestra también se dirige a poblaciones específicas, como niños, adolescentes, adultos y trabajadores, lo que indica un interés en comprender y abordar las cuestiones de salud mental en diferentes grupos demográficos, pero una gran mayoría se centra en niños y jóvenes así que se puede asumir que hay una gran preocupación por esta población.

Se observa la presencia muy frecuente de términos relacionados con la metodología de investigación, lo cual en parte es debido al tipo de textos pero también podemos interpretar como un interés en la evaluación sistemática y celosa de las técnicas usadas y preocupación por la calidad de los servicios, sistemas y herramientas diagnósticas que hay.

Se nos han colado algunos estudios poco relacionados con el tema o con enfoques únicos, que se reflejan en las representaciones como outliers.

4.3 Discusión

4.3.1 ¿Columnas Artículo o columnas Palabra?

Como se ha estado comentando a lo largo de esta sección, a pesar de que, por definición, el HJ biplot optimiza la calidad de representación tanto de las filas como de las columnas simultáneamente, hay una gran diferencia entre los resultados obtenidos usando las palabras como variables/vectores en la representación respecto a usar los artículos.

En el planteamiento inicial del problema, se pensó que las palabras eran, técnicamente, lo que determina las características y el significado de los textos, por eso se esperaba, tal vez intuitivamente, que la matriz que tiene de variables las palabras fuera más interesante metodológicamente que su contraparte traspuesta.

Sin embargo, lo normal en este tipo de datos esencontrar muchas más palabras que artículos y se da el problema anteriormente comentado de que los primeros ejes-dimensiones no absorben suficiente variabilidad, lo cual para una técnica basada en las componentes principales, como es el caso del biplot, quita mucha riqueza.

Otra consideración es que las técnicas de clúster se aplican a los individuos, no a las variables. Por tanto, teniendo en cuenta que nuestro objetivo principal era localizar núcleos temáticos mediante técnicas de clúster, es lógico emplear las palabras como individuos.

BIPLOT PARA ARTÍCULOS COMO VARIABLES	BIPLOT PARA PALABRAS COMO VARIABLES
Casi siempre va a presentar mayor número de individuos que de variables, permitiendo que los primeros ejes absorban más variabilidad. Al crear clústeres, se puede obtener a simple vista una idea de los temas mediante las agrupaciones de palabras.	Al haber demasiadas variables, el scree plot es demasiado plano y los primeros ejes no absorben una variabilidad considerable. Capacidad para agrupar artículos por temas mediante el clúster en torno a ciertas palabras, pero hay que comprobar manualmente o crear códigos previos.

Tabla 1: Matrices de palabras como variables vs matrices de artículos como variables

Queda por tanto recomendar, particularmente para este objetivo, el uso del HJ biplot con las palabras como individuos y los artículos como variables.

Sin embargo, la agrupación de los artículos en clúster no se puede descartar del todo y podría resultar interesante cuando 1. se quieren detectar artículos que se salen del tema de investigación 2. cuando hay menos palabras, o una mayor visibilidad de ciertas palabras, para proyectar los artículos sobre sus vectores.

4.3.2 Selección de técnicas de clúster

A pesar de las esperanzas puestas en este método, el clúster DBSCAN queda claramente descartado para el tipo de datos que presenta esta muestra. Analizando el motivo, este tipo de clúster, a diferencia de las otras dos técnicas, no tiene por qué "converger" en el sentido tradicional. Lo que parecía su mayor fortaleza para unos datos con ruido (no necesitar asignar todos los individuos a un cluster y excluir los outliers como ruido) es precisamente el motivo por el que no se adapta muy bien a nuestra muestra.

En cuanto a la comparación entre k-means y clúster jerárquico con método de ward, los resultados han sido un tanto inesperados también. En teoría el k-means es más sensible al ruido que el clúster jerárquico, porque la presencia de outliers afecta la posición del centroide, y por tanto, de todo el clúster. Sin embargo, k- means crea clústeres más uniformes en cuanto al número de individuos. Al basarse en la distancia entre los vecinos, el método jerárquico puede crear clústeres alejados con muy pocos individuos. Dada la naturaleza de nuestro objetivo (asignar una gran cantidad de palabras a unos pocos núcleos temáticos) tener clúster pequeños en cuanto a número de individuos resta riqueza a la técnica.

DBSCAN	K-MEANS	JERÁRQUICO con método de Ward
No siempre converge: no asigna a todos los individuos a un grupo.	Siempre converge	Siempre converge
No requiere un número previo de clústeres por parte del usuario	Requiere número de clústeres.	A menudo requiere número de clúster para establecer el punto de corte.
No es sensible al ruido	Muy sensible al ruido, genera clústeres uniformes en tamaño.	Sensible al ruido, genera clústeres dispares en tamaño.

Tabla 2. Comparación de técnicas de clasificación

Por tanto, se concluye esta comparación estableciendo que la técnica de clasificación más interesante en cuanto a este objetivo y tipo de datos es el clúster k-medias aplicado al biplot de palabras como individuos. Sin embargo, no hay que descartar que el clúster jerárquico podría ser una opción válida si se quieren detectar artículos, en la matriz traspuesta, que se desvíen del tema o si se sospecha en general que hay outliers que no encajan con el resto de datos por significado y no por frecuencia (una palabra muy frecuente pero sólo en unos pocos artículos).

4.3.3 El método Reinert o HJ biplot; usabilidad vs riqueza

ALCESTE	HJ BIPLOT + TÉCNICA DE CLASIFICACIÓN
No requiere de conocimientos previos de los datos para establecer un número de variables Existe software manual para personas no familiarizadas con estadística o programación	Puede requerir que el usuario decida un número de clústeres, particularmente en el caso de K-means. El software utilizado puede ser intimidante (R) pero se podría optimizar una interfaz en el futuro.
Sólo permite hacer agrupaciones de palabras y tiene cierta tendencia a agrupar la metodología en un clúster y el resto de temas en otros	Proporciona núcleos de significado más ricos y más información sobre los outliers. Permite la representación simultánea de palabras y artículos. Permite la clasificación de artículos. Da más opciones al investigador

Tabla 3. Comparación del método Reinert con esta propuesta

En cuanto al objetivo de encontrar una alternativa al método Reinert, recordemos algunas de sus limitaciones; para empezar, las altas frecuencias muchas veces no están detrás de la estructura léxica real que un profesional familiarizado con el tema sabe que hay detrás del corpus (Dalud-Vincent, 2011).

El análisis aquí propuesto también tiene la limitación de usar frecuencias; sin embargo hasta cierto punto se mitigan mediante el valor de caracterización de Caballero (2011) y el HJ biplot da muchas opciones al usuario a la hora de interpretar los resultados, a la vez que puede aportar insights también acerca de los artículos o componentes del corpus.

El resultado expuesto sobre que no se puede poner fácilmente un número automático de clústers parte es una desventaja para el usuario que no está familiarizado con el corpus y quiere extraer temas sin mucho esfuerzo; pero también es una ventaja a la hora de ajustar el análisis a los datos y a las observaciones cualitativas del usuario.

4.3.4 Algunos comentarios finales sobre las limitaciones de trabajar con frecuencias

A lo largo de este trabajo, hemos visto repetidas veces que cantidad no es lo mismo que significado. Y sin embargo, dado el crecimiento exponencial actual de las masas de datos textuales a analizar, se hace absolutamente patente la necesidad de aplicar técnicas cuantitativas a muestras y cuestiones que hace tiempo se habrían considerado claramente competencias de técnicas cualitativas.

La observación interesante que se puede extraer de este caso de de estudio es que las palabras frecuentes que retiene este tipo de análisis (una vez filtradas las palabras “vacías”, como establecimos en el apartado de metodología) sí aportan significado; pero es un significado “marco” que no nos interesa, porque ya lo conocemos. En una muestra de textos científicos, claramente va a ser frecuente la palabra “artículo”. En una búsqueda sobre redes sociales, claramente va a ser frecuente la palabra “internet”. Es previsible y es obvio.

Aislar aquellas palabras que pueden aportarnos insights sobre los “subtemas” que definen el tema que queremos investigar, podría, tal vez, requerir centrarse en aquellas palabras que tienen una frecuencia “media” o incluso baja si se quiere detectar artículos que se desvían del tema. En parte este análisis supera algunas de las limitaciones de equiparar significado con frecuencia gracias al valor de caracterización de Caballero (2011). Sin embargo, cabe más exploración futura de la relación entre la frecuencia y los “niveles” de significado dentro del esquema léxico de una estructura de textos y la diferencia entre la frecuencia general respecto a la frecuencia en unos pocos artículos o “individuos” dentro del corpus.

5. **CONCLUSIONES**

- La propuesta final de las técnicas examinadas respecto al objetivo de identificar temas en una gran cantidad de textos es el uso del HJ Biplot con artículos como vectores y palabras como individuos, con datos previamente ponderados por el valor de caracterización y posterior agrupción mediante clúster de k-medias especificando un número de clústeres con consideraciones del usuario.
- La aplicación del HJ biplot a matrices de frecuencias de palabras respecto a un conjunto de corpus resulta más interesante empleando, tal vez contraintuitivamente, a los textos como variables (columnas) y a las palabras como individuos/filas porque el tipo de datos y el hecho de que haya más palabras que textos distribuye la información (variabilidad) demasiado equitativamente a lo largo de los ejes y no se puede sintetizar en unos pocos. Así, el Biplot aplicado a las palabras como individuos resulta mucho más rico e interesante en cuanto a identificar temas y orientar al menos una parte de los textos hacia sus palabras más características.
- Esto también permite la aplicación de técnicas de clasificación posteriores para aislar núcleos temáticos.
- Sin embargo, el uso de las palabras como variables y los artículos como individuos permite aislar a veces algunos artículos particulares que se salen de la búsqueda general en la forma de outliers o encontrar agrupaciones de artículos del mismo tema.
- El HJ biplot es, por su doble representación, una herramienta interesante en general a la hora de clasificar y explorar núcleos temáticos en datos textuales.
- La aplicación de un mayor número de clústeres de lo establecido por las pruebas estadísticas habituales para datos numéricos cuando usamos un clúster aplicado a este tipo de datos puede proporcionar mayor riqueza y precisión a la hora de encontrar agrupaciones interesantes de los datos.
- El clúster de K medias resulta más interesante que el clúster jerárquico a la hora de agrupar una gran cantidad de palabras en unos pocos núcleos de significado.
- El clúster jerárquico con método de ward puede resultar interesante en el caso de tener unos pocos artículos (o palabras) que se salgan fuera del tema de búsqueda por su propiedad de formar clústeres menos similares entre sí en cuanto a número de individuos.

- Los métodos de clasificación basados en densidad (el clúster DBSCAN) presentan la ventaja de no requerir del usuario el establecimiento previo de un número de clústers (lo cual es positivo si no se conoce una cantidad a priori de núcleos temáticos) pero no son muy adecuados en este tipo de análisis porque hay una gran cantidad de individuos con valores bajos (palabras de baja frecuencia) cuya relevancia cualitativa puede ser importante, y este algoritmo, precisamente por su poca sensibilidad al ruido y a los outliers, puede aglomerarlos indiscriminadamente o
- Aplicar una clasificación de clúster jerárquico a un análisis de correspondencias múltiples (Método Reinert) es más fácil a nivel de usuario y con el software ya disponible puede proporcionar insights instantáneos sobre algunos de los temas del conjunto de textos. Sin embargo, el método propuesto en este trabajo muestra una mayor riqueza de la información que proporcionan los resultados, representando mejor las palabras que no son necesariamente las más frecuentes pero que caracterizan los términos de búqueda y el contenido de los artículos.
- En referencia al punto anterior, al representar los datos mediante un HJ biplot que ofrece la posibilidad de representar los artículos a la vez que las palabras se proporciona una cierta capacidad (si bien limitada y por refinar con pruebas posteriores) de asociar ciertos artículos a ciertos temas y de agruparlos en torno a sus palabras más frecuentes, cosa que el método Reinert no permite.
- El trabajo con frecuencias de palabras (palabra más repetida=palabra más importante) - tiene limitaciones a la hora de tratar de obtener resultados cualitativos en el caso de la exploración de búsquedas bibliográficas, porque todos los textos tienen un marco temático común en cuanto a la metodología. Esto probablemente es extrapolable a cualquier tipo de conjunto de textos en el que haya un tema común con subdivisiones de significado.
- El resultado de las técnicas aplicadas es sobre todo exploratorio.

FUTUROS OBJETIVOS

- Desarrollar un método de aislamiento y exclusión de las palabras referentes a la metodología en análisis similares a este. Sería conveniente o bien identificar previamente las palabras que constituyen el marco común entre los textos sin aportar

un significado real al análisis (en este caso, por ejemplo, todas las relacionadas con la metodología, como “artículo” o “revisión”) o bien tal vez limitar algunas frecuencias altas asumiendo que se referirán a la metodología (aunque esto supone peligros propios) y quedarse con un rango de frecuencias medio o con aquellas que sean más características de las columnas en y eliminar las que se repiten con frecuencia entre todos los textos

- Explorar si la clasificación en base a densidades se puede optimizar transformando la matriz o los datos y si esto tiene ventajas a la hora de tratar con los datos que tienen mucho ruido.
- Explorar otras técnicas de clúster para este tipo de datos.
- Desarrollar software más orientado al usuario para posibilitar la aplicación de las técnicas propuestas, particularmente el HJ biplot, a datos bibliográficos y textuales en general, a personas no familiarizadas con R o con la estadística multivariante.
- Analizar, mediante sistemas de clasificación previa, hasta qué punto las técnicas propuestas son útiles a la hora de agrupar textos en torno a temas. Es decir, comparar el resultado de esta técnica con una clasificación humana manual detallada para contrastar su potencial y sus limitaciones.
- Comparar este tipo de análisis con otros análisis de topic modeling, como el Latent Dirichet Allocation.
- Repetir este tipo de análisis y aplicar estas técnicas a otras muestras para comprobar que los hallazgos propuestos se mantienen y en qué contextos/tipos de muestra.

o Bibliografía:

1. Ao, S. I., Yip, K., Ng, M., Cheung, D., Fong, P.-Y., Melhado, I., Sham, P. C. (2005). CLUSTAG: hierarchical clustering and graph methods for selecting tag SNPs. *Bioinformatics, 21*(8), 1735–1736. https://doi.org/10.1093/bioinformatics/bti201
2. Benzecri, J. P. (1973) L'Analyse des données. Tome 2: L'Analyse des correspondances. *Dunod*. ISBN 2-04-007225-X
3. Caballero, D. (2011). El HJ Biplot como herramienta de análisis de grupos de discusión *Universidad de Salamanca*
4. Caballero, D., Galindo Villardón, P., & García, M-C. (2017). JK-Meta-Biplot y STATIS Dual como herramientas de análisis de tablas textuales múltiples. *Revista de Investigação em Sistemas e Tecnologias de Informação*, 25, 18-33. DOI: 10.17013/risti.25.18–33
5. Dalud-Vincent, M. (2011). Trial and Critique of Alceste as a Tool for Analyzing Semi Structured Interviews in Sociology. *Langage et société*, 135(1), 9-28. DOI: 10.3917/ls.135.0009
6. De Alba, M. (2004). El Método ALCESTE y su Aplicación al Estudio de las Representaciones Sociales del Espacio Urbano: El Caso de la Ciudad de México. *Papers on Social Representations. 20*.
7. Eitan, T., & Gazit, T. (2023). No social media for six hours? The emotional experience of Meta's global outage according to FoMO, JoMO, and internet intensity. *Computers in Human Behavior*, 138, 107474. https://doi.org/10.1016/j.chb.2022.107474
8. Ester, M., Kriegel, H-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *KDD-96 Proceedings. AAAI Press.*
9. Fisher, J. T., Hopp, F. R., Chen, Y., Weber, R. (2023). Uncovering the structure of media multitasking and attention problems using network analytic techniques. *Computers in Human Behavior*, 147, 107829. https://doi.org/10.1016/j.chb.2023.107829
10. Forgy, E. W. (1965). Cluster analysis of multivariate data: Efficiency versus interpretability of classifications. *Biometrika, 21*(3), 768-769. JSTOR 2528559.
11. Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika, 58*(3), 453

12. Gabriel, K. R. (1972). Analysis of meteorological data by means of canonical decomposition and biplots. *Journal of Applied Meteorology*, 11
13. Galindo-Villardón, P. (1986) Una alternativa de representación simultánea: HJ Biplot. *Questiió* _10(1):13-23
14. Gower, J. C., & Hand, D. J. (1995). Biplots (Vol. 54). *CRC Press*.
15. Lam, T. K., Vartanian, O., Hollands, J. G. (2022). The brain under cognitive workload: Neural networks underlying multitasking performance in the multi-attribute task battery. *Neuropsychologia*, 174, 108350. https://doi.org/10.1016/j.neuropsychologia.2022.108350
16. MacQueen, J. B. (1967). Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, 281–297. University of California Press.
17. Mann, R. B., & Blumberg, F. (2022). Adolescents and social media: The effects of frequency of use, self-presentation, social comparison, and self-esteem on possible self-imagery. Acta Psychologica, 228, 103629. https://doi.org/10.1016/j.actpsy.2022.103629
18. Nielsen, F. (2016). Hierarchical Clustering. 10.1007/978-3-319-21903-5_8.
19. Osuna, Z. (2006). Contribuciones al Análisis de Datos Textuales. *Universidad de Salamanca*.
20. Reinert, M. (1990) a. Une méthode d'analyse des données textuelles et une application: Aurélia de G. de Nerval. *Bulletin de méthodologie sociologique*, 26, 24–54.
21. Reinert, M. (1990) b. Système A.L.C.E.S.T.E: Une méthodologie d'analyse des données textuelles présentée à l'aide d'une application. *Journées internationales d'analyse statistique des données textuelles*, Barcelona, Spain.
22. Reinert, M. (1993). Les mondes lexicaux et leur logique à travers l'analyse statistique d'un corpus d'un récits de cauchemars. *Langage et société,* 66, 5-39
23. Trefethen, L. N., Bau, D. (1997). Numerical Linear Algebra. *SIAM*. ISBN: 0898713617
24. Ward, J. H. Jr. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58, 236–244

o Software:

Charrad M., Ghazzali N., Boiteau V., Niknafs A. (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. Journal of Statistical Software, 61(6), 1-36. URL: https://www.jstatsoft.org/v61/i06/

Hahsler, M., Piekenbrock, M., & Doran, D. (2019). dbscan: Fast Density-Based Clustering with R. *Journal of Statistical Software, 91*(1), 1-30. doi:10.18637/jss.v091.i01

Microsoft Corporation. (2018). *Microsoft Excel*. Retrieved from https://office.microsoft.com/excel

R Core Team. (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/

Ratinaud, P. (2009). IRAMUTEQ: Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires. Un logiciel libre construit avec des logiciels libres

Vicente-Villardon J.L. (2021). MultBiplotR: Multivariate Analysis Using Biplots in R (R package version 1.3.30). URL: https://CRAN.R-project.org/package=MultBiplotR.