

UNIVERSIDAD DE SALAMANCA

Departamento de Estadística

Doctorado en Estadística Multivariante Aplicada



**Integración de Minería de Texto y Técnicas
Multivariantes en el Entorno Digital, aplicado al Análisis
Organizacional PESTEL**

Autor:

Luis Manuel Pilacuan Bonete

Directores:

Dra. M. Purificación Galindo Villardón

Dr. Francisco Javier Delgado Álvarez

2023

**Integración de Minería de Texto y Técnicas
Multivariantes en el Entorno Digital, aplicado al Análisis
Organizacional PESTEL**



**DEPARTAMENTO DE ESTADÍSTICA
UNIVERSIDAD DE SALAMANCA**

Memoria para optar al Grado de Doctor,
por el Departamento de Estadística de
la Universidad de Salamanca, presenta:

Luis Manuel Pilacuan Bonete

Salamanca

2023



**DEPARTAMENTO DE ESTADÍSTICA
UNIVERSIDAD DE SALAMANCA**

Purificación Galindo Villardón

Catedrática del Departamento de Estadística de la Universidad de Salamanca

Delgado Álvarez Francisco Javier

Profesor del Departamento de Estadística de la Universidad de Salamanca

CERTIFICAN:

Que Don Luis Manuel Pilacuan Bonete ha realizado, en la Universidad de Salamanca, bajo su dirección, el trabajo titulado “Integración de Minería de Texto y Técnicas Multivariantes en el Entorno Digital, aplicado al Análisis Organizacional PESTEL”, para optar al título de Doctor en Estadística Multivariante Aplicada, autorizando expresamente su lectura y defensa.

Y para que conste, firman el presente certificado en Salamanca a 31 de mayo de 2023.

Dra. Purificación Galindo Villardón

Dr. Delgado Álvarez Francisco

AGRADECIMIENTO

Sin duda alguna un pilar fundamental durante este proceso de doctorado ha sido la Doctora Purificación Galindo Villardón; quién ha compartido sus valiosos conocimientos conmigo, aportando un gran valor a mis estudios doctorales. Además de la parte académica, de ella no solo obtuve sabiduría, sino también una amiga invaluable con la he podido compartir muchos momentos memorables que de seguro me servirán para el futuro.

No podía faltar el agradecimiento al Doctor Francisco Delgado Álvarez, quien me ha acompañado en cada paso de mi tesis doctoral. Su vasta experiencia en el área estadística fue un soporte fundamental en mi formación como doctor, lo cual me permitirá transmitir los conocimientos adquiridos dentro de mi alma máter la Universidad de Guayaquil.

A la Universidad de Guayaquil, por su respaldo y apoyo.

A todos mis compañeros y amigos del Doctorado de la USAL.

¡Gracias a todos!

DEDICATORIA

Doy gracias a Dios, quién me ha prestado salud y fortaleza para completar este arduo proceso. Además, ha sido el proveedor de todo lo necesario para finalizar con éxito esta meta.

A mi esposa Pilar Macías, quien ha sido mi soporte durante este tiempo, ya que ha cumplido su labor de madre, esposa y amiga durante este tiempo.

A mis hijos Santiago y Abdiel, mi motor para continuar con mis estudios y darles el ejemplo necesario para criar hombres de bien para el futuro y sobre todo por esperar con paciencia para retomar el tiempo junto a ellos.

A mis padres, Luis y Liliam, quienes siempre me han apoyado a cumplir con mis sueños y metas propuestas.

“La industria sin estadística es como un
barco sin timón, a la deriva en un mar de
incertidumbre.”

Edwards Deming

CONTENIDO

CONTENIDO

<u>PREÁMBULO.....</u>	<u>XVI</u>
------------------------------	-------------------

<u>CAPÍTULO I.....</u>	<u>20</u>
-------------------------------	------------------

<u>1. MINERÍA DE DATOS TEXTUALES.....</u>	<u>20</u>
--	------------------

1.1. INTRODUCCIÓN	21
1.2. MINERÍA DE DATOS TEXTUALES	22
1.2.1. DEFINICIÓN.....	22
1.2.2. LÍNEA DE TIEMPO DE LA MINERÍA DE TEXTO.....	23
1.3. PROCESO DE LA MINERÍA DE DATOS TEXTUALES	25
1.3.1. RECUPERACIÓN DE LA INFORMACIÓN.....	27
1.3.2. DATOS TEXTUALES.....	29
1.3.3. PREPROCESAMIENTO DE DATOS TEXTUALES.....	31
1.3.3.1. TRANSFORMACIÓN DE TEXTO.....	31
1.3.3.2. TOKENIZACIÓN.....	31
1.3.3.3. ELIMINAR PALABRAS VACÍAS.....	32
1.3.3.4. NORMALIZACIÓN DE LAS PALABRAS.....	32
1.4. ESTRUCTURA DEL TEXTO	36
1.4.1. N-GRAMAS	36
1.4.2. CONCEPTO.....	38
1.4.3. TESAURO	38
1.4.4. REDES SEMÁNTICAS	39
1.5. REPRESENTACIÓN MATRICIAL DEL CORPUS.....	41
1.5.1. PONDERACIÓN DE PALABRAS.....	43

<u>CAPÍTULO II</u>	<u>47</u>
---------------------------------	------------------

<u>2. TÉCNICAS PROBABILÍSTICAS DE LA MINERÍA DE TEXTO.....</u>	<u>47</u>
---	------------------

2.1. INTRODUCCIÓN	48
2.2. ANALISIS SEMANTICO LATENTE	48
2.3. DESCOMPOSICIÓN DE VALORES SINGULARES.....	50
2.3.1. ESPACIO SEMÁNTICO	52
2.4. MODELADO DE TÓPICOS.....	53

2.4.1. ANÁLISIS PROBABILÍSTICO SEMÁNTICA LATENTE.....	55
2.4.2. ASIGNACIÓN LATENTE DE DIRICHLET	58
2.4.2.1. INTERPRETACIÓN GEOMETRICA DEL LDA	59
2.4.2.2. MODELO GENERATIVO LDA	60
2.4.2.3. INFERENCIA EN LDA.....	65

CAPÍTULO III.....67

3. METODOS BILOT67

3.1. INTRODUCCIÓN	68
3.2. LOS MÉTODOS BILOT.....	68
3.2.1. BILOT CLÁSICOS	69
3.2.1.1. GH-BILOT.....	70
3.2.1.2. JK-BILOT	70
3.2.2. HJ-BILOT	71
3.2.3. INTERPRETACIÓN HJ BILOT	73
3.3. BILOT EN EL ANÁLISIS DE DATOS TEXTUALES.....	74

CAPITULO IV78

4. PESTEL78

4.1. INTRODUCCIÓN.....	79
4.2. PESTEL.....	79
4.3. FACTORES PESTEL.....	80

CAPÍTULO V.....82

5. HJ-BILOT COMO HERRAMIENTA PARA DAR UN IMPULSO ANALÍTICO ADICIONAL AL MODELO DE ASIGNACIÓN LATENTE DE DIRICHLET82

5.1. INTRODUCCIÓN.....	83
5.2. METODOLOGÍA LDABILOTS.....	83
5.3. INTEGRACIÓN DE MÉTODOS BILOT AL ANÁLISIS DEL MODELADO DE TÓPICOS LDA	86
5.3.1. TRATAMIENTO DE MATRICES DEL LDA	86

5.4. HJ-BILOT COMO REPRESENTACIÓN DE MATRICES DE PROBABILIDAD DEL LDA.	87
---	----

CAPÍTULO VI.....90

6. LDABILOTS.....90

6.1. INTRODUCCIÓN	91
6.2. PAQUETE LDABILOTS.....	91
6.2.1. DESCRIPCIÓN DEL LDABILOTS	94
6.2.1.1. OBTENCIÓN DE INFORMACIÓN	94
6.2.1.2. PREPROCESAMIENTO DEL CORPUS	96
6.2.1.3. INFERENCIA DE TÓPICOS	99
6.2.1.4. OBTENCIÓN DEL MODELO LDA	101
6.2.1.5. REPRESENTACIONES BILOT	103
6.2.1.6. REPRESENTACIÓN HJ-BILOT DEL ENTORNO PESTEL	105

CAPÍTULO VII.....108

7. APLICACIÓN PRÁCTICA LDABILOTS.....108

7.1. INSTALACIÓN DEL PAQUETE LDABILOTS	109
7.2. MENÚ IMPORTACIÓN DE DATOS	109
7.3. MENÚ VISUALIZACIÓN DE MATRIZ DTM	113
7.4. MENÚ INFERENCIA DE TÓPICOS.....	114
7.5. MENÚ LDA Y BILOT	116
7.5.1. MENÚ PROCESO LDA.....	116
7.5.2. MENÚ REPRESENTACIONES BILOT.....	125
7.5.3. MENÚ HJ-BILOT_PESTEL.....	129

CONCLUSIONES.....138

CONTRIBUCIONES CIENTÍFICAS.....141

BIBLIOGRAFÍA.....159

LISTADO DE FIGURAS

<i>Figura 1.1. Evolución de la Minería de Datos Textuales.</i>	25
<i>Figura 1.2. Secuencia Hipotética de exploración de colección de textos</i>	26
<i>Figura 1.3. Aplicación de la MDT, según Miner (2012)</i>	26
<i>Figura 1.4. Esquema de MDT, según Kodratoff (1999).</i>	27
<i>Figura 1.5. Representación de la extracción de Información Textual</i>	29
<i>Figura 1.6. Contenido del Corpus textuales.</i>	30
<i>Figura 1.7. Ejemplo contenido diccionario para Lematización</i>	34
<i>Figura 1.8. Sección del Diccionario para Lematización de Mechura (2016)</i>	34
<i>Figura 1.9. Palabras originales de un documento después de Stemming y Lematización</i>	36
<i>Figura 1.10. Ejemplo de una red semántica de palabras.</i>	40
<i>Figura 1.11. Tipos de Matrices de Documentos.</i>	41
<i>Figura 2.1. Proceso de la SVD, basado en Martin y Berry (2007).</i>	51
<i>Figura 2.2. Representación de Modelado de Tópicos, basado en Steyvers & Griffiths (2007).</i>	54
<i>Figura 2.3. Representación de Placas del PLSA simétrico, Basado en Hofmann (1999).</i>	57
<i>Figura 2.4. Representación Geométrica del LDA</i>	60
<i>Figura 2.5. Enfoque del proceso generativo del modelado de tópicos (Steyvers & Griffiths, 2007)</i>	61
<i>Figura 2.6. Representación de Placas de LDA, basado en Blei & Lafferty, 2009</i>	63
<i>Figura 3.1. Representación Gráfica de Biplot de Gabriel (1971).</i>	69
<i>Figura 3.2. Marcadores de los Biplot de Gabriel (1971).</i>	71
<i>Figura 3.3. Marcadores Fila y Columna del HJ-Biplot. (Cubilla Montilla, 2019).</i>	72
<i>Figura 3.4. Interpretación Gráfica del HJ-Biplot. Tomado de (Ballesteros, 2022)</i>	74
<i>Figura 3.5. Publicaciones en Scopus y WOS de USAL en MDT a abril 2023.</i>	74
<i>Figura 3.6. Instituciones con Publicaciones en WOS, referentes a MDT a abril 2023.</i>	75
<i>Figura 3.7. Áreas de Publicaciones en MDT de WOS a abril 2023</i>	77
<i>Figura 4.1. Factores PESTEL.</i>	81
<i>Figura 5.1. Representación Metodología LDABILOTS</i>	85
<i>Figura 5.2. Esquema de Obtención de Theta y Phi del LDA de Blei (2003)</i>	87
<i>Figura 5.3. Esquema de transformación de matriz theta, para el HJ-Biplot.</i>	88
<i>Figura 6.1. LDABiplots publicado en CRAN de R</i>	93
<i>Figura 6.2. Código de Webscraping del LDABiplots.</i>	95

Figura 6.3. Estructura de formato xlsx a Importar	96
Figura 6.4. Código para creación de tablas del texto extraído de la Web.	97
Figura 6.5. Código de preprocesamiento del LDABiplots	98
Figura 6.6. Sección de Código para la Inferencia de K.	101
Figura 6.7. Código del LDABiplots para obtener el LDA del corpus.	102
Figura 6.8. Función HJ-Biplot del LDABiplots	104
Figura 6.9. Función GH-Biplot del LDABiplots	104
Figura 6.10. Función JK-Biplot del LDABiplots	104
Figura 6.11. Sección de Código de generación de Biplot del paquete.	105
Figura 6.12. Sección del Léxico PESTEL	107
Figura 6.13. Código de generación de clúster PESTEL en el HJ-Biplot	107
Figura 7.1. Códigos de Instalación de LDABiplots en RStudio.	109
Figura 7.2. Pantalla de Menú del LDABiplots	109
Figura 7.3. Página web El Mundo. Web Scraping de Etiqueta <a>	110
Figura 7.4. Página web 20Minutos. Web Scraping de Etiqueta <a>	110
Figura 7.5. Página web El País. Web Scraping de Etiqueta <h2>	111
Figura 7.6. Importación de datos desde archivo xlsx	111
Figura 7.7. Información de Datos textuales extraídos.	111
Figura 7.8. Parámetros para obtener DTM en el LDABiplots.	112
Figura 7.9. DTM obtenida en el LDABiplots.	112
Figura 7.10. Barplot de términos del DTM obtenido en LDABiplots.	113
Figura 7.11. Wordcloud de 31 términos del DTM obtenido en LDABiplots.	113
Figura 7.12. Coocurrencia de términos del DTM obtenido en LDABiplots.	114
Figura 7.13. Inferencia de K para el modelo LDA en el LDABiplots	115
Figura 7.14. Menú de análisis LDA, Biplot, PESTEL	116
Figura 7.15. Parámetros para obtener el LDA.	117
Figura 7.16. Matriz de Resumen del Modelo LDA.	118
Figura 7.17. Matriz Theta del LDA obtenida en el LDABiplots.	119
Figura 7.18. Matriz PHI del LDA obtenida en el LDABiplots.	119
Figura 7.19. Nube de Palabras por tópico.	120
Figura 7.20. Mapa de Calor de la Probabilidad de que los Diarios pertenezcan a cada tópico	121
Figura 7.21. Selección de Método de distancia para los clústeres de los tópicos del LDA.	122

<i>Figura 7.22. Menú de selección de tipo de Clúster.</i>	123
<i>Figura 7.23. Clúster rectangular de los tópicos del LDA.</i>	123
<i>Figura 7.24. Clúster circular de los tópicos del LDA.</i>	124
<i>Figura 7.25. Clúster filogenético de los tópicos del LDA.</i>	124
<i>Figura 7.26. Menú Representaciones Biplot.</i>	125
<i>Figura 7.27. Valores Propios del HJ-Biplot de Matriz Theta del LDA.</i>	126
<i>Figura 7.28. Varianza explicada por el HJ-Biplot de Matriz Theta del LDA.</i>	126
<i>Figura 7.29. Valores de Cargas obtenidos en el HJ-Biplot de Matriz Theta del LDA.</i>	126
<i>Figura 7.30. Coordenadas de los tópicos en el HJ-Biplot.</i>	127
<i>Figura 7.31. Coordenadas de los Diarios en el HJ-Biplot.</i>	128
<i>Figura 7.32. Representación gráfica HJ-Biplot de matriz theta del LDA.</i>	128
<i>Figura 7.33. Menú de opciones para representaciones gráficas Biplot.</i>	129
<i>Figura 7.34. Menú HJ-Biplot_PESTEL de la Matriz Phi del LDA</i>	130
<i>Figura 7.35. Representación LDA_HJ-Biplot_PESTEL de las palabras con respecto a los tópicos obtenidos.</i>	130
<i>Figura 7.36. Contribuciones de los Tópicos en el LDA_HJ-Biplot_PESTEL</i>	131
<i>Figura 7.37. Parte de las contribuciones de las Palabras en el LDA_HJ-Biplot_PESTEL</i>	132
<i>Figura 7.38. Proporción de contenido PESTEL por Tópicos.</i>	133
<i>Figura 7.39. Términos Políticos por Tópicos.</i>	134
<i>Figura 7.40. Términos Económicos por Tópicos</i>	134
<i>Figura 7.41. Términos Sociales por Tópicos</i>	135
<i>Figura 7.42. Términos Tecnológicos por Tópicos.</i>	135
<i>Figura 7.43. Términos Ambientales por Tópicos.</i>	136
<i>Figura 7.44. Términos Legales por Tópicos.</i>	136

LISTADO DE TABLAS

<i>Tabla 1.1. Preprocesamiento de texto de Documentos.</i>	42
<i>Tabla 1.2. Matriz TDM resultante de Tabla 1.1.</i>	42
<i>Tabla 1.3. Estructura de Matriz TDM.</i>	43
<i>Tabla 3.1. Bondad de Ajuste de Biplot de Gabriel (1971), y de Galindo (1986).</i>	73
<i>Tabla 7.1. Etiqueta de Tópicos obtenidos en el LDA</i>	118

CONTRIBUCIONES CIENTÍFICAS

Contribución Científica 1. Publicación publicada en revista científica Mathematics 142

PREÁMBULO

PREÁMBULO

La presente tesis doctoral aborda, desde lo fundamental, el estudio y aplicabilidad del análisis estadístico de datos textuales (AEDT) a partir de la minería de datos de texto (MDT); atendiendo a las técnicas de investigación más activas a nivel mundial, así como a nuevas perspectivas en el área adelantadas en el Departamento de Estadística de la Universidad de Salamanca.

Cada vez es más frecuente un notable aumento en publicaciones, a nivel mundial, que dan cabida al tratamiento de datos textuales en diferentes disciplinas. Al respecto, diversos estudios presentan a consideración de la comunidad científica, la aplicación de distintos enfoques metodológicos para la adquisición, estructuración y análisis de conocimiento a partir de información obtenida desde repositorios digitales en la web.

Efectivamente, múltiples metodologías se han desarrollado entorno al AEDT. Se remontan desde las generadas por la escuela francesa, donde se postuló el análisis factorial de correspondencia (AFC) para estudiar las tesis de Chomsky sobre la lengua (Benzécri, 1964). Continuando con técnicas como las desarrolladas por la escuela anglosajona, como el análisis semántico latente (LSA) (Deerwester et al., 1990), el cual incorpora la semántica latente de los textos analizados.

En la actualidad, con el incremento de aportes en relación con las técnicas en el campo del aprendizaje automático, la escuela americana ha desarrollado la técnica conocida como Asignación Latente de Dirichlet (LDA) (D. M. Blei et al., 2003). Se trata de un método de aprendizaje no supervisado utilizado para descubrir tópicos ocultos en grandes conjuntos de datos, usándose en el campo de la minería de datos textuales, análisis de sentimientos y recuperación de información.

En correspondencia, la presente investigación asume como propósito fundamental el desarrollo de una estrategia metodológica basado en los métodos Biplot para dar un impulso analítico al modelo de Asignación Latente de Dirichlet, integrando la adquisición de información a partir del entorno digital Web, con aplicación al análisis organizacional PESTEL.

De esta manera, nuestra investigación pretende contribuir con el desarrollo de una aplicación escrita en lenguaje R (Posit, 2023; R Development Core Team, 2000), denominada LDABiplots (Pilacuan-Bonete, Galindo-Villardón, Delgado-Álvarez, et al., 2022). Destacamos especialmente la utilización del HJ-BIPLLOT, que permite generar representaciones Biplot de las matrices de probabilidad transformadas mediante el cálculo de una medida de centralidad del modelado de tópicos LDA, a partir del procesamiento de los datos no estructurados y extraídos desde la web de noticias de Google e integrando el análisis del entorno organizacional PESTEL al HJ-Biplot. Esto representa una ventaja significativa, porque se constituye en una representación conjunta de filas o sujetos objeto de estudio y columnas o variables de estudio. Por ende, proporciona una representación visual intuitiva de la estructura del modelo, permitiendo identificar patrones y tendencias ocultas y ayudando en la selección de términos o palabras, así como de documentos relevantes.

En correspondencia a estas premisas, se ha configurado la estructura de esta tesis de la manera siguiente:

Capítulo I. Minería de Datos Textuales; se describen los principales procesos de la minería de datos textuales. A su vez, se desarrollan conceptos y técnicas para la extracción, procesamiento y transformación de los datos textuales importantes para el presente estudio.

Capítulo II. Técnicas Probabilísticas de la Minería de Texto; el mismo, está orientado a describir las principales técnicas probabilísticas para el análisis de datos textuales aplicando el aprendizaje automático. De la misma manera, se describe el proceso para la generación de tópicos, operaciones algebraicas, propiedades de las matrices y modelo generativo de la asignación LDA.

Capítulo III. Métodos Biplot; presenta los Biplot clásicos propuesto en 1971 por Gabriel, y el propuesto por Galindo en 1986; usados en el presente estudio para el análisis de datos textuales.

Capítulo IV. PESTEL; se describen los conceptos de la metodología PESTEL usada para la evaluación de los factores externos que afectan a la organización, y como se aplicara en esta tesis.

Capítulo V. HJ-Biplot como herramienta para dar un Impulso analítico adicional al modelo de Asignación Latente De Dirichlet; desarrolla la descripción del proceso metodológico que integra la extracción de datos textuales desde la web, procesamiento del texto y modelado de tópicos; con la finalidad de generar representaciones Biplot de las matrices transformadas del modelo LDA. A ello, se suma la inferencia sobre cómo generar agrupaciones integrando listado de palabras a las representaciones Biplot.

Capítulo VI. LDABiplots; se describe el paquete desarrollado en lenguaje R, su funcionamiento y explicación de la metodología integrada en el software. De la misma manera, se explica la integración del listado PESTEL, para la agrupación de las palabras en la representación HJ-Biplot, así como las funciones integradas en el paquete para el tratamiento de los datos textuales.

Capítulo VII. Aplicación Práctica LDABiplots, presenta la aplicación práctica de la metodología propuesta, con base a las técnicas analizadas. Estas, han sido codificadas en el paquete LDABiplots.

CAPÍTULO I

MINERÍA DE DATOS TEXTUALES

1.1. INTRODUCCIÓN

En la actualidad existen múltiples plataformas tecnológicas que permiten el acceso a datos, en particular, a datos textuales. En este orden, con la masificación de redes sociales, prensa digital, digitalización de documentos y avances en las comunicaciones, se vuelve imprescindible el empleo de herramientas tecnológicas y técnicas estadísticas que aporten ventajas significativas en la interpretación y análisis de grandes cantidades de datos textuales, así como en la extracción de conocimiento a partir de información relevante.

Desde esta circunstancia, existe la necesidad de contar con información relevante por parte de investigadores y organizaciones, por lo cual, es importante procesar estas masas de datos textuales mediante una gestión tecnológica que permita el procesamiento masivo de texto. Esto se lleva a cabo por medio de técnicas de procesamiento de lenguaje natural (PLN), que forman parte del aprendizaje automático conocido en inglés como machine learning, de una manera ágil y rápida, para convertirlos en conocimiento, con tiempos de análisis reducidos, que faciliten la rápida toma de decisiones.

Este apartado, reviste especial relevancia, al desvelar aspectos significativos de la minería de datos textuales. Dados estos escenarios, Berry y Kogan destacan un abanico de herramientas para la minería de datos textuales que buscan extraer conocimiento a partir del texto incorporado en diferentes plataformas físicas y digitales (Berry & Kogan, 2010). Entre ellas, destacan: a) PLN, que es considerada una técnica semiestructurada de la inteligencia artificial que se enfoca en la comprensión y generación del lenguaje humano; b) Modelado de tópicos o temas, que extrae los temas principales en los conjuntos de texto y; c) Análisis de redes semánticas, que implica el análisis semántico de las palabras.

Siendo uno de los métodos no supervisados más relevantes, el modelado de tópicos no necesita información previa, como etiquetas para generar una comprensión más profunda de los textos analizados. Sin embargo, el procesamiento de los datos textuales es similar a otras técnicas, antes de la generación del modelo a partir de estos datos.

Partiendo de estas premisas, se aborda el proceso de datos textuales, en un recorrido que incluye la disertación de sus distintas fases: extracción, tratamiento, transformación y representación de estos; considerando tiempos y recursos computacionales necesarios, ya

que el análisis de grandes conjuntos de datos puede demandar una infraestructura de procesamiento más complejo, como la computación de alto rendimiento o el uso de bases de datos almacenadas en la nube, así como una gran cantidad de recursos invertidos.

De esta manera, se abordarán los postulados teóricos que fundamentan a la minería de datos textuales y su línea de tiempo. Así mismo, se acometerá un análisis sobre el proceso que incluye la recuperación de la información, datos textuales y los aspectos que involucran su preprocesamiento. A ello, se suma el estudio de la estructura del texto y sus aspectos vinculantes, así como las temáticas relacionadas a la representación matricial del corpus.

1.2. MINERÍA DE DATOS TEXTUALES

1.2.1. DEFINICIÓN

En sus inicios, la minería de datos textuales (MDT) nace como un proceso de clasificación y agrupamiento de textos para el reconocimiento de patrones, mediante la investigación en el texto contenido en libros (Luhn, 1958). Su evolución hasta la actualidad es asumida como una técnica metodológica que permite la extracción automática de nuevo conocimiento, basado en una gran cantidad de datos textuales a partir de un conjunto de información textual no procesada como texto de libros, correos electrónicos, redes sociales, cartas, entre otros posibles documentos (Hearst, 1999).

Esta concepción, según Hearst, involucra las siguientes etapas.

- Recuperación de la Información, la cual tiene como objeto identificar documentos de interés dentro de un grupo de documentos textuales (Luhn, 1958), a partir de una representación formal y con la ayuda de métodos computarizados (Choueka, 1980). Involucra la extracción de contenido textual desde varios medios físicos digitalizados o desde formatos no estructurados web, con asistencia computacional (Gayo-Avello et al., 2004).

- Tratamiento Estadístico de la información textual recuperada, que utiliza técnicas estadísticas para el tratamiento de los corpus de texto, generando conocimiento estructurado o no estructurado (Lebart & Salem, 1988). En este sentido, cuando se tienen grandes

cantidades de datos textuales, los sistemas computacionales permiten la generación de resultados con mayor rapidez y fiabilidad, conforme avanzan los años (Kostoff et al., 2001).

- Análisis del texto, donde se evalúa el rendimiento de la información, para valorar si los resultados son proporcionados de forma coherente y relevante (Hearst, 1999).

La definición de la MDT ha variado a través del tiempo, con la incorporación de nuevas tecnologías y los avances computacionales de la inteligencia artificial (Chellappandi & Vijayakumar, 2018; Viera & Viera, 2017) y se ha definido como aquella técnica que busca encontrar patrones en el texto, por medio del uso de algoritmos de identificación y extracción de patrones de datos no estructurados desde medios físicos y digitales. También se puede definir la MDT como una técnica de agrupamiento de un gran número de documentos, para el análisis de similitud y coherencia entre palabras, creando categorías no definidas, permitiendo organizar grupos definidos para los documentos y generando información nueva a partir de la analizada.

1.2.2. LÍNEA DE TIEMPO DE LA MINERÍA DE TEXTO.

La minería de texto puede concebirse como una combinación de elementos computacionales y humanísticos que permiten la conversión de textos en datos cuantitativos y analizarlos por medio de la tecnología.

Este proceso tiene sus inicios entre 1940 y 1960, donde se desarrollaron aspectos muy significativos, como la descripción del lenguaje natural asistido por computadoras (Luhn, 1958), así como el análisis de varios tipos de contenido de libros, discursos, entre otros (Doyle, 1961).

Posteriormente, para la década de los 80, se incorporó la semántica latente, la cual supone que ciertas palabras aparentemente independientes estén relacionadas por temas subyacentes no observados. Es introducido como un método de reducción de dimensiones e identificación de los factores latentes aplicado a datos textuales (Dumais et al., 1988). Al respecto y con la incorporación del conocimiento en bases de datos, se configuraron condiciones para ordenar y dar sentido a los mismos (Feldman & Dagan, 1995).

A ello se suman los avances en la década de los 90 que produjeron la creación de conocimiento en un marco de texto estructurado. Surge así la minería de texto con base en el conocimiento de bases de datos textuales (Fayyad et al., 1996), a las cuales se sumó la aplicación de métodos de aprendizaje automático (Dong & Agogino, 1997; Hearst, 1999), ganando protagonismo y aplicabilidad en diferentes ciencias, y generando otros avances a partir del 2000, como el modelamiento de tópicos (D. M. Blei et al., 2001; Steyvers & Griffiths, 2007) y el análisis de sentimiento (Pang et al., 2002).

Desde este contexto, es a partir del año 2010 cuando se han podido aplicar estos métodos en grandes bases de datos, gracias al avance computacional (Shayaa et al., 2018). Estos avances fueron combinados con técnicas de Inteligencia Artificial, como el aprendizaje profundo o Deep Learning, permitiendo el desarrollo de aplicaciones tales como redes neuronales recurrentes aplicadas a texto (J. Zhang & Zong, 2015), que brindan la opción de guardar memoria y procesar secuencias de palabras para la generación de texto o traducción de palabras, así como el uso de redes neuronales convolucionales (Severyn & Moschitti, 2015), las cuales, en sus principios, estaban concebidas para el tratamiento de imágenes, lo que generó la complementariedad texto-imágenes.

Por otro lado, con el avance de investigaciones en la inteligencia artificial, se han podido generar modelos de lenguaje pre-entrenados, permitiendo la extracción, procesamiento y entrenamiento de grandes cantidades de datos textuales con el fin de ajustar el texto a tareas específicas, como chatbots, traducción, generación de texto automático (Topal et al., 2021). Al respecto, se destacan dos aportes muy significativos.

Desde esta perspectiva, entre los modelos usados con mayor frecuencia en la actualidad, tenemos el de Transformadores Bidireccionales Profundos para la comprensión del Lenguaje, conocido en inglés como Bidirectional Encoder Representations from Transformers (BERT). Este desarrollo, creado por Google, permite entrenar previamente texto sin etiquetas usando técnicas del PLN para crear modelos de generación de texto, utilizando información del contexto y de la relación de palabras (Devlin et al., 2018). Permite la generación de representaciones contextuales de palabras, abordando la polisemia del lenguaje; así también permite clasificación de texto, con el fin de ser usado para resumen de texto, respuestas de preguntas, entre otras funciones.

Así mismo, se cuenta con los aportes ofrecidos por el modelo ChatGPT, desarrollado por OpenIA, que funciona utilizando una red neuronal profunda basada en la arquitectura GPT 3-5. ChatGPT tiene la característica de actualizarse y mejorarse continuamente de manera automática a medida que el usuario utiliza el modelo (OpenAI, 2023). Con funcionalidades similares al BERT, este modelo se caracteriza por la generación de texto autónomo y coherente, conectando preguntas o textos anteriormente usados por el usuario para dar respuestas más coherentes al contexto. En la figura 1.1, se puede apreciar esta evolución de la minería de texto con sus avances.

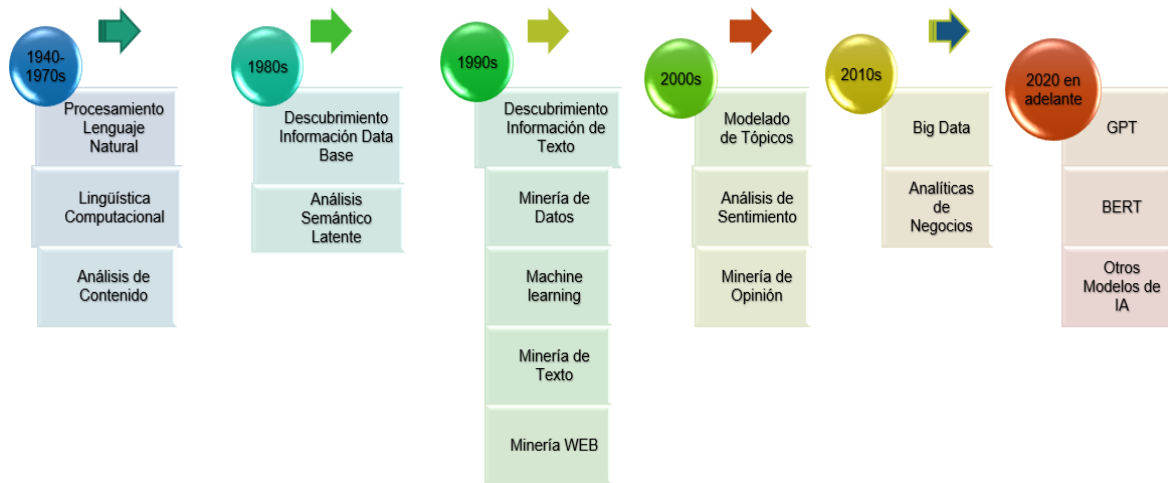


Figura 1.1. Evolución de la Minería de Datos Textuales.

1.3. PROCESO DE LA MINERÍA DE DATOS TEXTUALES

Partiendo de los avances y evolución de la MDT en todo este tiempo, se han realizado aportes en el desarrollo de los procesos vinculados a la extracción, procesamiento y análisis de texto (Billheimer et al., 2003; Weiss et al., 2005). En este sentido, cabe destacar que la estructura de un proceso para la minería de texto no se ha podido estandarizar, como consecuencia de la amplitud del rango de aplicación y su disparidad, por lo cual, varios autores han generado procesos a partir de la coherencia y sentido práctico.

En este orden de ideas, Hearst (Hearst, 1999), fundamentó un proceso donde establece diferencias entre: a) la recuperación de la información, cuyo objeto es ayudar al usuario a la búsqueda de la información previamente conocida incluida en los documentos y, b) La lingüística computacional, la cual se enfoca en el estudio de los resultados estadísticos obtenidos para el descubrimiento de patrones. En la figura 1.2 se observa un ejemplo

planteado por Hearst de como a partir de la ayuda computacional se puede extraer palabras de múltiples documentos.

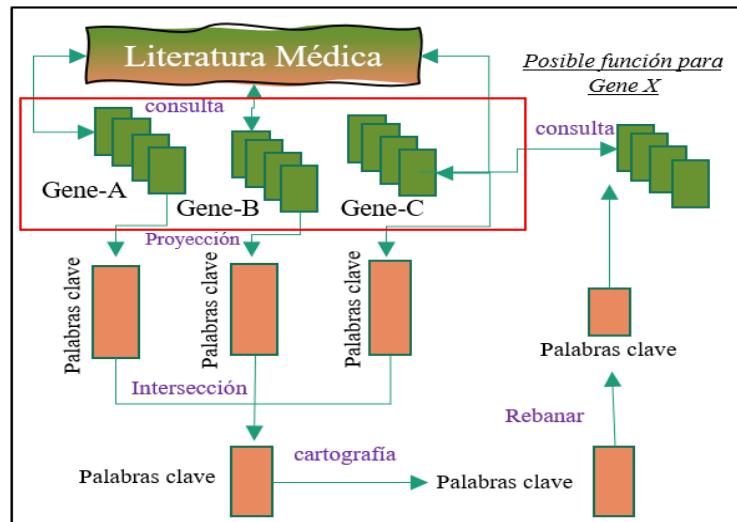


Figura 1.2. Secuencia Hipotética de exploración de colección de textos

Miner, por su parte, propuso un enfoque en el que se consideran siete posibles áreas prácticas de aplicación de la MDT, como la extracción de conceptos, minería web, recuperación de la información, agrupamiento y clasificación de documentos, extracción de Información y el procesamiento del lenguaje natural (Miner et al., 2012)

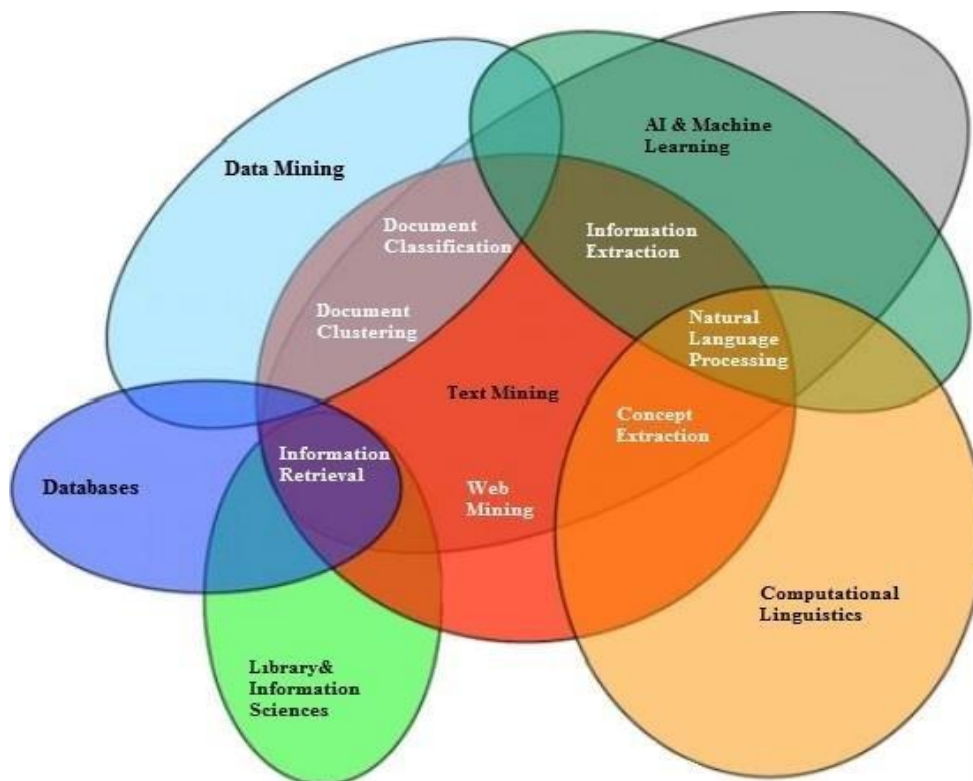


Figura 1.3. Aplicación de la MDT, según Miner (2012)

En la figura 1.4 se esquematiza un proceso de MDT, obtenido a partir de las investigaciones de Kodratoff y Tan (Kodratoff, 1999; Tan, 1999)

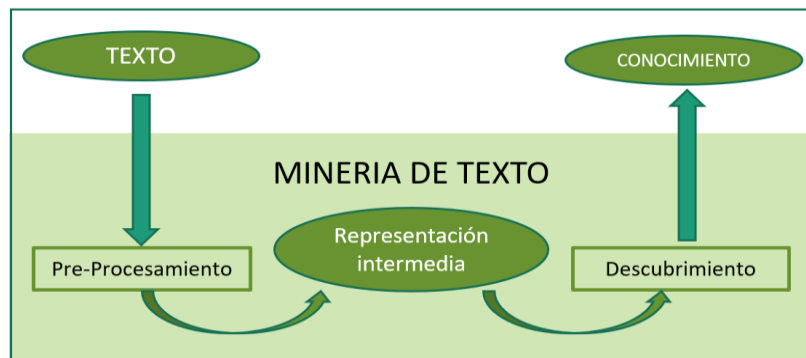


Figura 1.4. Esquema de MDT, según Kodratoff (1999).

Basándonos en lo descrito, el presente estudio esquematiza el proceso de la MDT, basándonos en la propuesta de Kodratoff (1999), enfatizando que los textos pueden ser recuperados de medios digitales o desde la Web. En este capítulo se analiza el procesamiento del texto para poder ser transformado en datos estructurados previo al descubrimiento del conocimiento.

1.3.1. RECUPERACIÓN DE LA INFORMACIÓN

Como primer paso en la MDT, tenemos la recuperación o extracción de la información (IR Information Recovery), que es parte del proceso de descubrimiento del conocimiento (KD Knowledge Discovery). Al respecto, la IR es usada en la MDT para descubrir conocimiento de base de datos textuales (KDT Knowledge Discovery Textual) a partir de una gran cantidad y variedad de información. (Liddy, 1998).

En este sentido, tiene el propósito de extraer de diferentes fuentes de datos los documentos que satisfacen la necesidad de información de la investigación. Ello involucra la capacidad de los sistemas para representar correctamente el contenido de los documentos, esquemas de ponderación apropiados e incluir las consultas apropiadas de los usuarios, considerando las consultas complejas de estos (Liddy, 1998).

Así, la IR se aplica desde varias bases o fuentes de datos que contienen texto natural no estructurado de manera libre, como enciclopedias, libros de texto, periódicos, entre otros;

por lo que, procesos de recuperación, integran algoritmos que permiten convertir estos datos obtenidos de las fuentes en estructurados. Al respecto, en el capítulo 6 se precisa con más detalle algunos de estos algoritmos de extracción.

En este contexto, la mayor información textual se concentra en bases multimedia o portales digitales, como la World Wide Web (WWW). Así, para la recuperación de la información de estas fuentes se requiere de diferentes algoritmos que permitan poder estructurar de acuerdo con el conocimiento requerido en la investigación (Baeza-Yates & Ribeiro-Neto, 1999).

En la actualidad, los algoritmos más comunes tienen aplicaciones estadísticas, dado que en los enfoques actuales de IR se pueden deducir como estadísticos, ya que permiten estructurar los datos de manera vectorial, probabilística, inferencial, con una combinación de lingüística computacional (Liddy, 1998).

En la figura 1.5 se observan las diferentes bases de donde puede ser extraída la información; la misma que se puede obtener de repositorios digitales ordenados, así como de páginas web por medio de métodos denominados *web scraping*, donde esta información extraída se encuentra no estructurada, y que luego se convertirá en un *corpus* estructurado el cual se puede preprocesar para poder proceder con los análisis requeridos.

La extracción web y de medios digitalizados son los procesos que nos van a permitir desarrollar los diferentes métodos objeto de estudios de este trabajo. El proceso metodológico de la extracción de texto desde la web y de medios digitales será revisado con un ejemplo de aplicación en el capítulo 7.

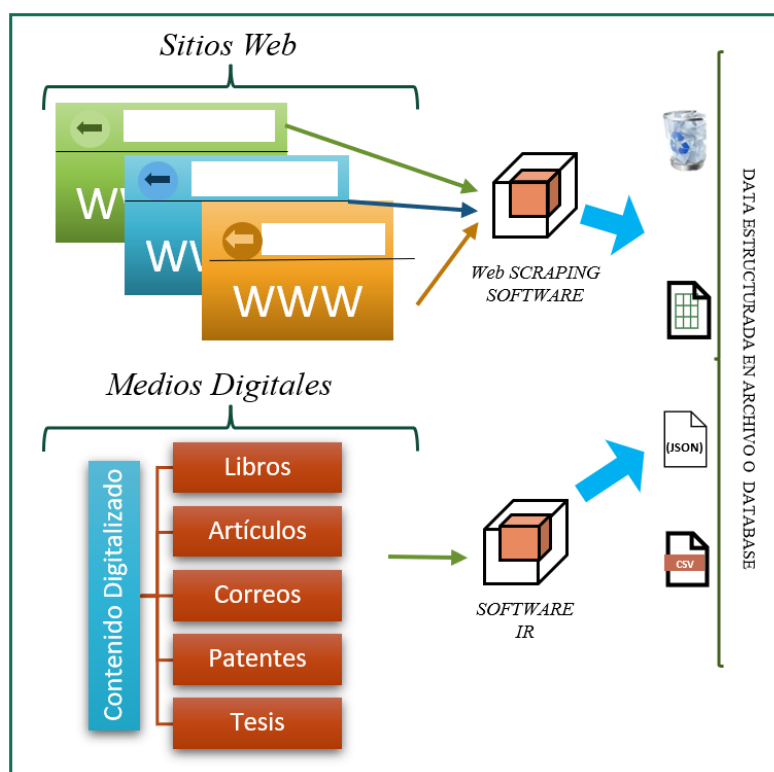


Figura 1.5. Representación de la extracción de Información Textual

1.3.2. DATOS TEXTUALES.

Luego de la obtención de la información que se va a analizar, mediante las operaciones del IR, el siguiente proceso consiste en la construcción de un corpus compuesto de datos textuales. En este punto cabe destacar que, en el campo de la lingüística, se denomina corpus al conjunto de documentos escritos (J. M. Sinclair, 1966). Así, el objetivo de la MDT, subyace en la conversión de un corpus en unidades léxicas tratables de documentos por palabra para procesar mediante técnicas estadísticas (Dumais et al., 1988).

En la identificación de estas unidades se requiere seguir un proceso, en el cual se identifiquen elementos básicos del corpus que permitan realizar el tratamiento estadístico e informático. Estos elementos básicos los podemos observar en la figura 1.6, precisando los siguientes:

- *Contexto*, formado por el corpus y los textos o documentos derivados del contexto de análisis (J. M. Sinclair, 1966, 1991). Partiendo de ello, se tiene: a) **Corpus**, el cual, proviene del latín, que significa cuerpo. Se define como un conjunto de textos o documentos que ha sido creado para ser sometido a un análisis lingüístico, por lo cual está estructurado y planificado con la idea de obtener conocimiento a partir del análisis; no hay un tamaño

óptimo del texto, el mismo viene derivado de la representatividad exigida y, b) **Documentos**, formado por un conjunto de párrafos que buscan expresar una o varias ideas, y forma parte del corpus.

– **Contenido Semántico**, formado por la agrupación de palabras en frases y párrafos. Permiten entender una idea dentro de un contexto de estudio, facilitando interpretar los enunciados de acuerdo con las contribuciones de las unidades lingüísticas (Dumais, 1992; Dumais et al., 1988; Leonetti & Escandell, 2004). Con base a esto, se tiene: a) **Párrafo**, conjunto de frases que terminan en un punto final, buscan expresar una o varias ideas al agrupar diferentes frases coherentes entre sí y b) **Frases**, es una secuencia de palabras, con un nexo semántico en un determinado texto.

– **Contenido morfológico**, este se refiere a los aspectos gramaticales y estructurales de las palabras en un texto. Es parte de la lingüística y pieza fundamental de la lingüística computacional en la MDT (Doyle, 1961; Luhn, 1958). Está formada por: a) **Palabras**, considerada la unidad textual primaria; permite añadir información morfológica y riqueza léxica al documento, se mantiene una dependencia al contexto de la idea principal y b) **Caracteres**, son unidades elementales, las cuales pueden ser numéricas, textuales o de símbolos que forman parte de cualquier documento; se consideran los pilares para la formación de las palabras.

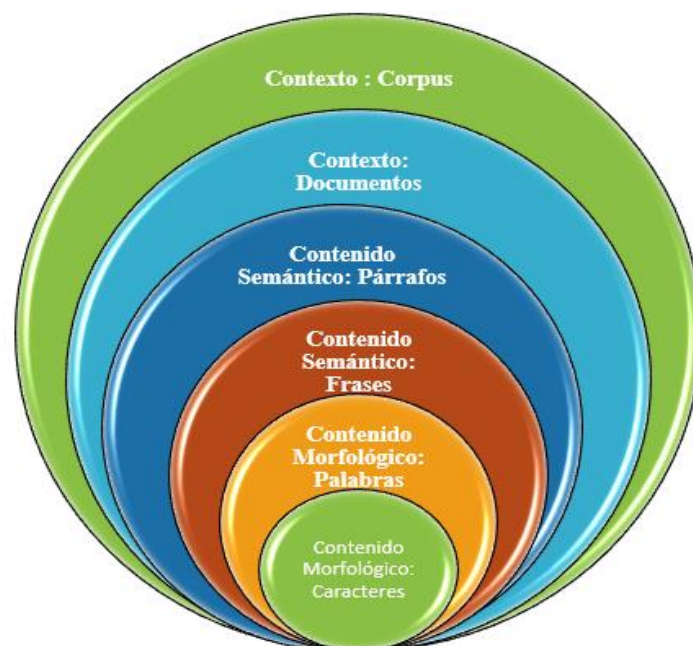


Figura 1.6. Contenido del Corpus textuales.

1.3.3. PREPROCESAMIENTO DE DATOS TEXTUALES

Una vez obtenidos datos estructurados de la IR y, habiendo identificado los componentes léxicos del corpus, estos deben ser sometidos a un procesamiento, con el fin de depurar la información hasta convertir los datos en una estructura denominada tabla léxica, que es la base del análisis textual desarrollado más adelante. Entre los tratamientos a considerar se tienen:

1.3.3.1. TRANSFORMACIÓN DE TEXTO.

En este punto, los corpus de textos son sometidos a una transformación donde se puede convertir mayúsculas, signos de puntuación, números, caracteres especiales; todo ello con la finalidad de permitir un correcto análisis de las formas lingüísticas y evitar que los algoritmos de procesamiento analicen incorrectamente la información (Weiss et al., 2005).

Ejemplo: Se desea reemplazar un carácter especial como @, en un espacio en blanco.

Este proceso de transformación es esencial para el presente proyecto, ya que nos va a permitir la reducción de dimensionalidad de los corpus, eliminando palabras vacías o sin contenido importante, así como la extracción de características relevantes de los textos, mediante procesos de tokenización y normalización.

1.3.3.2. TOKENIZACIÓN.

Luego de llevar los corpus a unidades léxicas tratables similares en forma, estos corpus deben de ser tokenizados, lo cual es asumido como un proceso de segmentación de un texto en unidades léxicas conocidas como “token”; es decir, dividir un texto legible por humanos, en un componente legible por una máquina. Esta separación va a depender del idioma en el que se encuentra el texto ya que, en ciertos idiomas, como el chino escrito, no hay marcas de separación entre palabras en una frase o texto; al contrario del español o el inglés, donde el espacio es el separador natural entre palabras (Gil Pascual, 2021).

Por Ejemplo: una frase podría ser “Buenas Noches España”; que estaría compuesto por los Tokens “Buenas”, “Noches”, “España”.

Existen múltiples formas de tokenizar, esto va a depender del tipo de estudio e investigación a realizar, ya que se puede realizar por espacios, o por símbolos, o se puede generar no solo por palabras, sino también por caracteres (Liu & Curran, 2006).

1.3.3.3. ELIMINAR PALABRAS VACÍAS.

Este proceso conocido en inglés como *Stop Word*, se refiere a eliminar aquellas unidades léxicas conocidas como auxiliares tales como las preposiciones, conjunciones, palabras modales y otras, que presenten una alta frecuencia en el corpus, pero que no aportan información relevante al análisis o estudio que los investigadores realizan.

Este tipo de palabras se conoce como vacías, por su nula o baja contribución al objeto de estudio. Además, suelen ocupar una gran cantidad de espacio de almacenamiento, por lo que es importante para la MDT su eliminación, con objeto de optimizar la eficiencia operativa del análisis. Para este proceso es necesario contar con un listado de palabras vacías, que se van a eliminar. Con el auge de los paquetes y algoritmos de análisis textual, se han creado diferentes listados de palabras vacías ya validados en diferentes idiomas (Raulji et al., 2016).

1.3.3.4. NORMALIZACIÓN DE LAS PALABRAS.

El siguiente proceso, denominado *normalización de las palabras*, consiste en fusionar las diferentes formas léxicas de una palabra en una forma única, que permita mejorar la eficiencia del procesamiento de texto. Así se minimiza significativamente el problema de la escasez de unidades léxicas frecuentes, causada por la representación de características del léxico en ciertas palabras de los textos las cuales, durante el procesamiento, pueden ser eliminadas al ser consideradas menos importantes por su limitada frecuencia (Salton & Buckley, 1988)

En este sentido, los sistemas computacionales aportan agilidad y rapidez en los procesos de normalización ya que, para estos, al igual que en la *eliminación de palabras*, se requiere de

un diccionario que contenga todas las posibles formas de normalizarlas en el idioma requerido. Este es el caso, por ejemplo, del diccionario freeling de la Universidad Politécnica de Cataluña (Hurtado et al., 2015), donde constan diferentes analizadores lingüísticos, con el fin de optimizar el proceso del análisis léxico y morfológico de las palabras.

El proceso de normalización comprende dos conceptos muy importantes como son la lematización y el Stemming:

a. Lematización.

Consiste en la restauración de las palabras arbitrariamente deformadas en sus formas originales. Tiene por objeto el de reducir el número de palabras índice que semánticamente sean parecidas (Zellig S, 1952). Este proceso lo hace buscando variaciones morfológicas de las palabras, extrayendo la raíz de los morfemas de estas y haciendo un análisis morfosintáctico de la misma (Zellig S, 1991). Es decir, se asigna en forma de una etiqueta, un *LEMA* (*forma de citación de una palabra*) a una palabra, tal y como se encuentra en un discurso textual (Martí & Llisterri, 2002).

Así, en cualquier idioma, el proceso de lematizar consiste en transformar:

- Los adjetivos al masculino singular
- Las formas verbales al infinitivo
- Los sustantivos al singular

Por tanto, cuando se lematiza un texto, se reemplaza cada palabra del texto por su lema; es decir, un texto lematizado, contendrá todas las formas verbales representadas por su infinitivo, todas las formas sustantivas representadas por su singular y los adjetivos en su forma masculino singular. Como se observa en la figura 1.7, las diferentes palabras que forman una conjugación de un verbo son sustituidas por su infinitivo. Para este proceso estas conjugaciones con sus verbos se encuentran en múltiples diccionarios, que con la ayuda de los sistemas computacionales permiten realizar este proceso de manera más eficiente (Hechavarría Díaz & Pérez Suárez, 2006).

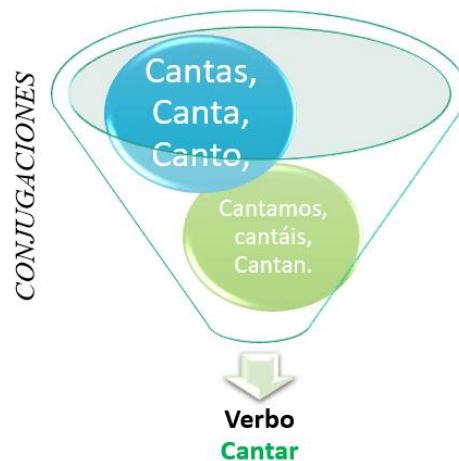


Figura 1.7. Ejemplo contenido diccionario para Lemmatización

Realizar este proceso de manera manual sería significativamente ineficiente, por lo cual, los diccionarios léxicos incorporados en los diferentes paquetes de múltiples softwares computacionales permiten su realización de forma más eficiente. Al respecto, en la figura 1.8 se observa una parte del listado en inglés generado por Mechura (Mechura, 2016) que forma parte del paquete “lexicón” escrito en lenguaje de programación R (Rinker, 2019).

Token	lemma
abbreviated	abbreviate
abbreviates	abbreviate
abbreviating	abbreviation
abcesses	abcess
abdicated	abdicate
abdicates	abdicate
abdicationing	abdicate
abdications	abdication
abdomens	abdomen
abdomina	abdomen

Figura 1.8. Sección del Diccionario para Lemmatización de Mechura (2016)

b. Stemming.

El stemming es el proceso de eliminar el afixo flexivo y derivativo para la obtención de la palabra raíz léxica o morfema conocida como *stem*, las cuales, no son necesariamente capaces de expresar una semántica completa (Bauer, 1983). Esto se hace con el objeto de analizar variaciones de una palabra como una sola.

Por ejemplo: Asumimos que se tiene un conjunto de palabras como: (moderniza, moderna, moderno, modernos, modernización). El proceso de stemming permite obtener el stem (modern, modern, modern, modern, modern) del conjunto dado de palabras.

En la actualidad, existen varios algoritmos que realizan esta operación en los diferentes programas de análisis de texto. En el caso de librería SnowballC de R, el cual presenta el algoritmo wordstem (Bouchet-Valat, 2020), que funciona con 7 conceptos básicos:

- Un identificador de regla, donde se pone cada lema y su stem
- El sufijo para identificar en el proceso de stemming, los paquetes y software usan mayoritariamente la técnica del algoritmo de Porter para la identificación del sufijo, con el fin de que la morfología de las palabras no penalice la frecuencia de estas (Porter, 1980).
- El texto por el cual debe ser reemplazado al encontrar el sufijo
- El tamaño del sufijo adecuado
- El tamaño del texto de reemplazo
- El tamaño mínimo que debe tener la raíz resultante luego de aplicar cada concepto, con el fin de no procesar palabras con pocos caracteres.
- Una función que verifica si se debe aplicar la regla una vez identificado el sufijo.

Durante el proceso de Stemming, los algoritmos generalmente presentan dos errores (Gil Pascual, 2021), precisados de la siguiente manera:

- *Overstemming*, o sobreescritura. Es un error en el que dos palabras flexionadas se derivan de la misma raíz, pero que no debió haber sucedido; es considerado un falso positivo.
- *Understemming*, o subestimación. En este error las palabras flexionadas separadas deben derivarse a la misma raíz, pero esto no sucede; es considerado un falso negativo.

De esta manera, los algoritmos actuales buscan minimizar la probabilidad de incurrir en estos errores, aunque resulta complicado reducir las dos condiciones a la vez durante los análisis. En la figura 1.9 se observa como un documento inicialmente compuesto por varias palabras, queda como resultante, luego de aplicar el stemming y la lematización



Figura 1.9. Palabras originales de un documento después de Stemming y Lematización

1.4. ESTRUCTURA DEL TEXTO

Para el análisis de los corpus de texto es importante considerar la estructura de las unidades léxicas, con objeto de poder realizar un análisis adecuado y efectivo. A continuación, se presentan aspectos a considerar en la MDT, dentro de la estructura léxica de los textos. Los mismos que aportan en la estructura de los textos a ser analizados bajo la propuesta del presente proyecto.

1.4.1. N-GRAMAS

Los n-gramas son secuencias de elementos tal cual aparecen en los documentos. Los n-gramas tradicionales son una subcadena de caracteres, etiquetas o cualquier otra medida que se encuentra una tras otra en los textos (Adamson & Boreham, 1974; Robertson & Willett, 1998). Estos representan información sintagmática y se utilizan ampliamente en procesos de lingüística computacional. Donde **n**, en el término n-grama, corresponde al número de elementos que se toman en una secuencia (Sidorov, 2019).

Por ejemplo, en la frase “La estadística es una materia vinculada a todas las ciencias”, se pueden formar Unigramas donde $n = 1$, como: “La”, “estadística”, “es”, “una”, “materia”, o bigramas con un $n = 2$ como: “La_estadística”, “estadística_es”, “una_materia”, “materia_vinculada”; o trigramas para un $n = 3$, como: “la_estadística_es”, “estadística_es_una”.

Así, la construcción de los n-gramas se basa en asignar una probabilidad a una secuencia de palabras en un contexto determinado. Sea $Pr(S^K)$ la probabilidad condicional de ocurrencia de una secuencia de K palabras (w_1, w_2, \dots, w_k) .

Para el cálculo de la probabilidad de $Pr(S^K)$ se puede usar aproximación de la probabilidad condicional $Pr(w_i/h_i)$. En este sentido, cuando hablamos de Unigramas, donde $n = 1$, se está hablando de palabras. Para la mayoría de las aplicaciones de minería de texto se consideran $n = 2$ o $n = 3$. Por tanto, tomando como base el modelo de Jelinek (Jelinek, 1976), la aproximación de la $Pr(S^K)$ sería:

$$Pr(w_i/h_i) = Pr(W_i/W_1, \dots, W_{i-1}) \approx Pr(W_i/W_{i-n+1}, \dots, W_{i-1}) \quad (1)$$

Para $n = 2$, donde se denomina bigramas (2-gramas), tendríamos la ecuación:

$$Pr(w_i/h_i) = Pr(W_i/W_1, \dots, W_{i-1}) \approx Pr(W_i/W_{i-1}) \quad (2)$$

Donde, $Pr(W_i/W_1, \dots, W_{i-1})$, es la probabilidad de que la palabra W_i aparezca después de las $i - 1$ palabras W_1, W_2, \dots, W_{i-1} . Y la secuencia de las palabras transmitidas previamente $h_i = W_1, W_2, \dots, W_{i-1}$ es llamado el histórico de la palabra.

Se puede inferir que, en la construcción de los n-gramas se considera el orden en que aparecen los términos o palabras en el texto. Ello no toma en cuenta la información sintáctica, por lo que, múltiples estudios han generado otros modelos de n-gramas, como el sn-gramas (n-gramas sintáctico), donde el número de palabras que se toman del texto se basa en relaciones sintácticas de estas y en donde cada palabra tomada está ligada a sus palabras vecinas reales, permitiendo así, considerar el significado sintáctico de la palabra (Sidorov et al., 2014).

Por ejemplo, en las frases de un documento “comer con cuchara plástica” y “comer con cuchara metálica” para obtener los bigramas con un $n = 2$, tenemos para un bigrama tradicional “comer_con”, para un bigrama sintáctico tenemos “comer_con”, “con_cuchara”.

En contra de lo mostrado en el ejemplo anterior, para construir los n-gramas o sn-gramas es recomendable no considerar las palabras auxiliares (stop word).

1.4.2. CONCEPTO

El término “concepto” denomina a las unidades más básicas de toda forma de conocimiento humano, representando la expresión de un pensamiento mediante una palabra (Putman, 1975). Es asumido como una idea abstracta o captación de la realidad, resultado de la interacción o experiencias con el entorno que se expresa finalmente con una única palabra o “unidad de conocimiento” (Díez & Moulines, 1997; Putman, 1975).

Por ejemplo, el concepto del término “Variable”, viene definido por la Real Academia Española, como: “magnitud que puede tener un valor cualquiera de los comprendidos en un conjunto”

En la MDT los conceptos se deben considerar de acuerdo con el idioma en el que se generan los análisis. Generalmente estos son avalados por la comunidad científica, reales academias de las lenguas e institutos lingüísticos. En la MDT son importantes, ya que nos permiten establecer o distinguir la polisemia y la semántica entre palabras a partir de los conceptos.

1.4.3. TESAURO

Un Tesauro es una lista de términos o palabras empleada para representar diferentes conceptos (Joyce & Needham, 1958). Se define como un vocabulario controlado y estructurado formalmente e integrado por términos que guardan relación semántica y genérica entre sí, como sinónimos. (Cavieres Abarca et al., 2010; Rada & Bicknell, 1989).

Así, para la construcción de Tesauros, será necesario conocer aspectos básicos que orienten y ayuden a su creación, como los siguientes:

- Una lista de términos preferentes, ordenados en forma alfabética, temática. equivalente y jerárquica.

- Una lista de sinónimos de los términos preferentes, conocidos como descriptores, cuando se presentan más de uno, uno de ellos se elige término preferente y se emplea siempre en la indización.
- La identificación de “términos generales” y “términos más específicos”, mediante una jerarquía o relación entre los términos.
- Determinar la ambigüedad entre términos, para facilitar la selección de estos.
- Definir las reglas para el uso del tesoro, de acuerdo con el objetivo de este.

Algunos de los tesauros prácticos que se encuentran en la literatura, de los varios que existen, puede ser:

- Tesoro de términos de indización de psicología (APA)
- Tesoro de términos de Indización de Sociología
- Tesoro de Arte y Arquitectura (ATT)
- Tesoro GeoRef
- Términos de Indización Legislativa (CRS)
- Tesoro de descriptores ERIC
- Tesoro de Indización de términos de Agricultura (AGROVOC)

Ahora bien, para el uso en sistemas computacionales, es necesario que estos tesauros sean automatizados y que, sin dejar de lado las relaciones analizadas entre términos, ayuden a que el usuario pueda crear y gestionar la terminología adecuada a los propósitos de su investigación concreta.

1.4.4. REDES SEMÁNTICAS

Las redes semánticas se asumen como un caso particular de una red jerárquica de conceptos o taxonomía de términos (Quillian, 1967). Es una forma de representación de conocimiento lingüístico en la que los conceptos y sus interrelaciones se representan mediante un grafo dirigido, aprovechando las relaciones de las palabras comunes para crear categorías de sinónimos o hipónimos (Collins & Loftus, 1975; Quillian, 1969).

En la figura 1.10, se representa un grafo donde los elementos semánticos se representan por nodos. Se pueden diferenciar dos tipos de nodos:

- *Token*: nodos de nivel inferior que heredan las características de los nodos de capa superiores y tienen características propias.
- *Type*: Representan a la clase de individuos, son nodos de nivel superior.

De esta manera, para cada nodo existen enlaces a otros nodos y, a su vez, están ligados a otros. Estos nodos están unidos mediante una línea, enlace o arista y para grafos dirigidos se requieren de flechas, los cuales representan las relaciones conceptuales entre términos (Collins & Loftus, 1975).

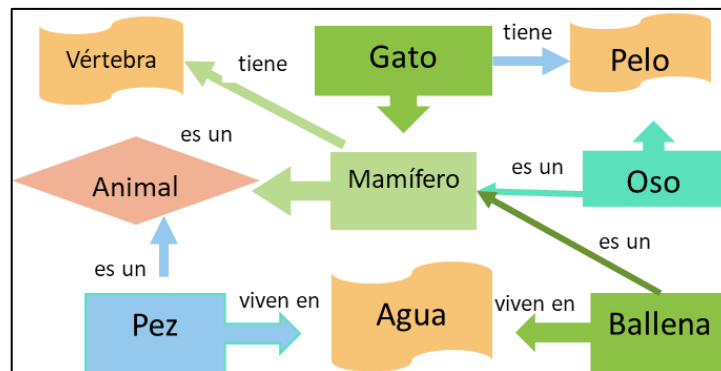


Figura 1.10. Ejemplo de una red semántica de palabras.

Por ejemplo, si mamífero es un concepto, gato y ballena serían hipónimos de mamífero, puesto que ambos son mamíferos.

En este orden de ideas, las redes semánticas inician identificando los sentidos posibles de cada concepto. Cuando estos se identifican como sinónimos o hipónimos, se agrupan en una sola categoría. Por ejemplo, la técnica de red semántica crearía una única categoría que contenga a pez, gato, oso, ballena, puestos que la red contiene la información de que estos son animales. Y estos a su vez se conectan con otros nodos que representan sinónimos o hipónimos.

1.5. REPRESENTACIÓN MATRICIAL DEL CORPUS

En este punto, convertir nuestros corpus en palabras (tokens), es el primer paso para posibilitar la realización de los análisis textuales. Este proceso de transformación es conocido como vectorización (Gil Pascual, 2021). Para su cumplimiento, primero, el texto debe ser preprocesado de acuerdo con la *eliminación de palabras vacías*, descrito en el punto 1.3.3, para luego ser estructurado en concordancia con los pasos del *proceso de la minería de datos textuales*, enfatizado en la sección 1.3, concluyendo en la elaboración de una tabla léxica.

Estas tablas, pueden ser construidas para representar frecuencias absolutas, pesos de las palabras por documentos, probabilidades de palabras por documentos, entre otras posibilidades (Anandarajan et al., 2019). Entre las diferentes tablas léxicas tenemos:

- *La Matriz Término Documento, Term Document Matrix (TDM)*. - La cual muestra la frecuencia absoluta de las palabras en los diferentes documentos que conforman el corpus.
- *La Matriz Documento Término, Document Term Matrix (DTM)*. - es la matriz traspuesta de una TDM, es decir en las filas se ubican los documentos del corpus y como variables los términos o tokens.

DTM		Término 1	Término 2	Término 3
	Documento 1			
	Documento 2			
	Documento 3			

TDM		Documento 1	Documento 2	Documento 3
	Término 1			
	Término 2			
	Término 3			

Figura 1.11. Tipos de Matrices de Documentos.

De acuerdo con estas tablas léxicas (figura 1.11), puede inferirse que el hecho común implica la representación de las frecuencias (Dumais, 1991). En la tabla 1.1 se observan 10 textos diferentes, que luego de ser preprocesados, se construye la matriz TDM mostrada en la tabla 1.2

Documentos

<i>Numero</i>	Texto	Preprocesamiento Texto
1	My favorite do gis fllyffy and tan	[favourite] [dog] [fluffy] [tan]
2	The do gis Brown and cat is Brown	[dog] [Brown] [cat] [Brown]
3	My favorite ha tus Brown and coat is Pink	[favorite]] [hat] [Brown] [coat] [Pink]
4	My dog has a hat and leash	[dog] [that] leash]
5	He has a fluffy coat and Brown coats	[Fluffy] [coat] [Brown] [coat]
6	The do gis Brown and fluffy and has a Brown coat	[dog] [Brown] [fluffy] [Brown] [coat]
7	My do gis White with Brown spots	[dog] [White] [Brown] [spot]
8	The White dog has a Pink coat and the Brown dog is fluffy	[White] [dog] [Pink] [coat] [Brown] [dog] [fluffy]
9	The three fluffy dogs and two Brown hats are my favorites	[fluffy] [dog] [Brown] [hat] [favorite]
10	My fluffy dog has a white coat and hat	[fluffy] [dog] [White] [coat] [hat]

Tabla 1.1. Preprocesamiento de texto de Documentos.

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Brown	0	2	1	0	1	2	1	1	1	0
Cat	0	1	0	0	0	0	0	0	0	0
Coat	0	0	1	0	2	1	0	1	0	1
Dog	1	1	0	1	0	1	1	2	1	1
Favorite	1	0	1	0	0	0	0	0	1	0
Fluffy	1	0	0	0	1	1	0	1	1	1
Hat	0	0	1	1	0	0	0	0	1	1
Leash	0	0	0	1	0	0	0	0	0	0
Pink	0	0	1	0	0	0	0	1	0	0
Spot	0	0	0	0	0	0	1	0	0	0
Tan	1	0	0	0	0	0	0	0	0	0
White	0	0	0	0	0	0	1	1	0	0

Tabla 1.2. Matriz TDM resultante de Tabla 1.1.

En este orden, mediante el uso de algunos conceptos de algebra matricial, donde una matriz es un arreglo bidimensional con m filas y n columnas, la matriz TDM contiene los datos con la estructura de la tabla 1.3, donde f_{ij} es la frecuencia de ocurrencia de la i -ésima palabra en el j -ésimo documento.

Palabra \ Documento	Doc 1	Doc2	Doc _j
P1				
P2				
.				
.				
.				
P _i				F _{ij}

Tabla 1.3. Estructura de Matriz TDM.

Este tipo de modelo matricial complica los estudios debido a la variabilidad de las dimensiones de las frecuencias en las matrices de obtenidas, por lo cual, en una manera de reducir las dimensionalidades es utilizada la ponderación de las palabras, profundizada a continuación.

1.5.1. PONDERACIÓN DE PALABRAS

Una vez generada la matriz de palabras, entre las formas de mitigar valores diversos de las frecuencias, se encuentra la ponderación de las palabras, la cual es consistente en asignar una función de peso, o ponderación, a la frecuencia absoluta, reemplazando en la matriz de partida la frecuencia por el peso (Dumais, 1992).

Así, para el cálculo del peso, se abordará el estudio de la ponderación local, ponderación global y ponderación combinatoria, descritas a continuación:

- *Ponderación Local*, es la asignación de un peso considerando un documento particular del corpus (Berry et al., 1995).

- *Ponderación Global*, es la importancia general del término en la colección completa del corpus (Berry et al., 1999a). Al aplicar la ponderación global, el resultado de los cálculos será un vector de valores que es la longitud del número total de términos.
- *Ponderación Combinatoria*, en una combinación de la ponderación local y global (Salton & Buckley, 1988).

De la misma manera, entre las alternativas de ponderación local, tenemos:

- *Frecuencia de Términos (tf)*, se lleva a cabo sobre la base de la probabilidad estadística, en donde se toma la frecuencia de palabras que transmiten un mayor significado en el corpus (Luhn, 2010).
- *Frecuencia logarítmica (lf)*, permite reducir el efecto de grandes diferencias en las frecuencias (Dumais, 1991). La frecuencia logarítmica $lf_{i,j}$ del término i en el documento j . Calculada como:

$$lf_{i,j} = \ln (tf_{i,j} + 1) \quad (3)$$

- *Frecuencia Binaria/Boleana (if)*, esta captura si una palabra aparece en un documento, sin tener en cuenta cuantas veces aparece (Salton & Yang, 1973). La frecuencia binaria $n_{i,j}$ del término i en el documento t :

$$n_{i,j} = \begin{cases} 1, & \text{if } tf_{i,j} > 0 \\ 0, & \text{de lo contrario} \end{cases} \quad (4)$$

Las ponderaciones globales más relevantes tenemos:

- *Frecuencia de Documentos (df)*, se puede obtener sumando las filas del TDM binario ponderado en frecuencia. Cuenta el número de documentos que contiene un término al menos una vez. Es un número entero entre 0 y D, número de documentos en el corpus (K. S. Jones, 1972). Se calcula como:

$$df_i = \sum_{j=1}^D n_{i,j} \quad (5)$$

Donde, $n_{i,j}$ es la matriz ponderada en frecuencia binaria y D es el número total de documentos.

- *Frecuencia Global (gf)*, mide la frecuencia de los términos en todos los documentos y es calculado mediante la ecuación:

$$gf_i = \sum_{j=1}^D tf_{i,j} \quad (6)$$

Donde $tf_{i,j}$ es la frecuencia del término i en el documento j y D es el número de documentos.

- *Frecuencia de Documentos Inversa (idf)*, en este caso las frecuencias de documentos se convierten en frecuencia de documentos inversa (K. S. Jones, 1972) , donde los términos raros tienen pesos más altos y los términos frecuentes tienen pesos más bajos (Dumais, 1991). Es calculado como:

$$idf_i = \log_2 \left(\frac{n}{df_i} \right) + 1 \quad (7)$$

Donde n es el número de documentos en el corpus, y df_i es el número de documentos donde aparece la palabra i -ésima.

En la actualidad, se precisan un grupo de ponderaciones que se usan frecuentemente, como son las ponderaciones combinatorias:

- *Frecuencia de Términos – Frecuencia de Documentos Inversa (tf-idf)*, esta ponderación cuantifica el grado de utilidad de los términos para caracterizar el documento en el que aparecen. Combina la ponderación local de la frecuencia de término y la ponderación global de la frecuencia del documento inversa al multiplicarlas (Salton, Yang, et al., 1975; Salton & Buckley, 1988; Spärck Jones, 1972). Se calcula mediante la formula:

$$tf - idf_{i,j} = tf_{i,j} \times idf_i \quad (8)$$

Así, cuando un término aparece muchas veces en unos pocos documentos, el TF-IDF es alto, y es bajo cuando un término aparece en todos, en la mayoría o en muchos documentos. Se podría indicar, que, si una palabra aparece con frecuencia en un documento o en la colección de documentos, tendría sentido considerar que el término es importante. Sin embargo, cuanto más frecuentemente aparece un término en los documentos, menos ayuda realmente a comprender el contenido textual (Jessup & Martin, 2005).

Cabe destacar que la elección del método de ponderación va a depender tanto de los datos como del modelado a generar y de los objetivos de investigación. Por ejemplo, para explorar el análisis semántico latente (LSA), descrito en el capítulo dos, se puede usar la ponderación tf-idf. Por otro lado, para el análisis de modelado de temas mediante la asignación latente de Dirichlet (LDA), a tratar en el capítulo dos, se requiere de una matriz TDM o DTM.

CAPÍTULO II

*TÉCNICAS PROBABILÍSTICAS DE LA
MINERÍA DE TEXTO*

2.1. INTRODUCCIÓN

El capítulo 1 se abordó el análisis de datos textuales, desde el preprocesamiento del texto hasta la generación de las matrices DTM o TDM. Ahora bien, desde este contexto, cabe analizar las tablas léxicas desde dos amplias perspectivas:

En el primer campo se encuentran los análisis de reducción de dimensionalidad, como el análisis de componentes principales (ACP) descrito por Pearson (Pearson, 1901) o el análisis Factorial de correspondencia (AFC) (Benzécri, 1973) y algunas técnicas derivadas de estas, permitiendo describir los datos en espacios vectoriales.

En el segundo campo de análisis se cuenta con las técnicas de análisis probabilísticos, como el análisis semántico latente probabilístico (PLSA) (Saul & Pereira, 1997b), que se deriva del análisis semántico latente (LSA) (Deerwester et al., 1990; T. K. Landauer & Dumais, 1997) y el modelado de tópicos, como el latente Dirichlet Allocation (LDA) (D. M. Blei et al., 2001) entre otras.

Las mismas perspectivas se enmarcan en las llamadas técnicas de “machine learning”, que sirven de base para algunas técnicas de la inteligencia artificial aplicada a la minería de datos textuales.

Atendiendo a estas premisas abordaremos en el presente capítulo, las técnicas del segundo campo amplio, es decir, las generadas a partir del análisis de probabilidad aplicadas a datos textuales.

2.2. ANALISIS SEMANTICO LATENTE

Una de las primeras técnicas usadas en la MDT es el análisis semántico latente (LSA). Esta se constituye en una herramienta matemática desarrollada inicialmente como un abordaje para la recuperación de información mediante la indexación semántica latente (Deerwester et al., 1990), cuyo objeto era el de indexar texto, a fin de solventar las grandes limitaciones que presentaban los motores de búsqueda en base de datos. Posteriormente, LSA ha sido

usado para la representación de nuevo conocimiento, a partir de frecuencia de palabras en los documentos.

Al respecto, puede inferirse la concepción de LSA como un modelo estadístico que permite comparar similitudes semánticas entre diferentes partes de los textos analizados (Foltz, 1996). En este orden de ideas, haciendo un ejercicio de adaptación de este concepto, desde la perspectiva de la psicolingüística, puede definirse como una teoría y método para la extracción y representación del significado contextual en el uso de las palabras, por medio de la computación estadística aplicada a un corpus textual (T. K. Landauer & Dumais, 1997).

De esta manera, LSA procesa los documentos del corpus lingüístico, dada una matriz, en la cual, las filas y columnas contienen los términos que aparecen en los documentos del corpus. En esta matriz de frecuencias se muestran en las filas las distintas palabras del corpus y, en las columnas, los diferentes documentos. Esta tabla léxica contiene el número de veces que cada palabra aparece en determinado texto.

Puede inferirse como un modelo de espacio vectorial, el cual permite suponer que ciertas palabras, aparentemente independientes, estén relacionadas por temas subyacentes no observados, efectuando una ponderación en la que se resta importancia a las palabras notablemente más frecuentes y aumenta la de las palabras moderadamente infrecuente. Esto es debido a que puede asumirse que las palabras excesivamente frecuentes no servirían para discriminar la información de cada documento (T. K. Landauer & Dumais, 1997).

Efectivamente, de manera específica, LSA, resulta en una aplicación de la descomposición de valores singulares (Singular Value Decomposition, SVD) (Berry et al., 1999b; Eckart & Young, 1936; Golub & Van Loan, 1996), que permite identificar el significado latente en los documentos a través de la reducción de dimensiones.

Así, LSA no sólo reduce la dimensionalidad, sino que también identifica las dimensiones latentes en función de valores singulares sin perder información relevante. Por tanto, el objeto de la SVD es ponderar cada término en función de su capacidad para representar un documento (Berry et al., 1999b).

2.3. DESCOMPOSICIÓN DE VALORES SINGULARES

En este apartado revisaremos el proceso para la SVD, dado que es usada en el LSA para obtener los valores singulares de la matriz termino- documento. Así como también, es la base de los métodos Biplot analizados en el capítulo 3. Propuesta inicialmente en 1936, la SVD consiste en descomponer una matriz rectangular X_k con dimensiones $n \times p$, de n palabras y p documentos, como el producto de tres matrices: a) Una matriz que contiene los vectores izquierdos singulares, b) Una matriz de vectores derechos singulares y, c) Una matriz con los valores singulares (Eckart & Young, 1936).

De esta manera, la SVD se expresa matemáticamente como:

$$X_k = U_k \Sigma_k V_k^T \quad (9)$$

Donde X_k es una matriz $n \times p$ de rango $r \leq \min(n, p)$, donde la dimensión de U_k es una matriz ortogonal $n \times k$ cuyas columnas son los vectores izquierdos singulares. Σ_k es una matriz diagonal $n \times p$ cuyos elementos diagonales son los valores singulares. Y V_k^T es una matriz traspuesta ortogonal $k \times p$ de la matriz de vectores derechos singulares. Así mismo, los valores singulares son números no negativos, es decir valores positivos o nulos en su diagonal.

Además, los términos o palabras vendrán representadas por las filas de la matriz $n \times k$:

$$U_K \cdot \Sigma_k \quad (10)$$

Mientras que los documentos lo estarán por las columnas de la matriz $k \times p$:

$$\Sigma_k \cdot V_k^T \quad (11)$$

Efectivamente, la matriz truncada creada, a través de SVD tiene cuatro propósitos importantes: a) Obtener el significado latente (es decir, Tópicos o Temas), b) Reducción de ruido, c) Obtener la coocurrencia de alto orden y, d) Reducción de escasez (Turney & Pantel,

2010). En la figura 2.1, se puede representar esquemáticamente la SVD en el LSA (Martin & Berry, 2007).

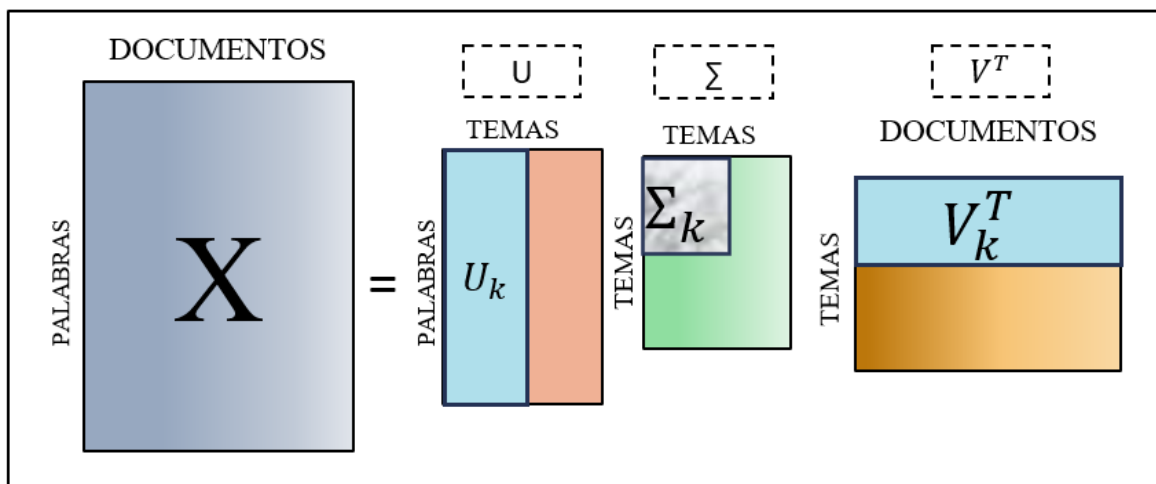


Figura 2.1. Proceso de la SVD, basado en Martin y Berry (2007).

Cabe destacar que, durante la SVD, se genera cierta pérdida de información, lo cual se considera como una reducción de ruido que provocan los factores que no eran relevante en las relaciones entre términos y documentos, y que han sido eliminados. Después del proceso de la SVD se obtiene una información que no está directamente disponible en la matriz original, sino que está latente en ella (T. K. Landauer et al., 2009).

Estas matrices pueden ser representadas en un espacio vectorial semántico, como también por las técnicas Biplot, este proceso será analizado en el capítulo 3. Así, de las matrices reducidas de vectores singulares obtenidas se considera a las filas de estas como coordenadas de los puntos que representan a los documentos y términos o palabras, en un espacio r -dimensional cuyos ejes están reescalados por cantidades relacionadas con los valores de la matriz diagonal.

Una de las técnicas de representar estas matrices es la visualización por medio del modelo de espacio vectorial. De acuerdo con Berry (1999), a partir del producto de los puntos obtenido de las ecuaciones (10) y (11), las palabras y documentos pueden ser comparados de la misma manera en que se comparan vectores: los productos escalares (coseno del ángulo entre vectores) entre los puntos nos darán las relaciones de similitud entre los diversos puntos (Berry et al., 1999b). Diferenciando de las representaciones Biplot por su medida de distancia.

2.3.1. ESPACIO SEMÁNTICO

De las matrices obtenidas en la SVD se determina que los resultados se pueden representar en un espacio semántico. Esta representación semántica de los textos se ha generalizado y se puede obtener por medio de la semántica distribucional o de la semántica composicional (Mitchell & Lapata, 2010). Desde esta concepción, la semántica composicional puede fundamentarse como una técnica de representación de los datos textuales que modela el significado semántico de las palabras utilizando el espacio vectorial. La semántica composicional permite enfocarse en modelar el significado no sólo de las palabras, sino de frases y oraciones.

Desde este ámbito, en la minería de datos textuales (MDT) y en el procesamiento del lenguaje natural (PLN) puede inferirse que uno de los modelos más usados para la representación gráfica de datos textuales es el espacio vectorial semántico (VSM). El VSM permite representar los documentos a través de vectores de términos en un espacio de alta dimensionalidad (Salton, Wong, et al., 1975), asumiendo que cada palabra se representa como un vector en un espacio n-dimensional, donde n es el número de términos o características que se consideran importantes para la representación del lenguaje. Así, los documentos se identifican como un vector de rasgos en un espacio en el cual cada dimensión corresponde a términos indexados distintos (palabras) (Arco, 2008).

De la misma manera, otra técnica de uso frecuente es la representación en un espacio semántico euclídeo, donde figuran los resultados obtenidos en la SVD del LSA. En este tipo de espacios cada vector representa el significado de palabras o documentos, considerando la información léxica (Kintsch, 2002), donde las palabras que tienen un significado similar estarán más cercanas entre sí en el espacio semántico latente euclídeo. Por ejemplo, al analizar las palabras “perro” y “gato”, estas estarán más cercanas entre sí, que las palabras “perro” y “coche” (Clark, 2014).

Ahora bien, en los espacios semánticos, de cara al establecimiento de la similitud del significado de los términos, se utiliza la longitud del vector y el coseno del ángulo entre vectores (T. K. Landauer & Dumais, 1997). Por ejemplo, en el VSM se utiliza la similitud coseno para establecer la distancia entre dos vectores, a diferencia del euclídeo, que mide la distancia euclídea, donde:

- *La longitud del vector*, se obtiene al sumar todos los componentes de un vector, permitiendo medir la cantidad de información semántica que el LSA tiene respecto de un dominio de conocimiento (Berry et al., 1995).

- *La similitud semántica*, representada por el coseno del ángulo entre vectores (Kintsch, 2002), considera que cuanto más próximos están dos vectores, mayor es la similitud entre ellos. Cuando el ángulo es menor, la similitud es mayor y en consecuencia el coseno del ángulo es mayor (Seijo et al., 2011).

Ambas propiedades, así como la SVD también se manifiestan en los modelos Biplot, ya que estos son la base de dichas representaciones, pero con ciertas diferencias que se revisaran en el próximo capítulo.

Cabe destacar que existen algunas medidas de similitud en el estudio de la minería de texto, adicional al coeficiente de similitud coseno. Así, entre los más usados tenemos: a) La medida de similitud de Jaccard (Jaccard, 1912), b) Medida de Sorensen (Sorensen, 1948), entre otras. Así mismo, se considera la propuesta de la similitud de distribución de tópicos (Steyvers & Griffiths, 2007). Sobre esta última, trataremos en el capítulo 6.

2.4. MODELADO DE TÓPICOS

En la actualidad, la significativa cantidad existente de información textual, ubicable entre diversidad de repositorios digitales y entornos web de la WWW, aunado al avance computacional, ha permitido que se generen nuevas técnicas estadísticas para la MDT.

En este sentido, se dispone de metodologías que permiten poder realizar resúmenes automáticos de documentos, mediante la generación del descubrimiento de tópicos o temas relevantes en estos. Este proceso se conoce como modelado de tópicos, o por sus siglas en inglés TM (*topic model*). El TM se constituye como una técnica del aprendizaje no supervisado no jerárquico, utilizada en el ADT. Se enfoca en el descubrimiento de tópicos a partir de patrones y relaciones más relevantes dentro de un conjunto de documentos y asignar cada elemento de esta agrupación a uno, o a varios, de los tópicos o temas principales (Asmussen & Møller, 2019a).

El TM es también conocido como modelado de tópicos probabilístico (TMP), el cual permite extraer y descubrir características y estructuras comunes (Tópicos) dentro de los datos textuales. Ello, gracias al análisis de las palabras de los corpus originales para explorar los temas latentes u ocultos que los contiene, y como se relacionan entre sí. (Roberts et al., 2014).

En este orden de ideas, los tópicos se pueden mostrar como una distribución de probabilidad sobre los términos (Tokens o palabras) (Yi & Allan, 2009); donde un documento del corpus puede estar asociado a un solo tópico (membresía única), o este puede ser una mezcla de múltiples tópicos (membresía mixta). En la figura 2.2, se muestra una representación del modelado de temas descrito por Steyvers (Steyvers & Griffiths, 2007).

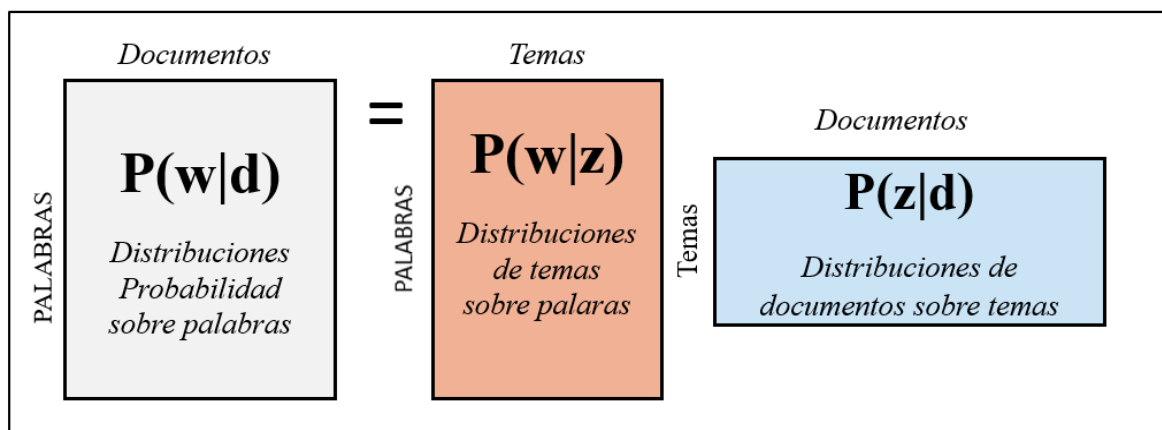


Figura 2.2. Representación de Modelado de Tópicos, basado en Steyvers & Griffiths (2007).

Así, las matrices TDM o DTM, se dividen en dos distribuciones principales cuando se analizan con técnicas de TM, que son: a) La distribución de tópicos sobre términos $P(w_j|z_k)$, que nos indica la importancia del término en cada tópico y, b) La de documentos sobre tópico $P(z_k|d_i)$, que nos determina la importancia de estos en los documentos (D. M. Blei et al., 2002).

Al respecto, entre algunos de los algoritmos usados para el modelado de tópicos, tenemos:

El algoritmo para análisis probabilístico de la semántica latente (PLSA)(Hofmann, 1999) analizado en el apartado 2.4.1. También el algoritmo de Factorización No-Negativa (NMF) en la cual permite descomponer una matriz de frecuencia de palabras, en dos matrices de

menor rango, las cuales representan los tópicos y las palabras asociadas a estos (Paatero & Tapper, 1994).

Así mismo, encontramos el algoritmo de Asignación latente de Dirichlet (Latente Dirichlet Allocation, LDA) usado para identificar los temas latentes en un corpus de texto, donde se asume que cada documento del corpus es una mezcla de varios tópicos (D. M. Blei et al., 2002), este método se analiza en el punto 2.4.2, del presente trabajo. Entre otros algoritmos que parten de la propuesta de Blei tenemos, a) el Modelado de tópicos correlacionados (CTM) que permite identificar los tópicos que se superponen o que se encuentran relacionados de alguna manera (D. M. Blei & Lafferty, 2005), b) el algoritmo del proceso jerárquico de Dirichlet (Hierarchical Dirichlet Process, HDP) aplicado al estudio de grupos de datos, donde cada grupo se modela de un proceso de Dirichlet de nivel dos y todos los grupos comparten la distribución base extraída de un proceso Dirichlet nivel uno (Teh et al., 2006).

Cabe destacar que, en la presente intención investigativa, abordamos el modelo de membresía mixta LDA, donde el número de tópicos, denominado K , es un número fijo que se asigna antes de construir el modelo.

2.4.1. ANÁLISIS PROBABILÍSTICO SEMÁNTICA LATENTE

El análisis probabilístico de la semántica latente (Probabilistic Latent Semantic Analysis, PLSA) es propuesto en 1999 por Hofmann como la versión probabilística del LSA, donde Hofmann declara que el LSA posee ciertas limitaciones estadísticas para encontrar los temas latentes (Hofmann, 1999). Al respecto, este se desarrolló a partir de conceptos del álgebra lineal, propuesto inicialmente como modelo de aspecto por Saul y Pereira en 1997 y que, para efectos del ámbito del modelado del lenguaje, se le nombra modelo agregado de Márkov (Saul & Pereira, 1997a).

De esta manera, PLSA se utiliza para encontrar la estructura latente de un conjunto de documentos del corpus, es decir, identifica temas o tópicos que subyacen en estos y, así, descubrir la semántica de tópicos ocultos en el corpus (Ren & Han, 2014). En este sentido, busca identificar y encontrar las probabilidades de ocurrencia de cada tópico en cada uno de

los documentos; así como las probabilidades de ocurrencia de cada palabra, en cada uno de los tópicos, construyendo una matriz de tópicos-documentos y una matriz de tópicos-palabras.

En este modelo de datos coocurrentes es asociado un factor latente, o variable no observable, $Z_k \in (z_1, z_2 \dots z_k)$ a cada observación; siendo esta observación la ocurrencia de un término en un documento. Por lo tanto, según Hofmann, para representar el algoritmo que se utiliza en la estimación de la distribución de probabilidad condicional en el PLSA, consideraremos:

- $P(d_i)$, que representa la probabilidad que será observada de que un término aparezca en un documento d_i .
- $P(w_j|z_k)$, que es la probabilidad de un término w_j condicionada a la variable Z_k
- $P(z_k|d_i)$, es la distribución de probabilidad de un documento d_i en el espacio latente de las variables $z_1, z_2 \dots z_k$.

Con base en estos conceptos, se puede generar las coocurrencias de términos-documentos, para lo cual el PLSA, se puede expresarse basado en el siguiente esquema:

- Primero, seleccionamos un documento d_i con probabilidad $P(d_i)$,
- Luego, tomamos un factor oculto Z_k con probabilidad $P(z_k|d_i)$,
- Finalmente, generar un término w_j con probabilidad $P(w_j|z_k)$.

De esta manera, como resultado, se obtiene una probabilidad ocurrente de un par observado (d_i, w_j) que adopta al factor denominado variable Z_k , quedando está descartada.

Este esquema se expresa en el siguiente modelo de probabilidad:

$$P(d_i, w_j) = P(d_i) P(w_j|d_i), \text{ donde } P(w_j|d_i) = \sum_{k=1}^K P(w_j|z_k) P(z_k|d_i) \quad (12)$$

Aplicando la fórmula Bayesiana en el modelo, obtenemos la ecuación (13), que es una versión parametrizada equivalente, en el que la probabilidad conjunta resulta perfectamente simétrica para documento y términos.

$$P(d_i|w_j) = \sum_{k=1}^K P(z_k)P(d_i|z_k) P(w_j|z_k) \quad (13)$$

Cabe destacar que, para maximizar la probabilidad total, es necesario generar repetidamente las probabilidades condicionales de $P(z_k)$, $P(d_i|z_k)$, $P(w_j|z_k)$, usando los datos de partida.

Así, el ajuste de las dos distribuciones de probabilidad generadas en el PLSA se realiza mediante el algoritmo de Expectación-Maximización (EM). Este algoritmo permite encontrar estimadores de máxima verosimilitud (Dempster et al., 1977) y se compone de dos pasos iterativos.

- Paso de Expectación (E), donde se calculan las probabilidades a posteriori a partir de los factores latentes, basado en las estimaciones actuales de la probabilidad condicional.
- Paso de Maximización (M), donde las probabilidades condicionales estimadas se actualizan, e intentan maximizar la probabilidad basada en las probabilidades a posteriori computadas en el paso de expectación.

De esta manera, la implementación iterativa del paso E y M se repite hasta que se converge hacia un límite de locales óptimos, es decir, los resultados calculados representan las estimaciones de la probabilidad óptima de los datos observados. En la figura 2.3 se representa el modelo PLSA cuando es simétrico; donde, dado un documento d_i , se encuentra un tema o tópico z_k presente en el documento, con una probabilidad $P(z_k|d_i)$; y dado el tema z_k se extrae la palabra o término w_j de este tema con una probabilidad $P(w_j|z_k)$.

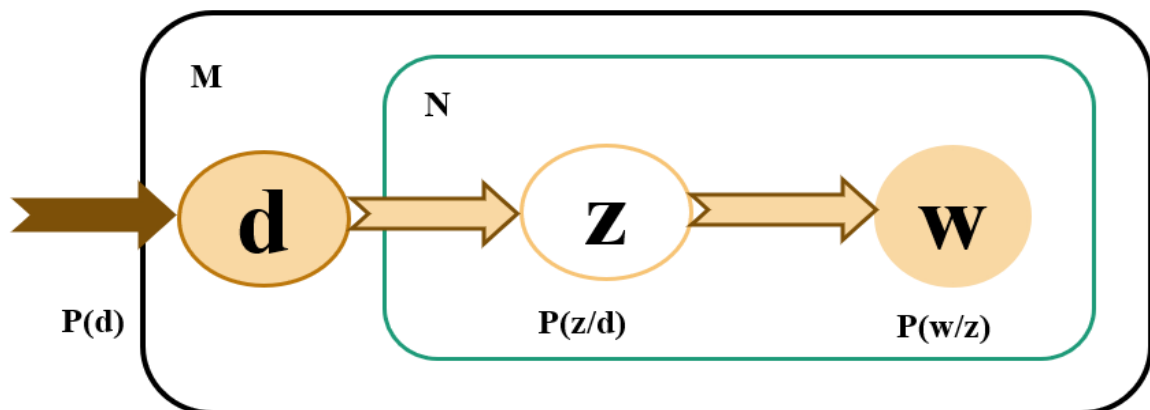


Figura 2.3. Representación de Placas del PLSA simétrico, Basado en Hofmann (1999).

Por tanto, relacionando el LSA con el modelo probabilístico PLSA, se puede reescribir el modelo probabilístico de manera matricial. Obteniendo las tres matrices:

- $U = (P(d_i|z_k))_{i,k}$
- $V = (P(w_j|z_k))_{j,k}$
- $\sigma = \text{diag} (P(z_k))_k$

Donde, la probabilidad conjunta P del PLSA, se expresa como el producto de las tres matrices, expresada como:

$$P = U \sigma V^t \quad (14)$$

Hay que destacar que, el algoritmo PLSA, se ha usado tanto en múltiples estudios textuales, como en la identificación de secuencia de sentimientos en correos electrónicos (Srinivasarao & Sharaff, 2022), en la investigación de traducción al inglés (Shen & Guo, 2022), en la predicción económica a partir de actas del comité federal de mercado abierto de Estados Unidos (Huang & Kuan, 2021) y en el análisis de habilidades laborales (Ao et al., 2023), entre otros.

En virtud de ello, este algoritmo presenta ventajas significativas, producto de su condición para presentar un claro significado probabilístico y las direcciones del espacio latente, que son interpretables como distribuciones condicionales de términos que definen un contexto temático. Ahora bien, resulta imperativo señalar que requiere una mayor complejidad computacional.

2.4.2. ASIGNACIÓN LATENTE DE DIRICHLET

La Asignación Latente de Dirichlet (Latent Dirichlet Allocation, LDA), propuesto por Blei (D. M. Blei et al., 2003) es una técnica estadística que se utiliza para identificar los temas o tópicos latentes en un conjunto de documentos. LDA permite asumir que cada documento es una mezcla de temas latentes y que cada tema o tópico es una distribución de probabilidad sobre un conjunto de palabras; admitiendo determinar la mezcla de temas latentes y las palabras asociadas en un conjunto dado de documentos del corpus.

Esto permite asumir que el objeto del LDA es inferir o estimar las variables latentes mediante un modelo probabilístico generativo no supervisado de un corpus, donde los documentos se representan como mezclas aleatorias sobre tópicos latentes y donde cada tema se caracteriza por una distribución condicionada entre palabras (D. M. Blei et al., 2003).

LDA reduce considerablemente el número de parámetros a considerar y permite de forma clara el cálculo de la probabilidad para documentos arbitrarios. Al respecto, en el LDA, las unidades básicas de los datos son las palabras, las cuales, quedan representadas mediante vectores unitarios de dimensión V con una única componente igual a 1 y el resto de componentes tendrá valores iguales a cero. Así, los documentos se representan por la secuencia de palabras que lo componen, expresado por un vector $d = (w_1, w_2 \dots w_N)$, y por el corpus $D = (d_1, d_2 \dots d_M)$. De esta manera, la técnica asume que los documentos son intercambiables, lo que significa que no hay un orden secuencial significativo de los documentos en la colección (D. Blei et al., 2010).

Cabe destacar que la principal diferencia con el PLSA es que esta utiliza la descomposición matricial para modelar la distribución conjunta de palabras y tópicos, mientras el LDA utiliza una distribución de Dirichlet para modelar la distribución de tópicos en los documentos y la de palabras en los tópicos (D. M. Blei & Lafferty, 2009).

2.4.2.1. INTERPRETACIÓN GEOMETRICA DEL LDA

Para entender el LDA imaginemos un enfoque geométrico, con una colección de documentos $D = (d_1, d_2 \dots d_M)$, donde estos, por un lado, se encuentran asociados con uno o varios temas latentes como: Política, Social y Económica y, por otro un triángulo donde las esquinas o vértices son los temas latentes.

Así, el algoritmo LDA colocará los documentos dentro del triángulo, de tal manera, que los documentos se agrupen cerca de la esquina que corresponde al tema y los documentos que tengan varios temas, se distribuirán en el centro del triángulo, como se observa en la figura

2.4

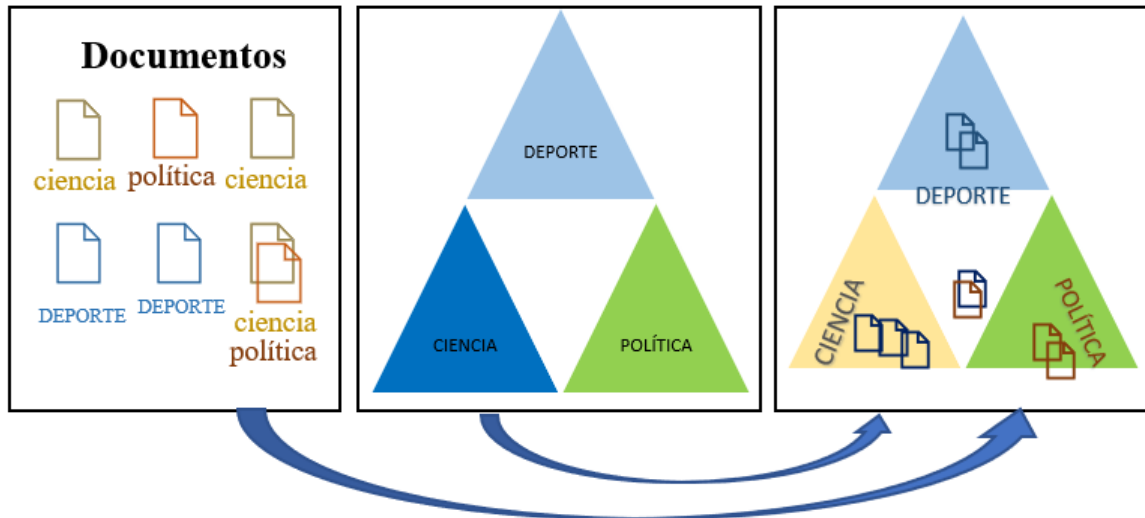


Figura 2.4. Representación Geométrica del LDA

A esta representación geométrica de la distribución de los documentos en el interior del triángulo se la conoce como distribución de Dirichlet (Good, 1965; Lindley, 1964). Se denota por $Dir(\alpha)$, la cual, es parametrizada por un vector alfa perteneciente al conjunto de los números reales positivos. La misma que se expresa como:

$$P(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \theta_1^{\alpha_1-1} \dots \theta_k^{\alpha_k-1} \quad (15)$$

Donde, una variable aleatoria K-dimensional de Dirichlet θ puede tomar valores en el $(k-1)$ -simplex, esto es, $\theta_i \geq 0$ y $\sum_{i=1}^k \theta_i = 1$, y, donde el parámetro de concentración α es un vector de dimensión k con componentes $\alpha_i \geq 0$, y donde $\Gamma(\alpha)$ es la función Gamma.

2.4.2.2. MODELO GENERATIVO LDA

El LDA, asume que cada Documento d del corpus D se genera siguiendo un modelo generativo, basado en reglas de muestreo probabilístico (D. M. Blei, 2012). El objeto de seleccionar un modelo generativo es en encontrar el mejor conjunto de variables latentes que pueden explicar los datos originales observados (Steyvers & Griffiths, 2007).

En la Figura 2.5 se observa el enfoque de TM basado en un modelo generativo, donde este es ilustrado con dos tópicos. Cada uno con un conjunto de palabras que contiene diferentes

distribuciones de probabilidad sobre las palabras. De estos, se generan tres documentos. Los documentos 1 y 3 fueron generados de los tópicos 1 y 2 respectivamente, mientras, el documento 2 es una mezcla de ambos.

Se infiere que el modelo define la no existencia de la noción de exclusividad mutua que restringe las palabras, para ser parte de un tema solamente, capturando así la polisemia, que la misma palabra pueda tener diferentes sentidos.

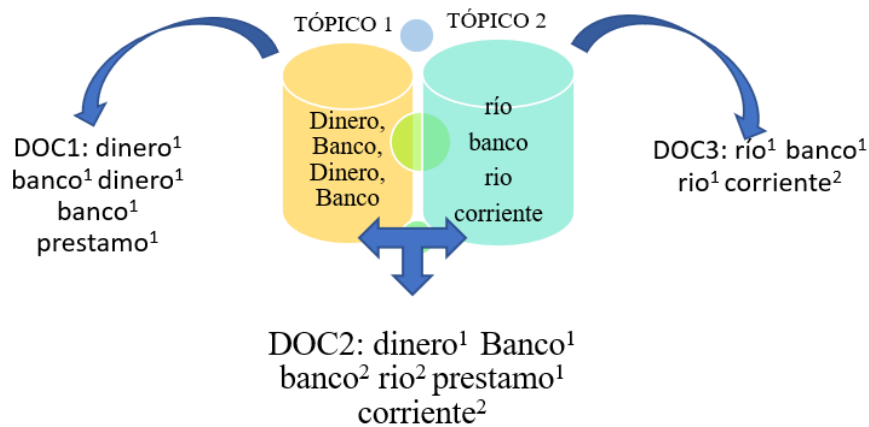


Figura 2.5. Enfoque del proceso generativo del modelado de tópicos (Steyvers & Griffiths, 2007)

Así, la LDA, considerando las premisas expuestas por Blei, puede asumirse como un modelo jerárquico de tres niveles, de los cuales, cada nivel subyacente es tomado como un conjunto finito de su nivel inferior respectivo (D. M. Blei et al., 2003). Estos niveles son:

- *Palabra*, que es la unidad básica discreta de los datos textuales, se define como un elemento del vocabulario $\{1, \dots, V\}$
- *Documento*, es la secuencia de N palabras indicado como $m = (w_1, w_2 \dots w_N)$
- *Corpus*, definido como la colección de M Documentos denotado por $D = (m_1, m_2 \dots m_M)$

Así, el proceso generativo del LDA se plantea en los siguientes pasos:

1. Para cada tópico:

a. Elija $\varphi_k \sim Dir(\beta)$, donde φ_k es el parámetro de la distribución de términos dado el tema k -ésimo.

2. Para cada Documento:

Elija $\theta_m \sim Dir(\alpha)$, donde θ_m es el parámetro de la distribución de tópicos en el m -ésimo documento.

Para cada posición de palabra en cada documento:

- i. Elija un tópico para la posición actual: $Z_{m,n} = Cat(\theta_m)$.
- ii. Con Base en el tópico $Z_{m,n}$, elija un término para la posición $w_{m,n} = Cat(\varphi_{Z_{m,n}})$.

Así, dada la ecuación;

$$\theta_m = [p(Z_1|m_M), p(Z_2|m_M), \dots, p(Z_K|m_M)] \quad (16)$$

Que denota la distribución condicional del tópico en el m -ésimo documento, y dada la ecuación;

$$\varphi_k = [p(w_1|Z_k), p(w_2|Z_k), \dots, p(w_V|Z_k)] \quad (17)$$

La cual, representa a la distribución condicional de términos en el K -ésimo tópicos.

De esta manera, los parámetros θ_m y φ_k se consideran variables aleatorias y se extraen de dos distribuciones de Dirichlet, en lugar de tratarse como variables determinísticas como se evidencia en el PLSA. En este orden, los parámetros en el LDA no crecen con el tamaño del corpus y es capaz de generalizar fácilmente a los nuevos documentos. Por ello, las probabilidades de las palabras están parametrizadas por una matriz β de dimensión $k \times V$.

Este TM se puede representar gráficamente como una notación de placas, como la observada en la figura 2.6, donde cada nodo es una variable aleatoria y, los nodos ocultos proporciones de los tópicos.

Asignaciones y tópicos no se encuentran sombreados. Los nodos observados y las palabras de los documentos están sombreados. Se usan rectángulos, como notación de placa para indicar la replicación del modelo donde, por un lado, la placa N denota la colección de palabras dentro de los documentos y, por el otro, la placa M denota la colección de documentos en el Corpus. Así mismo, la placa K representa los tópicos del modelo (D. M. Blei & Lafferty, 2009).

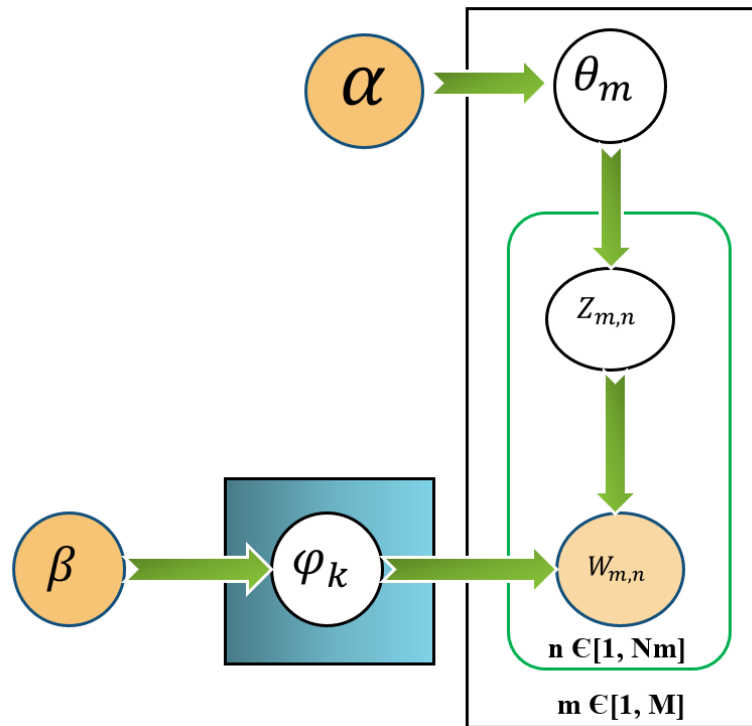


Figura 2.6. Representación de Placas de LDA, basado en Blei & Lafferty, 2009

Considerando lo descrito, y para una representación más formal, se puede deducir la siguiente distribución conjunta de las variables ocultas y observada, expresada como:

$$P(w, z, \theta, \varphi | \alpha, \beta) = P(\theta_m | \alpha) P(z_m | \theta_m) P(\varphi_k | \beta) P(w_k | z_k, \varphi_k) \quad (18)$$

Donde,

- α , es un hiperparámetro previo de la distribución de Dirichlet con respecto a θ_m
- β , es un hiperparámetro previo de la distribución de Dirichlet con respecto a φ_k
- N_m , es la longitud de m -ésimo documento

- V , es el número de términos o palabras
- $Z_{m,n}$, que representa al tópico asignado a la n -ésima palabra posicionada en el m -ésimo documento
- $w_{m,n}$, representa al término asignado a la n -ésima palabra posicionada en el m -ésimo documento

De la ecuación (18), tenemos:

- $P(\theta_m | \alpha)$, que representa la distribución de tópicos por documento y se expresa en la ecuación (19) a partir del parámetro Dirichlet, explicado en la ecuación (15). Este es establecido como un vector K -dimensional con componentes $\alpha_k > 0$ y considerando el punto 2.a del modelo generativo, se tiene:

$$P(\theta_m | \alpha) = \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \theta_m^{\alpha_k - 1} \quad (19)$$

- $P(z_m | \theta_m)$, tomando el punto 2.b.i del modelo generativo, donde el tópico de la n -ésima palabra posicionada en el m -ésimo documento $Z_{m,n}$; se expresa como la probabilidad del tópico z para todos los documentos y, también, de los tópicos en términos del número de palabras $n_{m,k}$; donde, en la ecuación (20), se expresan los múltiples ensayos, a la cual se ha asignado un tema K , a cualquier palabra, que se encuentra en el documento m .

$$P(z_m | \theta_m) = \prod_{n=1}^{N_m} P(z_{m,n} | \theta_m) = \prod_{k=1}^K \theta_{m,k}^{n_{m,k}} \quad (20)$$

- $P(\varphi_k | \beta)$, representa a las distribuciones de los tópicos según la palabra. Teniendo en cuenta todo el corpus φ_k , se obtiene de una distribución de Dirichlet, con parámetro β , de acuerdo con el punto 1.a del modelo generativo. Así, la probabilidad proporcionada por $\varphi_{k,N}$, donde la palabra N se adquiere de acuerdo con el tópico seleccionado, expresado como:

$$P(\varphi_k | \beta) = \frac{\Gamma(\sum_{n=1}^N \beta_{k,n})}{\prod_{n=1}^N \Gamma(\beta_{k,n})} \prod_{n=1}^N \varphi_{k,n}^{\beta_{k,n} - 1} \quad (21)$$

- $P(w_k | z_k, \varphi_k)$, representa la probabilidad de un corpus w_k , denotado por la ecuación (22), donde w_k es una secuencia de términos en el que el tópicos en cada posición es K en todo el corpus y, z_k , es la secuencia correspondiente de tópicos donde cada elemento es el tema K , dado el parámetro φ_k .

$$P(w_k | z_k, \varphi_k) = \prod_{v=1}^V \varphi_{k,v}^{n_{k,v}} \quad (22)$$

Todos los procesos explicados en el modelo generativo del LDA presentan problemas para ser tratados computacionalmente, ya que se requiere de procesos de inferencia, por lo cual, se debe hacer uso de algoritmos de aprendizaje automático a fin de obtener una aproximación cercana al verdadero valor de la probabilidad a posteriori.

2.4.2.3. INFERENCIA EN LDA

Este fundamento está referido al proceso de inferir los valores de las variables ocultas o latentes, en función de las variables observadas. Hay que destacar que, en el proceso del LDA, se debe inferir la distribución de temas latentes dadas las palabras observadas $P(Z_{m,n} | w_{m,n})$, según los procesos de inferencia bayesiana y estimar la distribución a posteriori de θ_m y φ_k (Steyvers & Griffiths, 2007).

En este sentido, el LDA usa algunas técnicas de aprendizaje automático (machine learning), para tratar lo complicado de generar estas inferencias y obtener una inferencia aproximada.

Al respecto, Blei, en su artículo publicado en el 2003, utilizó el algoritmo de Maximización de la Expectativa Bayesiana Variacional (VBEM, por sus siglas en inglés). Así, entre los diferentes algoritmos para la inferencia en el LDA, se tienen:

- *Maximización de la Expectativa Bayesiana Variacional (VBEM)* (Attias, 1999): en el LDA se utiliza para inferir los valores de los parámetros de las distribuciones de tópicos y de palabras en cada documento, a partir de los datos de partida observados. VBEM, iterativamente, actualiza valores de los parámetros a través de dos pasos: *el de expectación*, por un lado, que estima los valores esperados de las distribuciones latentes y, por el otro, *el*

paso de maximización, que maximiza la función de verosimilitud con respecto a los parámetros (D. M. Blei et al., 2003).

- *Propagación de Expectativas (EP)* (Minka, 2001), genera un refinamiento iterativo de las aproximaciones más precisas; utiliza una serie de aproximaciones locales para propagar la información entre los diferentes factores del modelo LDA y, busca minimizar la divergencia entre la distribución a posteriori y la factorizada (Sumba & Bouguila, 2020).

- *Monte Carlo vía Cadena de Márkov (MCMC)* (Metropolis et al., 1953; Metropolis & Ulam, 1952), se usa para estimar la distribución a posteriori del modelo. Utiliza una cadena de Márkov para generar una secuencia de muestras de la distribución a posteriori, que se utiliza para aproximar la distribución real. EL MCMC es un algoritmo computacionalmente intensivo, que tiene la ventaja de poder aproximar distribuciones a posteriori arbitrariamente complejas y es generalmente robusto a las suposiciones del modelo. Se inicia con una configuración de partida aleatoria de los parámetros del modelo y luego aplica un proceso de muestreo estocástico para actualizar los parámetros en cada iteración. Este muestreo se basa en la distribución condicional de los parámetros dada la configuración actual, la misma que se calcula a partir de las ecuaciones del modelo LDA (Nguyen et al., 2014).

- *Muestreo de Gibbs* (Geman & Geman, 1984), es una técnica utilizada para estimar la distribución a posteriori de un modelo. Se trata de un algoritmo de simulación que permite muestrear, de manera iterativa, los valores de las variables latentes del modelo dado un conjunto de variables observadas. Este muestreo de Gibbs fue propuesto como una aproximación de inferencia para el modelo LDA; es un método representativo de cadenas de Márkov Monte Carlo (MCMC) (Griffiths & Steyvers, 2004).

Cabe destacar que, todas las técnicas de inferencia antes mencionadas requieren, efectivamente, un proceso computacional de inversión financiera muy significativa.

A efectos de la presente investigación, nos apoyaremos en la propuesta Griffiths y Steyvers, ya que permite obtener un modelo más estricto, aun con variables complejas. Su aplicabilidad es abordada, con la profundidad pertinente, en el capítulo 6.

CAPÍTULO III

MÉTODOS BIPLLOT

3.1. INTRODUCCIÓN

La MDT, como se ha mencionado, es un proceso que se aplica generalmente a una gran cantidad de datos textuales, en la cual, se busca identificar patrones, tendencias y relaciones significativas. En la actualidad, es una herramienta de apoyo necesaria, dado el creciente aumento de la información textual en repositorios digitales a nivel global. En este contexto, el AEDT se ha convertido, sin duda, en un mecanismo imperativo para el análisis de la información extraída a partir de grandes conjuntos de datos textuales.

Dadas estas condiciones, la aplicabilidad de las técnicas estadísticas, han devenido en auge en el contexto de la MDT, destacando, entre otras, los métodos Biplot. Estos, se constituyen en un abordaje muy significativo para la exploración de la relación entre palabras utilizadas, en el conjunto de documentos del corpus; permitiendo visualizar la correspondencia entre palabras y documentos o entre, documentos y temas latentes; fortaleciendo la capacidad de respuesta en la identificación de patrones y relaciones.

Es por ello por lo que las técnicas develadas, en los capítulos previos, nos proporcionan el sustento teórico pertinente, en la configuración de la perspectiva científica que sustente las bases para la extracción y tratamiento de la información textual; con el propósito de coadyuvar en la construcción de las tablas léxicas y encontrar temas latentes ocultos. Es así como, en el presente apartado, se profundizará sobre los métodos Biplot como herramienta estadística en el análisis de datos textuales, a partir de las tablas léxicas de probabilidad del LDA, en función de una representación gráfica de variables (Tópicos) y Filas (Documentos) de manera conjunta.

3.2. LOS MÉTODOS BILOT

Los métodos Biplot, tienen su origen en la década de los 70. Gabriel, los introduce con el objeto de describir aproximadamente una matriz rectangular, utilizando una representación gráfica en baja dimensión; que permita visualizar las interrelaciones entre filas y variables, así como las relaciones entre ambos conjuntos (Gabriel, 1971).

En 1981, Gabriel compara el Biplot propuesto con otras técnicas multivariantes (Gabriel, 1981) y, Cox y Gabriel, en conjunto, realizan una comparación con el Análisis exploratorio

de datos (AED) en 1982, evidenciando que los Biplot generan representaciones más intuitivas para interpretar un conjunto de datos.

Así, al igual que un diagrama de dispersión muestra la distribución conjunta de dos variables, los Biplot representan, gráficamente, tres o más variables en un espacio de dimensión reducida (Gabriel & Odoroff, 1990). De la misma manera, los aportes de Galindo en 1986 trazan un Biplot con alta calidad de representación para las columnas y para las filas conjuntamente (Galindo-Villardón, 1986).

3.2.1. BIPLLOT CLÁSICOS

Efectivamente, los Biplot propuestos por Gabriel son denominados GH y JK Biplot. Estos, se basan en la aproximación de la matriz de datos de partida $X_{n \times p}$ (arreglo rectangular con n filas y p columnas) por una de menor rango q ($q < r$), por medio de la DVS. Es decir, $X \cong U \Sigma V'$, donde U y V son matrices de vectores singulares ortonormales y Σ es la matriz diagonal, que contiene los r primeros valores singulares de X. En este orden, para garantizar la unicidad en la representación, se realiza una factorización en las matrices de marcadores fila y de marcadores columnas tal que, $X \cong (U \Sigma^s)(\Sigma^{1-s} V') = AB'$, siendo A y B las matrices que contienen las coordenadas de los (n+p) vectores o marcadores filas a_i y, columnas b_j a utilizar en la representación gráfica (Gabriel, 1971).

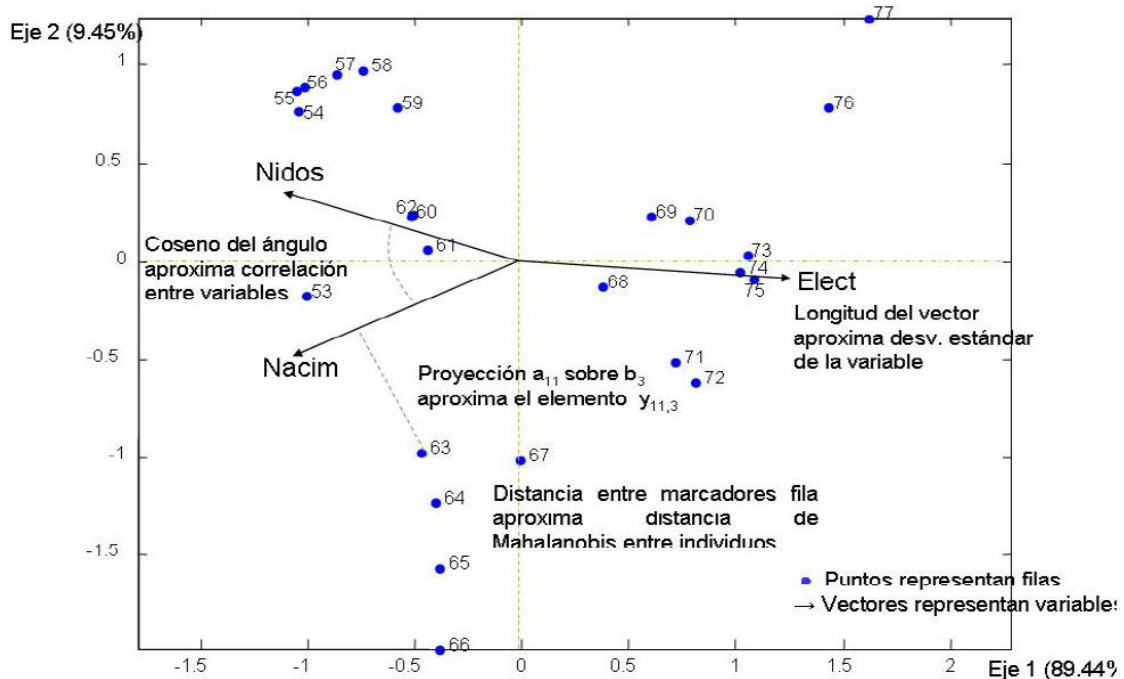


Figura 3.1. Representación Gráfica de Biplot de Gabriel (1971).

Así, dependiendo de los valores dados al escalar s en la factorización, se pueden obtener los Biplot clásicos propuestos por Gabriel, el GH Biplot y el JK Biplot. De esta manera, las propiedades de los marcadores filas y columnas, están dadas por la factorización elegida, la cual, depende de la métrica introducida en el espacio de las filas o en el espacio de las columnas.

3.2.1.1. GH-BILOT

El GH Biplot, tiene notación G , en referencia a la matriz de los marcadores fila y, H , para la matriz de los marcadores columna. En el GH se aproxima la matriz X bajo la restricción $U'U = I$ (siendo I la matriz identidad, tal que $U'U = V'V = I$), en este caso, los marcadores columnas b_j son los resaltantes; ya que preservan la métrica entre las columnas y tienen una calidad de representación óptima.

$$G = U \quad H = V\Sigma \quad (23)$$

3.2.1.2. JK-BILOT

El JK Biplot, tiene notación J para los marcadores filas y, H , en referencia a los marcadores columna. En el JK se aproxima la matriz X bajo la restricción $V'V = I$, siendo los más resaltantes los marcadores filas a_i ya que preservan la métrica entre filas; demostrándose que tienen una calidad óptima de representación.

$$J = U\Sigma \quad K = V \quad (24)$$

En la figura 3.2, se observa la representación de los marcadores en cada uno de los Biplot propuestos por Gabriel.

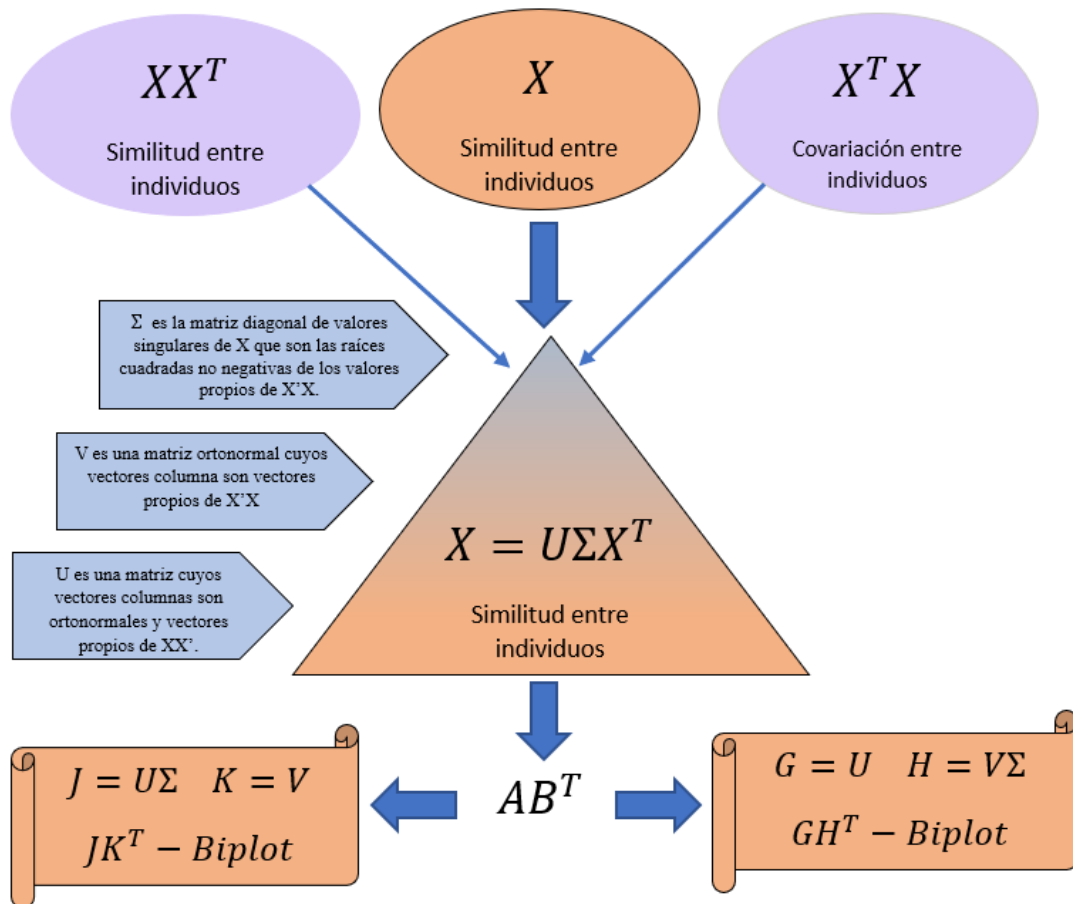


Figura 3.2. Marcadores de los Biplot de Gabriel (1971).

3.2.2. HJ-BIPLLOT

Teniendo como antecedente los aportes sobre Biplot propuestos por Gabriel, Galindo en 1986 (Galindo-Villardón, 1986) plantea una alternativa para proyectar, en un espacio de baja dimensión y con una alta calidad de representación, las filas y columnas simultáneamente.

$$H = V\Sigma \quad J = U\Sigma \quad (23)$$

Donde, H representa las buenas propiedades de los marcadores columnas del GH, y J las buenas propiedades de los marcadores filas del JK. Así, el HJ Biplot se constituye en una representación gráfica multivariante de las líneas de una matriz de partida $X_{n \times p}$ mediante los marcadores j_1, \dots, j_n para sus filas y h_1, \dots, h_p para sus columnas; elegidos de manera que, ambos marcadores, puedan ser superpuestos en un mismo sistema de referencia con máxima calidad de representación (Galindo-Villardón, 1986). Así, Galindo y Cuadras, demuestran que las relaciones entre las nubes de punto son las relaciones baricéntricas análogas a las del

AFC; interpretando las posiciones de las filas y las columnas, mediante las contribuciones del factor al elemento y del elemento al factor (Galindo-Villardón & Cuadras, 1986).

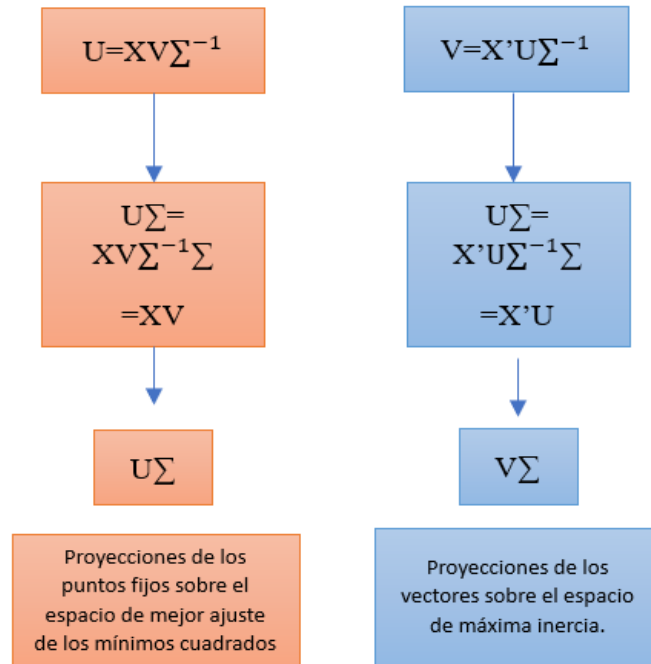


Figura 3.3. Marcadores Fila y Columna del HJ-Biplot. (Cubilla Montilla, 2019).

Así, $U\Sigma$ coincide con la proyección de los n puntos, que representan a las filas sobre el espacio de menor ajuste a esa nube de puntos en el sentido de los mínimos cuadrados. Por lo tanto, los marcadores para las filas en el HJ Biplot, coinciden con las coordenadas de las filas respecto a los ejes factoriales.

$V\Sigma$ coincide con la proyección de los puntos que representan a las variables, sobre el espacio de máxima inercia. Es decir, los marcadores columnas, coinciden con las coordenadas de las columnas respecto a los ejes factoriales.

En un HJ Biplot, no tiene sentido hablar de bondad de ajuste para los elementos de la matriz original X , ya que no se considera como una aproximación de bajo rango de esta. Sin embargo, tiene la ventaja de que es una representación simultánea en sentido estricto, dada las buenas propiedades, alta calidad de representación de filas y columnas y, la posibilidad de interpretar las posiciones de las filas, de las columnas y de las relaciones fila-columna como en el análisis de correspondencia (Galindo-Villardón & Cuadras, 1986). En la tabla 2.1 se observa la bondad de ajuste tanto para filas y columnas.

Representación Simultanea	Coordenadas Filas	Coordenadas Columnas	Bondad Ajuste Filas	Bondad Ajuste Columnas
GH-Biplot	U	$V\Sigma$	$\frac{2}{r}$	$\frac{\lambda_1^2 + \lambda_2^2}{\sum_{\alpha=1}^r \lambda_\alpha^2}$
JK-Biplot	$U \Sigma$	V	$\frac{\lambda_1^2 + \lambda_2^2}{\sum_{\alpha=1}^r \lambda_\alpha^2}$	$\frac{2}{r}$
HJ-Biplot	$U \Sigma$	$V\Sigma$	$\frac{\lambda_1^2 + \lambda_2^2}{\sum_{\alpha=1}^r \lambda_\alpha^2}$	$\frac{\lambda_1^2 + \lambda_2^2}{\sum_{\alpha=1}^r \lambda_\alpha^2}$

Tabla 3.1. Bondad de Ajuste de Biplot de Gabriel (1971), y de Galindo (1986).

3.2.3. INTERPRETACIÓN HJ BILOT

Los métodos Biplot, son aplicables a diferentes matrices de datos, no sólo para los casos donde las columnas representen variables y las filas a individuos, sino también en situaciones como: una matriz de covarianza (Gower, 1995), una tabla léxica de frecuencia de palabras (Z. M. Osuna, 2006), entre otras.

Usualmente, en los Biplot, las filas o individuos de la matriz se representan por puntos (marcadores filas) y las columnas con vectores (marcadores columna). La interpretación del HJ Biplot, sobre un plano Bidimensional, se basa en conceptos geométricos (Galindo-Villardón & Egido, 2009), como:

- La proximidad que existe entre los marcadores filas se interpreta como una función inversa de la distancia entre los mismos, lo que indica similitud entre ellos.
- La Longitud de los vectores, indican variabilidad en el aporte de información de las variables.
- El ángulo, entre vectores, indica la covariación de las variables; donde un ángulo pequeño, entre vectores, indica más relación directa entre ambos. Mientras que, si el ángulo es cercano a 180°, más relación inversa existe entre estos y, entre más cerca a 90° estén los vectores, más independencia se da.
- La proyección de cada marcador fila (punto) sobre los marcadores columna (vectores), permiten interpretar las relaciones entre filas y columnas, en términos de producto escalar. Si el punto se encuentra en la misma dirección que la punta del vector, esto indica valores altos sobre dicha variable.

– Los ejes factoriales del plano pueden interpretarse evaluando las contribuciones de cada variable al gradiente latente.

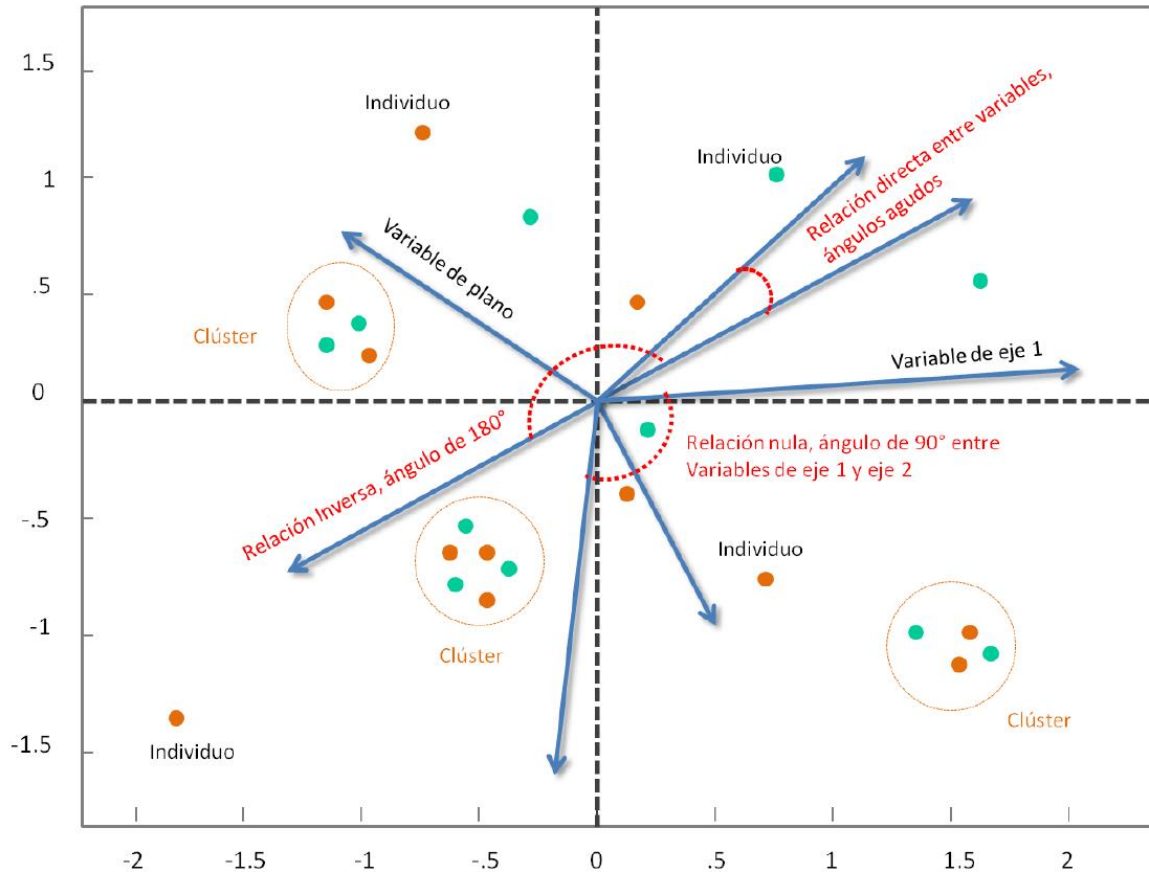


Figura 3.4. Interpretación Gráfica del HJ-Biplot. Tomado de (Ballesteros, 2022)

3.3. BIPLLOT EN EL ANÁLISIS DE DATOS TEXTUALES.

El uso de los métodos Biplot para representación de datos textuales ha venido en crecimiento en los últimos años siendo, la Universidad de Salamanca, una de las instituciones vanguardia en estos estudios, destacadas en publicaciones registradas en Elsevier (Scopus) y Web of Scienc (WOS).

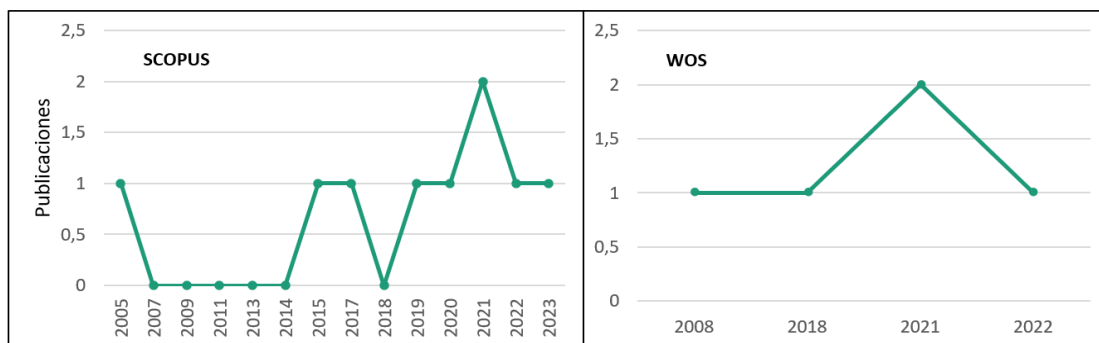


Figura 3.5. Publicaciones en Scopus y WOS de USAL en MDT a abril 2023.

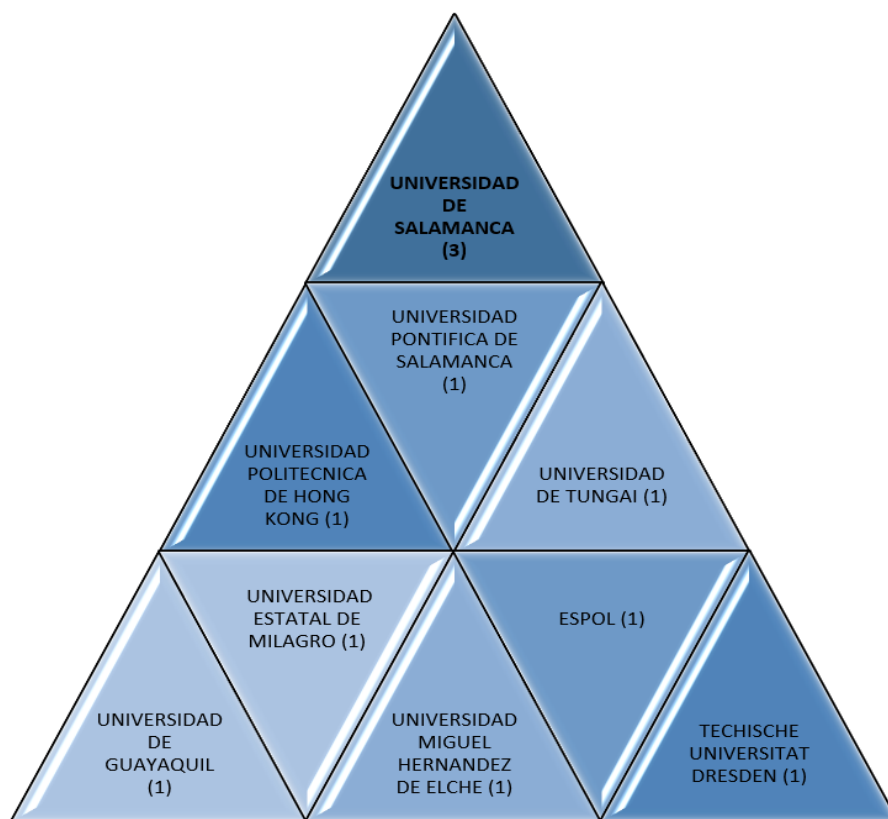


Figura 3.6. Instituciones con Publicaciones en WOS, referentes a MDT a abril 2023.

Así, las técnicas empleadas en el AEDT, parten de la DVS de una matriz conocida como tabla léxica. Lebart, expone al análisis de texto como una fuente cuantitativa de información, a través del tratamiento de los datos textuales (Lebart et al., 2000), la cual, estandariza, segmenta y lematiza el contenido de un texto. Así mismo, Zulaima Osuna, explica una descripción más detallada de las tablas léxicas y expone las principales técnicas que varios autores han preferido para el análisis de texto, tales como el análisis factorial de correspondencia. Adicionalmente, propone el uso del HJ-Biplot y el uso de Biplot robustos, a partir de tablas léxicas de frecuencia con las repeticiones de las unidades léxicas (frases cortas, lemas o palabras) (Z. M. Osuna, 2006). A diferencia del trabajo de Osuna, este proyecto amplió la aplicación del HJ-Biplot al análisis de matrices de probabilidad obtenidas del LDA, así mismo integra el HJ-Biplot con un léxico de factores para relacionar el macroentorno PESTEL con los temas latentes encontrados en el corpus.

De esta manera, dados los múltiples estudios de datos textuales, usando técnicas Biplot, publicados en WOS se tienen:

En este orden, Osuna, presenta una propuesta metodológica basada en el AEDT de corpus cronológicos, organizando, en una tabla de frecuencia de palabras, las declaraciones del Libertador Simón Bolívar de los años de 1812 y 1826, usando el AFC para examinar el texto (Z. Osuna et al., 2004). Posteriormente, Osuna, con base en estos aportes, aplicó el HJ-Biplot y Biplot robustos para mejorar las representaciones de las palabras con los discursos analizados (Z. M. Osuna, 2006).

De la misma manera, Pan 2008, usando el software atlas.ti, genera una representación Biplot de categorías de palabras referente al turismo en diferentes sectores de Asia. Ello, a partir de una tabla léxica que los autores construyeron definiendo como variables cuatro categorías que representaban a los 4 sectores de Asia y, como individuos, codificaron diferentes temas relacionados al turismo. La matriz representaba la frecuencia de los códigos en los diferentes sectores de Asia generando, a partir de esta tabla léxica, un Biplot y, estudiar así, las relaciones de los diferentes códigos afines al turismo en cada una de las zonas analizadas (Pan et al., 2008).

Así mismo, en el 2014, Caballero analizó grupos de discusión de calidad de vida de los ludópatas, a partir de una propuesta metodológica que combina la codificación cualitativa con un nuevo valor matemático de caracterización para las tablas léxicas de frecuencia de palabras, generando, a partir de estas nuevas tablas codificadas, una representación HJ-Biplot (Caballero-Julia et al., 2014). Seguidamente, Caballero, amplía la codificación del texto para una construcción de una nueva tabla léxica que permite calcular los pesos relativos de cada palabra en la tabla, según su presencia específica y, en un documento, en lugar de centrarse sólo en las frecuencias altas (Caballero-Julia & Campillo, 2021), para luego, representarlos por medio de un Manova Biplot.

Cabe destacar, también, el trabajo de Martín-Rodero, donde utilizó el Biplot logístico para evaluar la calidad de las revisiones metodológicas sobre enfermedades nutricionales y metabólicas. El método se usó a partir de una matriz donde se categorizó las diferentes tipologías documentales y clasifica a los documentos por su calidad metodológica (Martín-Rodero et al., 2018). Seguidamente, en el 2021 Kienberger, como parte de una investigación en tres universidades españolas de las estrategias implementadas para la enseñanza del alemán, realizó una representación Biplot a partir de tablas léxicas de frecuencia, para

estudiar las relaciones de las diferentes estrategias y de acuerdo con el significado de las palabras durante el aprendizaje (Kienberger et al., 2021).

Significativos son los aportes de Pilacuan, el cual propone incorporar en las matrices de probabilidad resultantes del modelo LDA, una transformación que permita incorporar, a esta tabla léxica de probabilidades, el uso de la representación HJ-Biplot (Pilacuan-Bonete, Galindo-Villardón, & Delgado-Álvarez, 2022). Este procedimiento que es parte del objeto de estudio de esta tesis será profundizado en el capítulo siguiente.

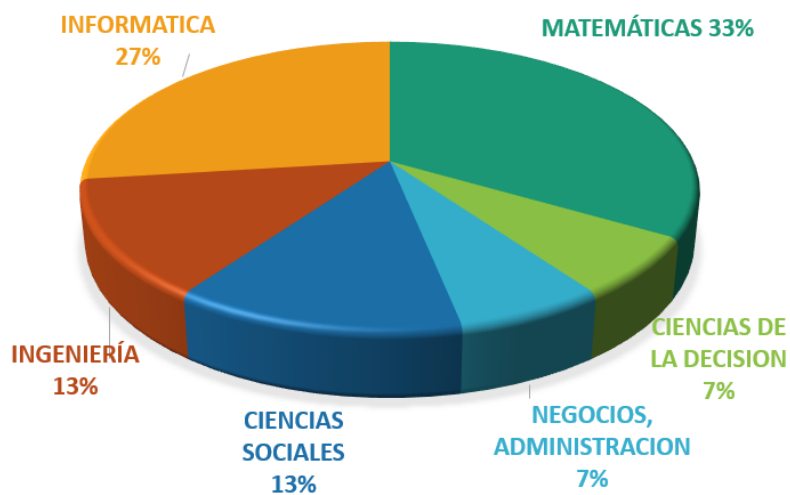


Figura 3.7. Áreas de Publicaciones en MDT de WOS a abril 2023

Como se puede observar en la figura 3.5, el AEDT, se ha usado en múltiples campos de la ciencia, combinando los métodos tradicionales de minería de texto e incorporando técnicas novedosas en la forma de construcción de las tablas léxicas para generar las representaciones Biplot en los análisis.

CAPITULO IV

PESTEL

4.1. INTRODUCCIÓN.

Sin duda, reviste importancia vital la evaluación de los entornos externos que afectan a las organizaciones y su desempeño. Ello puede afectar su funcionamiento y éxito en el ámbito organizacional y empresarial. Para Perera (2020), el estudio del entorno externo en el que operan las organizaciones permite identificar oportunidades y amenazas a las que se enfrentan, ayudándolas de esta manera a la toma de decisiones y al desarrollo de estrategias que se ajusten a las condiciones del cambiante entorno (Perera, 2020).

En función de esto, resulta imperativo el análisis de herramientas que permitan la realización de diagnósticos que puedan coadyuvar a las organizaciones en la toma de decisiones. En este orden de ideas, el presente apartado aborda el concepto, aplicación e interpretación de la metodología PESTEL y su potencial integración en el contexto de la minería de datos textuales web con las técnicas multivariantes BIPLLOT, y poder entender las relaciones de los factores con los tópicos.

4.2. PESTEL

En cualquier contexto, sea empresarial, corporativo, administrativo, público o político, siempre el tomar decisiones es un proceso más o menos complejo, donde se pueden ayudar con el uso de múltiples herramientas metodológicas de apoyo a la decisión (J. P. A. Fernandes & Guiomar, 2016).

En 1968, Fahey y Narayanan presentan un análisis macro ambiental en gestión estratégica aplicado al marketing, donde se expone una herramienta de diagnóstico enfocado en 6 magnitudes externas que afectan directa o indirectamente al giro de negocio de las organizaciones (Fahey & Narayanan, 1968).

Esta herramienta se denomina PESTEL, que es un acrónimo de los siguientes Factores que afectan cualquier estructura o incluso individuos:

- P => Factores Políticos.
- E => Factores Económicos.
- S => Factores Sociales.

- T => Factores Tecnológicos.
- E => Factores Ambientales (E = Environmental).
- L => Factores Legales.

Esta técnica es usada en la fase de diagnóstico del desarrollo de planes estratégicos y toma de decisiones de las organizaciones, donde en primer lugar se debe de realizar el análisis de las condiciones internas y externas que determinaran la capacidad para alcanzar las metas (Gillespie, 2014).

4.3. FACTORES PESTEL.

Cuando se habla del entorno o contexto donde las organizaciones interactúan, se refiere a aquellos factores externos que son relevantes para el éxito y buen funcionamiento de la empresa. Es así como el análisis de los diversos factores que atañen a la organización resulta vital para generar campañas y estrategias, tanto a corto, mediano o largo plazo. Estos factores Perera (2020) los detalla de la siguiente manera (Perera, 2020):

- 1. Factores Políticos**, relacionados con la política, el gobierno y las regulaciones que pueden influir en el entorno de un sector o una organización. Tales como, cambios en la legislación, políticas públicas o fiscales, entre otros.
- 2. Factores Económicos**, aquellos que pueden afectar a una organización en la parte financiera, como por ejemplo tasas de interés, inflación, desempeño, cambios en los mercados financieros, etc.
- 3. Factores Sociales**, engloba los aspectos sociales, culturales y demográficos que pueden tener un impacto en el entorno empresarial, como los cambios en estilo de vida, actitudes del consumidor, entre otras.
- 4. Factores Tecnológicos**, abarca avances tecnológicos y su impacto en la industria o en la forma en que se realizan los negocios, desarrollos tecnológicos, adopción de nuevas tecnologías, entre otros términos.
- 5. Factores Ambientales**, se relacionan con el medio ambiente o sostenibilidad, como regulaciones ambientales, impacto ambiental, cambio climático, etc.
- 6. Factores Legales**, estos incluyen los aspectos jurídicos y legales, tales como regulaciones laborales, normativas de salud y seguridad, leyes de propiedad intelectual, entre otras.



Figura 4.1. Factores PESTEL.

Los resultados del PESTEL normalmente son listados de factores agregados según cada temática, el cual se complementa con un ranking de su importancia en el contexto de análisis, así como su manejabilidad e impacto en el corto y largo plazo (J. P. Fernandes, 2019).

Este tipo de análisis ha sido usado en diferentes contextos tales como evaluar el desarrollo de la industria automotriz de vehículos eléctricos en Brasil (de Sousa & Castañeda-Ayarza, 2022), analizar los aprendizajes del brote del COVID, evaluar la gestión de residuos sanitarios (Thakur, 2021), incluso para el análisis de los entornos que afectan a Países como India (McManus et al., 2007b) o China (McManus et al., 2007a). En todos estos estudios ha sido usado para el diagnóstico del ambiente externo en el que se desarrollan las industrias, en los diferentes entornos de los Países o sectores analizados. Este método es generalmente combinado con otras técnicas como el análisis de fuerza de Porter o el análisis de Amenazas, Debilidades, Fortalezas y Oportunidades (DAFO). Juntos, se constituyen en valiosos mecanismos que permiten determinar estrategias para el desarrollo organizacional o Institucional.

En este sentido, el presente trabajo de investigación integra en el análisis textual, un léxico que va a permitir identificar palabras dentro de los términos que forman los tópicos generados, y así poder precisar los factores externos que están relacionados con los diferentes tópicos encontrados de la extracción de noticias localizadas en la web.

CAPÍTULO V

*HJ-BILOT COMO HERRAMIENTA PARA
DAR UN IMPULSO ANALÍTICO ADICIONAL AL
MODELO DE ASIGNACIÓN LATENTE DE
DIRICHLET*

5.1. INTRODUCCIÓN

En los apartados previos se han podido analizar y estudiar a profundidad los aspectos significativos relacionados a la MDT y al AEDT. Partiendo de estas premisas, en el presente capítulo se hace un ejercicio de integración conceptual de todos estos fundamentos, con el propósito de sentar las bases en la generación de una metodología que permita sustentar el impulso para el análisis de modelamiento de tópicos obtenidos a partir de modelos probabilísticos LDA. Efectivamente, este abordaje metodológico resulta particularmente significativo, ya que implementa la extracción de texto, a partir de la web, mediante técnicas de web scraping, desvelando la aplicabilidad sobre la transformación de las matrices resultantes del LDA, para la generación de las representaciones Biplot, con especial énfasis en el HJ-Biplot.

De la misma manera, nos permite generar una representación HJ-Biplot integrando un listado léxico que identifica el contenido Político, Económico, Social, Tecnológico, Ambiental y Legal (PESTEL), en correspondencia a los tópicos obtenidos a partir del análisis LDA.

Es importante resaltar que en la actualidad existe abundante información textual disponible en la web y que se puede encontrar en noticias, redes sociales, sitios institucionales y más. De esto se puede inferir la necesidad de seguir efectuando iniciativas, en las que investigadores y académicos centren sus esfuerzos en el análisis de la extracción de texto de la web o de bases de datos estructuradas. Con el fin de promover nuevas perspectivas para la generación de un enfoque integral a partir de los hallazgos teóricos de la presente investigación, aplicando metodologías que permitan la determinación de temas ocultos, sus relaciones e interacciones de estas con el entorno PESTEL, se desarrolla un proceso metodológico nombrado LDABiplots.

5.2. METODOLOGÍA LDABIPLOTS.

Se hace evidente que, en correspondencia con las técnicas revisadas a lo largo de esta investigación, estas permiten el análisis de datos textuales. Todos estos aportes técnicos se han tratado de manera individual o agrupadas con otras técnicas que no necesariamente se relacionan al objeto del presente trabajo.

Por ello, se infiere la incorporación de la integración del modelado de tópicos LDA con las representaciones Biplot, enfatizando en el HJ-Biplot, e incluyendo la extracción de texto a partir de la Web, así como un análisis de los factores del macroentorno PESTEL (Fahey & Narayanan, 1968) de los textos analizados.

En este orden de ideas, se propone el siguiente orden metodológico a aplicarse, denominado LDABiplots:

- i. Definir términos claves de búsqueda en la WEB.
- ii. Realizar el scraping de la Web para la extracción de la información que contenga los términos claves.
- iii. Estructurar la información obtenida del scraping.
- iv. Preprocesar el texto mediante la Tokenización, eliminación de palabras que no aportan a los análisis (stop word) y Normalización.
- v. Construir tablas léxicas
- vi. Realizar el modelado de tópico LDA.
- vii. Obtener las matrices de probabilidad del LDA
- viii. Transformar la matriz que indica la probabilidad de que cada palabra pertenezca a cada tópico generado.
- ix. Generar las representaciones Biplot, y HJ-Biplot incorporando el listado de palabras que representan el entorno PESTEL (HJ-Biplot_PESTEL).
- x. Generar representaciones tradicionales de los tópicos obtenidos.

En la figura 4.1, se observa el esquema de la metodología propuesta, para la integración de los métodos Biplot en el modelado de tópicos.

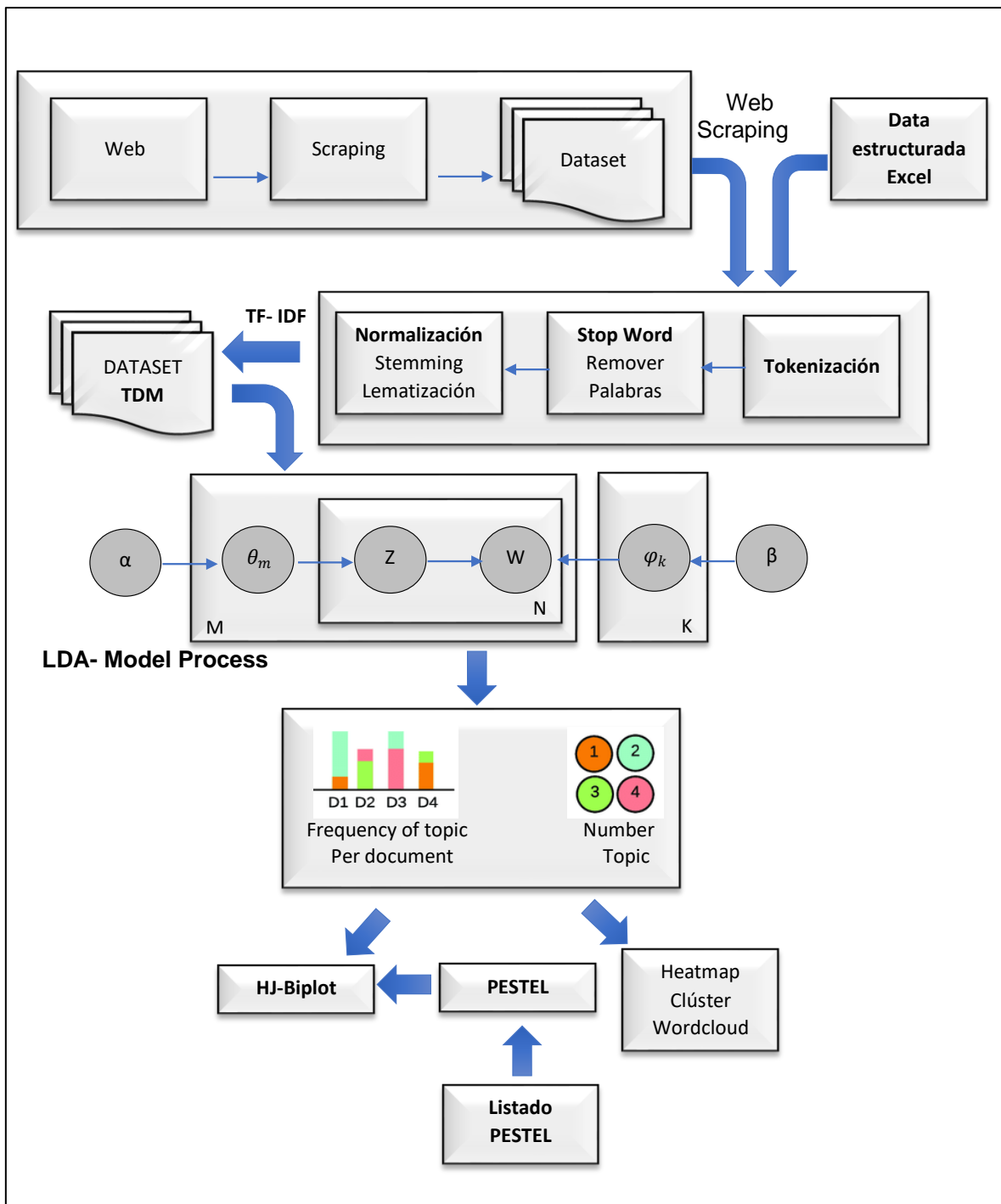


Figura 5.1. Representación Metodología LDABIPLOTS

LDABIPLOTS, nos va a permitir representar mediante un HJ-Biplot los tópicos generados por el LDA de un corpus de partida, para que, integrando la matriz phi obtenida y el léxico PESTEL, podamos observar gráficamente las relaciones de las palabras en su contexto del macroentorno con los tópicos del modelado. Así como, obtener en un espacio de baja

dimensión, una alta calidad de representación de los tópicos obtenidos y documentos analizados simultáneamente.

5.3. INTEGRACIÓN DE MÉTODOS BILOT AL ANÁLISIS DEL MODELADO DE TÓPICOS LDA

Una vez planteada la metodología, y antes de explicar la integración de estos métodos en un software que admita su aplicabilidad, se revisará, con detalle, el proceso de transformación de las matrices obtenidas del modelo LDA, así como la incorporación del listado de palabras asociadas al PESTEL.

5.3.1. TRATAMIENTO DE MATRICES DEL LDA.

Tal como se explicó en el capítulo dos, destaca el denominado LDA como uno de los métodos probabilísticos usados para el modelado de temas o tópicos. En este sentido, la presente investigación tiene como sustento fundamental el modelo propuesto inicialmente por Blei (D. M. Blei et al., 2003), que luego incorporó algoritmos eficientes para las inferencias (Steyvers & Griffiths, 2007), usando un modelo generativo para inferir la distribución de temas para documento, así como la distribución de palabras para cada tema (D. M. Blei & Lafferty, 2009).

Desde estas premisas, nos enfocaremos en las matrices generadas al aplicar el Teorema de Bayes. El propósito es inferir la distribución de temas en cada documento y distribución de palabras para cada tema, al maximizar la probabilidad posterior de los datos, dados los parámetros del modelo.

a. Matriz PHI.

La matriz que muestra la distribución de palabras para cada tema se conoce como Phi (φ). Esta resulta del cálculo denotado en la ecuación (21). Así, el algoritmo LDA asigna cada palabra de cada documento a un tema y luego calcula la probabilidad de que cada palabra aparezca en cada tema (D. M. Blei & Lafferty, 2009).

b. Matriz Theta.

La matriz theta (θ), es la resultante de la aplicación de la ecuación (19) en el LDA. Muestra la probabilidad de que cada documento pertenezca a cada tema o tópico. En este orden, el algoritmo calcula la probabilidad de que cada documento pertenezca a cada tema, a partir de la asignación de cada palabra, de cada documento, a un tópico.

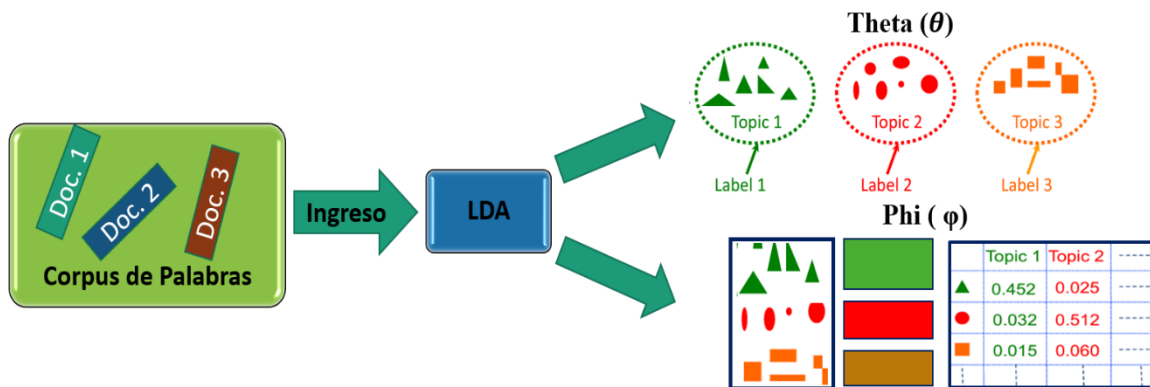


Figura 5.2. Esquema de Obtención de Theta y Phi del LDA de Blei (2003)

Desde este contexto, para la representación de estas matrices se han venido usando varias técnicas, como el LDavis (Sievert & Shirley, 2014), la cual, integra el escalado multidimensional (MDS) para visualizar la similitud de los tópicos, en referencia a los términos que lo contienen. También por medio de un dendograma de calor (HeatMap), que muestra los valores reordenados por la media o alguna medida dada por el usuario, de un conjunto de datos, en una escala de colores (Gu et al., 2016).

5.4. HJ-BILOT COMO REPRESENTACIÓN DE MATRICES DE PROBABILIDAD DEL LDA.

En la presente intención doctoral, se propone la representación HJ-Biplot, para la visualización de d documentos, los cuales, tienen características en común como, por ejemplo, autores, diarios de noticias, año en particular, revistas, entre otras. Es decir, el corpus está formado por múltiples d documentos que se clasifican por una característica en común, en la cual puede agruparse.

Así, para la realización del análisis propuesto, primero, se debe obtener la matriz θ , que nos muestra la probabilidad conjunta de que cada documento pertenezca a cada tópico. A partir

de la traspuesta de esta matriz, se calcula la probabilidad media de cada conjunto de documentos agrupados, de acuerdo con la característica en común analizada por el usuario (ejemplo múltiples noticias de un diario, donde cada noticia se considera un documento). Es decir, si se tienen cinco noticias de un periódico, cada noticia es un documento que tiene una probabilidad de pertenecer a un tópico. Se calcula la probabilidad media de los cinco documentos, que sería la probabilidad media de ese diario de pertenecer a un tópico.

$$\overline{d}_{ik} = \frac{d_{i1} + d_{i2} \dots + d_{in}}{D} \quad (24)$$

Donde;

\overline{d}_{ik} , es la probabilidad media de cada i-ésimo conjunto de documento pertenezca a k-ésimo tópico.

d_{in} , es la probabilidad de que un i-ésimo documento pertenezca a un n-ésimo tópico.

D, es la longitud del conjunto de documentos.

θ^T , Matriz traspuesta de theta.

A partir del cálculo de cada \overline{d}_{ik} , de todos los conjuntos de documentos de la matriz θ^T , estos valores de probabilidad media son representados en una matriz X; donde los individuos son los tópicos generados en el modelo LDA y las variables son los conjuntos de Documentos. De esta forma se obtiene la representación HJ-Biplot, que permite estudiar las relaciones entre tópicos generados o, entre los documentos o, las contribuciones que tienen los tópicos, con cada conjunto de documentos y en correspondencia con el procedimiento explicado en el capítulo 3.

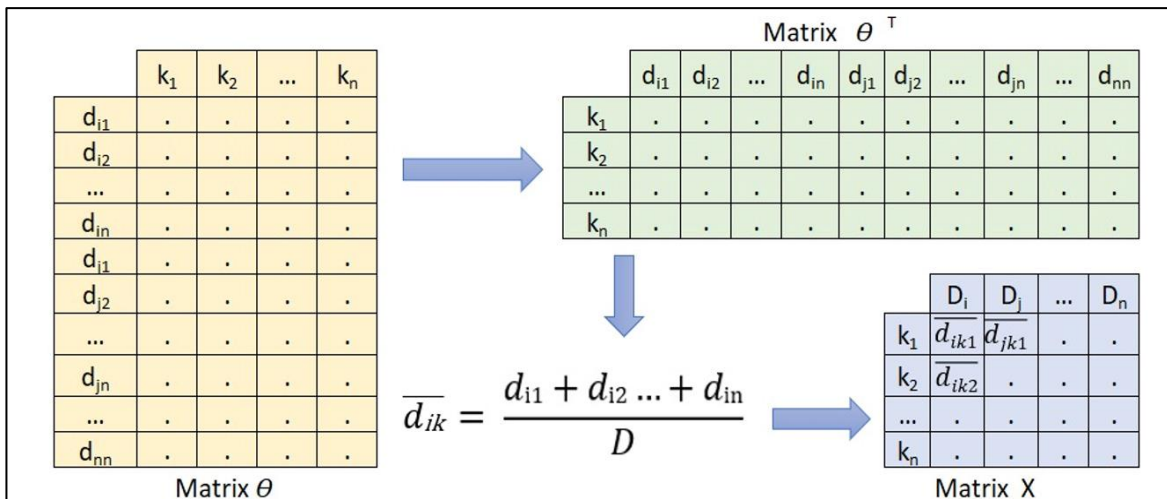


Figura 5.3. Esquema de transformación de matriz theta, para el HJ-Biplot.

Hasta lo analizado en este apartado, se ha generado un estudio práctico de las noticias de España, referentes al COVID-19, donde se analizaron 48112 noticias del 1 de enero del 2019 al 27 de mayo del 2022 de los tres diarios de mayor circulación (Pilacuan-Bonete, Galindo-Villardón, & Delgado-Álvarez, 2022), el cual se encuentra publicado en la revista *Mathematics* que actualmente es Q2 en ranking SJR y Q1 en ranking JCR (Aporte científico 1).

Cabe destacar que, la propuesta en estudio permite integrar un léxico de palabras relacionadas al análisis de los factores del macroentorno de una organización, tales como: Políticos, Económicos, Sociales, Tecnológicos, Ambientales y Legales. Esta integración de palabras se realiza a partir del resumen del modelo resultante de la matriz phi (φ), donde se detallan las palabras que componen cada modelo.

Así, para la representación HJ-Biplot de la matriz phi, identificando los factores PESTEL, se genera la visualización con seis clústeres; cada uno indica uno de los factores del léxico. De acuerdo con ello, cada palabra que conforma este listado se convierte en un identificador de cada palabra de la matriz phi que se analiza en el Biplot.

Esto nos permite analizar las relaciones entre palabras caracterizadas por los seis factores del macroentorno, así como las contribuciones de estas palabras con los tópicos. Adicionalmente, se pueden observar las relaciones que tienen los tópicos, considerando las palabras analizadas en el modelado LDA.

Es significativo destacar que todos los análisis explicados, hasta este momento de la investigación, requieren un soporte computacional para su desarrollo. Esto es debido a que el modelado LDA requiere del cálculo de una inferencia, explicada en el apartado 2.4.2.3, que detallaremos en el capítulo 6, así como la obtención del HJ-Biplot y la identificación de los clústeres, lo que requiere de un proceso ágil para su desarrollo.

Es así como todas estas técnicas serán propuestas en un software, de acceso abierto, para la comunidad de investigadores y científicos; como aporte significativo de esta iniciativa doctoral.

CAPÍTULO VI

LDABILOTS

6.1. INTRODUCCIÓN

Hemos descrito en apartados anteriores la teoría de los métodos analizados, así como su integración, con el objeto de poder generar representaciones Biplot. Estas representaciones nos van a permitir interpretar y analizar relaciones entre tópicos, contribuciones de los tópicos a los documentos, relaciones y similitud de documentos. Integran una clasificación, para identificar las palabras que contienen los tópicos, lo que permite relacionarlas con el entorno macro ambiental PESTEL.

Con la globalización, el avance de los procesos digitales y la masificación del internet en todas las áreas de la vida, se hace casi necesario que los procesos de investigación cuenten con herramientas que permitan realizar análisis más eficientes y rápidos. Los investigadores, en vista de esta necesidad, buscan mecanismos digitales que les permitan ser eficientes en obtener los resultados, por ello siempre persiguen herramientas que integren nuevos sistemas computacionales.

Estas operaciones, sin ayuda de sistemas computacionales se volverían tediosas y demoradas, haciendo complejos sus análisis. Por ello se propone integrar todas las técnicas expuestas en un ambiente computacional de acceso libre.

En este capítulo revisaremos todos los procesos de integración desde la extracción web de texto, hasta las representaciones Biplot, integrando el léxico PESTEL al HJ-Biplot. Analizaremos paso a paso las entradas requeridas para el modelado y las salidas resultantes de cada proceso, con la finalidad que el usuario pueda realizar los análisis de manera eficiente y rápida.

6.2. PAQUETE LDABILOTS

El uso de software libre se ha extendido entre desarrolladores, estadísticos, matemáticos e investigadores de diversas áreas. R es un lenguaje de programación de código abierto, conocido como un entorno de software libre para computación, estadísticas y representaciones gráficas, usado en una variedad de plataformas como Windows, MacOS y Unix (R Core Team, 2023).

En el repositorio de R, conocido como Comprehensive Archive Network (CRAN), se cuenta con 60 paquetes agrupados dentro del tópico de procesamiento de lenguaje natural (Fridolin, 2022), y en la biblioteca rdrv.io, donde se almacena un índice completo de paquetes de R, se cuenta con 284 paquetes relacionados con modelado de tópicos y 18 paquetes que permiten realizar representaciones Biplot.

Entre los más descargados para el análisis LDA, tenemos:

- **lda**, desarrollado por Chang en 2015, donde implementa LDA y modelos relacionados, usando el muestro de Gibbs colapsado para la inferencia (J. Chang, 2015).
- **ldatuning**, donde Nikita, implementa diferentes métricas para estimar el número de K óptimo más adecuado para los modelos (Nikita & Chaney, 2020).
- **topicmodels**, se proporciona una interfaz para el código c para LDA y modelo de temas Correlacionados (CTM), con un ajuste para los modelos basado en el muestreo de Gibbs (Grün et al., 2023).
- **LDavis**, permite crear visualizaciones interactivas basadas en la web de un modelo de tópicos mediante LDA, creando una interacción interactiva con D3.js en un navegador web (Sievert & Shirley, 2014).

No se encontraron paquetes en el CRAN de R que integren LDA con los métodos Biplot, adicionales al generado como producto de la presente investigación. Basado en esto, el objetivo de la creación del paquete LDABiplots es permitir que los usuarios generen modelos LDA y representaciones Biplot con mayor agilidad y de manera interactiva, en referencia a la metodología propuesta, los códigos del paquete se pueden encontrar en la CRAN y en siguiente enlace de GitHub, <https://github.com/Pilacuan-Bonete-Luis/LDABiplot>

Para generar un proceso interactivo de uso, se utiliza el paquete Shiny, el cual permite crear aplicaciones fáciles de usar para los usuarios, permitiendo que estos interactúen con los datos y sus análisis (W. Chang et al., 2022). Existen 170 paquetes interactivos desarrollados con shiny en el CRAN de R, algunas de estas permiten generar modelos LDA, como:

- **Shinylda**, el cual permite crear modelos LDA a partir de la carga de archivos con una sola entrada de texto por fila (Boyes, 2020).
- **LDashiny**, la cual proporciona una interfaz interactiva para realizar revisión de la

literatura científica usando solamente el modelado LDA (De La Hoz-M et al., 2021). Boyes y de la Hoz no aplican representaciones Biplot en sus paquetes shiny de R. Sin embargo, De la Hoz (2020) aplica en un estudio de las publicaciones científicas de acuicultura el Biplot dinámico, para representar los tópicos predominantes obtenidos del LDA en las diferentes revistas científicas de los años 2000 a 2019, estudiando la evolución de las publicaciones por los años a partir solamente de la matriz theta (De La Hoz Maestre, 2020) usando el paquete de R dynBiplotGUI (Egido, 2020). En nuestra propuesta se genera un paquete que integra el raspado web (webscraping), el modelado de tópicos LDA, las técnicas Biplots y un léxico de factores PESTEL, para representar las palabras del corpus de la matriz phi y etiquetadas por los factores del macroentorno para representarlas en un HJ-Biplot, esta representación la llamaremos LDA_HJ-Biplot_PESTEL.

El entorno de R, soportado por la infraestructura del CRAN, permite que las aplicaciones estén a disposición de toda la comunidad que usa este entorno para sus análisis. El paquete LDABiplots está ya disponible gratuitamente desde el CRAN de R (<https://cran.r-project.org/web/packages/LDABiplots/index.html>), y a la fecha se cuenta con más de 1850 descargas desde su publicación en Julio del 2022 (<https://cranlogs.r-pkg.org/badges/grand-total/LDABiplots>).

LDABiplots: Biplot Graphical Interface for LDA Models





Contains the development of a tool that provides a web-based graphical user interface (GUI) to perform Biplots representations from a scraping of news from digital newspapers under the Bayesian approach of Latent Dirichlet Assignment (LDA) and machine learning algorithms. Contains LDA methods described by Blei, David M., Andrew Y. Ng and Michael I. Jordan (2003) <<https://jmlr.org/papers/volume3/blei03a/blei03a.pdf>>, and Biplot methods described by Gabriel K.R(1971) <[doi:10.1093/biomet/58.3.453](https://doi.org/10.1093/biomet/58.3.453)> and Galindo-Villardón P(1986) <<https://diarium.usal.es/pgalindo/files/2012/07/Questiio.pdf>>.

Version: 0.1.2

Imports: [shiny](#), [shinyBS](#), [shinydashboard](#), [shinyWidgets](#), [shinyalert](#), [shinybusy](#), [shinyjs](#), [shinycssloaders](#), [dplyr](#), [ggplot2](#), [rvest](#), [DT](#), [highcharter](#), [tidyr](#), [SnowballC](#), [ldatuning](#), [topicmodels](#), [textmineR](#), [chinese.misc](#), [stringr](#), [htmlwidgets](#), [ggrepel](#), [textplot](#), [glasso](#), [ggraph](#), [Matrix](#), [utils](#), [factoextra](#), [quanteda](#)

Suggests: [rmarkdown](#), [knitr](#), [beep](#), [readxl](#)

Published: 2022-07-18

Author: Luis Pilacuan-Bonete  [cre, aut], Purificacion Galindo-Villardón  [aut], Javier De La Hoz Maestre  [aut], Francisco Javier Delgado-Álvarez  [aut]

Maintainer: Luis Pilacuan-Bonete <luis.pilacuanb@ug.edu.ec>

License: [GPL-3](#)

NeedsCompilation: no

Materials: [README](#)

CRAN checks: [LDABiplots results](#)

Documentation:

Reference manual: [LDABiplots.pdf](#)

Vignettes: [Tutorial_LDABiplots_English](#)
[Tutorial_LDABiplots_Spanish](#)

Downloads:

Package source: [LDABiplots_0.1.2.tar.gz](#)

Windows binaries: r-devel: [LDABiplots_0.1.2.zip](#), r-release: [LDABiplots_0.1.2.zip](#), r-oldrel: [LDABiplots_0.1.2.zip](#)

macOS binaries: r-release (arm64): [LDABiplots_0.1.2.tgz](#), r-oldrel (arm64): [LDABiplots_0.1.2.tgz](#), r-release (x86_64): [LDABiplots_0.1.2.tgz](#), r-oldrel (x86_64): [LDABiplots_0.1.2.tgz](#)

Figura 6.1. LDABiplots publicado en CRAN de R

La aplicación cuenta con soporte en español e inglés para uso de los investigadores, donde se explica con detalle su uso. Los tutoriales disponibles pueden ser consultados desde el CRAN, a través de los siguientes enlaces:

- Español: https://cran.r-project.org/web/packages/LDABiplots/vignettes/Tutorial_LDABiplots_Spanish.html
- Inglés: https://cran.r-project.org/web/packages/LDABiplots/vignettes/Tutorial_LDABiplots_English.html

6.2.1. DESCRIPCIÓN DEL LDABILOTS

El programa LDABiplots se basa en enfoques de modelado de tópicos y representaciones Biplot. Las principales contribuciones de este software son: a) proponer combinación de métodos y paquetes usados para el modelado de tópicos y visualizaciones de los Biplot, b) incorporar una transformación de la matriz theta del LDA para las representaciones Biplot, y c) incorporar un léxico PESTEL en la representación HJ-Biplot de la matriz phi para relacionar el macroentorno externo con los tópicos obtenidos.

Basado en las premisas desarrolladas y revisadas en capítulos anteriores, y en la metodología detallada en el punto 5.2 de esta tesis, se va a considerar el proceso en seis pasos: (1) Obtención de la información, (2) Preprocesamiento del corpus, (3) Inferencia de tópicos, (4) Obtención del Modelo basado en LDA, (5) Representaciones Biplot, (6) representación HJ-Biplot del entorno PESTEL.

6.2.1.1. OBTENCIÓN DE INFORMACIÓN

La herramienta desarrollada permite dos formas de obtener la información para ser preprocesada: a) El raspado web, también conocido como web scraping, y b) A partir de una data estructurada en archivo xlsx.

a) Web scraping.

La obtención de datos por medio del raspado web es una técnica utilizada para extraer información de sitios en la WWW de manera automática. Implica enviar solicitudes a un servidor de páginas web y analizar el código de lenguaje de marcado de hipertexto (Hyper Text Markup Language, HTML) retornado, este permite generar una serie de etiquetas o tags

que permiten definir la estructura y el formato de los elementos en una página web. Así mismo, se permite analizar el lenguaje de marcado extensible (Xtensible Markup Language, XML), que permite crear etiquetas personalizadas para definir elementos y atributos específicos relacionados con el contenido que se está representando (Markov & Larose, 2007).

LDABiplots, se centra en la extracción web de la página www.google.com, mediante la extracción de noticias web. Se basa en la estructura del análisis de nodos HTML de la página de Google, usando como base para la extracción el paquete Rvest (Wickham Hadley, 2019), y RCrawler (Khalil & Fakir, 2017), las cuales, a partir de las URL de Google, navegan en la web extrayendo la información de acuerdo con lo requerido por el usuario.

Dado que Google presenta la información paginada, es decir en diferentes páginas web, el código del paquete permite la extracción considerando esta característica. La información extraída es almacenada en formato separado por comas (Comma Separated Values, CSV) para el posterior preprocesamiento del texto adquirido en la web de noticias.

```
##-----scraping-----
observeEvent(input$runsearch, {
  req(input$search)
  if(input$language == "spanish"){
    url1 <- URLencode(paste0("https://www.google.com/search?q=",
                             gsub(" ", "+", input$search),
                             "&hl=es-419&tbm=nws&start="))

    news_array <- c()
    diarios_array <- c()
    for(i in 0:input$numberpag){
      page <- i * 10
      new_url <- paste(url1, toString(page), sep="")
      reddit_wbpg <- read_html(new_url)
      (diario<-reddit_wbpg %>%
        html_nodes('.BNeawe.UPmit') %>%
        html_text())

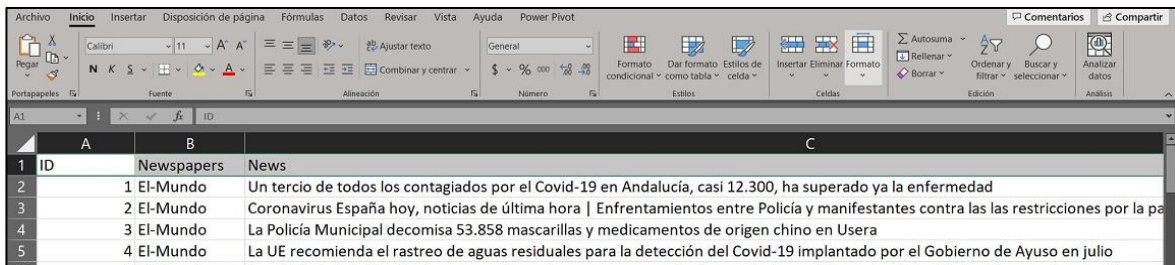
      (Noticias<-reddit_wbpg %>%
        html_nodes('.zBAuLc.l97dzf') %>%
        html_text())
      news_array <- append(news_array, Noticias)
      diarios_array <- append(diarios_array, diario)
    }
  }
}
```

Figura 6.2. Código de Webscraping del LDABiplots.

b) CARGA DE DATOS DESDE ARCHIVO

Esta opción del paquete nos permite cargar desde un repositorio local del usuario, archivos en formato xlsx, para lo cual se utiliza el paquete readxl (Wickham, Bryan, et al., 2023). Para el uso de esta opción, los datos deben estar estructurados como una matriz de tres columnas, las cuales se requieren, tanto para el análisis en español como en inglés, en el siguiente formato:

- Columna 1: ID, código de identificación de los documentos.
- Columna 2: Newspapers, el nombre de los diarios de noticias, también se puede aplicar para el nombre de revistas, u otras, de acuerdo con los datos de esta columna se pueden generar agrupaciones, por lo cual es importante que esta columna contenga características comunes de los documentos a analizar.
- Columna 3: News, es donde se encuentra contenido el texto a ser analizado por cada uno de los documentos.



	A	B	C
1	ID	Newspapers	News
2	1	El-Mundo	Un tercio de todos los contagiados por el Covid-19 en Andalucía, casi 12.300, ha superado ya la enfermedad
3	2	El-Mundo	Coronavirus España hoy, noticias de última hora Enfrentamientos entre Policía y manifestantes contra las las restricciones por la pa
4	3	El-Mundo	La Policía Municipal decomisa 53.858 mascarillas y medicamentos de origen chino en Usera
5	4	El-Mundo	La UE recomienda el rastreo de aguas residuales para la detección del Covid-19 implantado por el Gobierno de Ayuso en julio

Figura 6.3. Estructura de formato xlsx a Importar

6.2.1.2. PREPROCESAMIENTO DEL CORPUS

Una vez obtenidos los datos, los procesos descritos en el capítulo uno, apartado 1.3.3 y 1.4, nos permiten obtener la matriz léxica a ser procesada. Dada estas premisas, y a partir de la obtención del texto, bien sea de un archivo o de la web de noticias de Google, el programa permite realizar una serie de transformaciones y limpieza en los datos textuales, antes de generar el modelado de tópicos, Estos pasos son:

- **Limpieza de datos**, permite eliminar espacios en blanco en el texto, caracteres especiales, acentos y signos de puntuación. También nos ayuda a convertir el texto a minúscula y/o quitar documentos duplicados. El paquete base que nos permite realizar esta tarea es dplyr, que permite la manipulación de objetos gramaticales (Wickham, Francois, et al., 2023).

- **Tokenización**, consiste en dividir el texto en unidades más pequeñas (Liu & Curran, 2006), conocidas como token, que se pueden construir de acuerdo con el n-grama seleccionado (Sidorov, 2019). El programa nos permite la selección de Uni-gramas, Bi-gramas o Tri-gramas.
- **Remover números**, se ha considerado permitir esta opción al usuario de eliminar los números del texto, ya que no siempre los números contienen información relevante que explique el texto.
- **Remover palabras Vacías**, este proceso permite eliminar palabras que no aporten significado en la investigación, para lo cual usaremos el listado de stopword de Porter (Porter, 1980). Así mismo se pueden agregar palabras adicionales dentro del programa, dependiendo de las necesidades del usuario.
- **Normalización léxica**, conformado por los procesos de lematización y stemming, el programa usa los léxicos “libstemmer”, el cual usa el algoritmo de derivación de Porter para colapsar palabras en una raíz común (Porter, 1980) y, “stopWord” de los paquetes SnowballC (Bouchet-Valat, 2020) y stopwords (Benoit et al., 2021) respectivamente.
- **Sparcity**, término acuñado para referirse a los términos que aparecen en muy pocos documentos. Este proceso elimina las palabras que no han tenido mayor presencia en la mayor cantidad de documentos, para lo cual se debe cumplir con la condición:

$$df(t) > N(1 - Sparce) \quad (25)$$

Donde, $df(t)$, es la frecuencia del termino t , y N el número de términos. El valor de $Sparce$ va a depender del usuario. Su valor según Grimmer está entre 0.99 y 0.995, ya que el 0.5 % y 1% de los artículos deberían de descartarse. Si el valor es del 0.99, se tomarán solo los términos que aparecen en más del 1% de los documentos (Grimmer, 2010; Yano et al., 2012).

```
Diario<-data.frame(Diario=diarios_array)
Diario$Diario<-str_trim(Diario$Diario,side = "both")
Diario$Diario<-str_to_lower(Diario$Diario,locale = "es")
Diario$Diario<-chartr('áéíóúñ','aeioun',Diario$Diario)
Noticia<-data.frame(Noticia=news_array)
Noticia$Noticia<-str_to_lower(Noticia$Noticia,locale = "es")
Noticia$Noticia<-chartr('áéíóúñ','aeioun',Noticia$Noticia)
Data<-cbind(ID=row.names(Diario),Diario, Noticia)
values$Data <- Data %>% group_by(Noticia) %>% filter (! duplicated(Noticia))
NamesDiarios<- as.data.frame(Data$Diario)
values$out <- table(NamesDiarios)
```

Figura 6.4. Código para creación de tablas del texto extraído de la Web.

El preprocesamiento del corpus obtenido de la web, o de los archivos xlsx, nos permite obtener la bolsa de palabras para la construcción de la tabla léxica. Durante este proceso se van a eliminar algunas palabras de acuerdo con lo seleccionado por los usuarios, por lo que el mismo debe ser validado. Si bien no existe una forma científica de garantizar un procedimiento de “limpieza” idéntico al realizar una revisión exploratoria (Asmussen & Møller, 2019b).

Al finalizar este proceso se obtiene una matriz DTM, la cual nos va a servir para desarrollar los procesos posteriores. En la figura 5.5, se observa el código de R usado en LDABiplots para el preprocesamiento y obtención de la matriz.

```
.data <- dplyr::filter(values$Data, Diario %in% c(input$selectnews))
#View(.data)
diario <- .data
values$diario <- data.frame(diario =.data$Diario)
#View(values$diario)
stp <- unlist(strsplit(input$stopwords,","))
stp <- trimws(stp)
cpus <- parallel::detectCores()
ngram <- as.integer(input$ngrams)
Stemm <- trimws(input$Stemm)
odtm <- textmineR::CreateDtm(
  doc_vec = .data$Noticia,
  doc_names = .data$ID,
  ngram_window = c(1,ngram),
  lower = FALSE,
  remove_punctuation = FALSE,
  remove_numbers = FALSE,
  cpus = cpus)
if(input$checkStemming){
  dtm <- textmineR::CreateDtm(
    doc_vec = .data$Noticia,
    doc_names = .data$ID,
    ngram_window = c(1,ngram),
    stopword_vec = c(stopwords::stopwords(input$Language),letters,stp),
    lower = TRUE,
    remove_punctuation = TRUE,
    remove_numbers = input$removenumber,
    stem_lemma_function = function(x) SnowballC::wordStem(x, Stemm),
    cpus = cpus)}
else{dtm <- textmineR::CreateDtm(
  doc_vec = .data$Noticia,
  doc_names = .data$ID,
  lower = TRUE,
  stopword_vec = c(stopwords::stopwords(input$Language),letters,stp),
  ngram_window = c(1,ngram),
  remove_punctuation = TRUE,
  remove_numbers = input$removenumber,
  cpus = cpus)
}
```

Figura 6.5. Código de preprocesamiento del LDABiplots

6.2.1.3. INFERENCIA DE TÓPICOS

El modelado de tópicos LDA requiere que se preasigne un valor a priori de K (número de temas o tópicos). Se deberá tener en cuenta que un valor de K pequeño, cercano a 1, puede generar tópicos amplios y heterogéneos y un valor alto producirá tópicos demasiado específicos (Sbalchiero & Eder, 2020). Existen varias métricas para establecer el valor de K , como el de inferencia variacional, usado por Blei en su propuesta original (D. M. Blei et al., 2003); o el de coherencia, que se basa en la hipótesis de distribución conjunta que establece que las palabras con interpretaciones similares tienden a coexistir en contextos similares (Cao et al., 2009; Mimno et al., 2011). De acuerdo con Mimno, la coherencia del tópico está dada por:

$$\mathbf{C}(\tau, V^\tau) = \sum_{m=2}^M \sum_{l=1}^{m-1} \log \frac{D(t_m^\tau, t_l^\tau) + 1}{df(t_l^\tau)} \quad (26)$$

Donde, $df(t_l^\tau)$ es la frecuencia de documento de la palabra t , $D(t_m^\tau, t_l^\tau)$ la frecuencia de co-documento de las palabras t y t^* y, $\mathbf{C}(\tau, V^\tau)$ mide la coherencia del tema τ como la suma de la similitud de distribución por pares para las M palabras mas probables en el tópico.

Estos métodos se consideran de evaluación intrínseca, ya que para seleccionar el valor de K usan medidas internas basadas en las características de los datos. En resumen, no existe un proceso único y definitivo para determinar el valor óptimo de K , por lo que en ciertas situaciones se requiere del juicio de un experto que se ajuste a los objetivos y características del análisis.

LDABiplots incorpora el método intrínseco de coherencia, incorporado en el paquete TextmineR. Este proceso consiste en: (a) Entrenar el modelo con diferentes valores de K . (b) Extraer las palabras más relevantes de cada K generado. (c) Calcular la coherencia del tema midiendo la similitud semántica entre las palabras claves dentro de cada uno (Collins & Loftus, 1975). (d) Calcular el promedio de coherencia de cada tema. (e) Seleccionar el K con la coherencia más elevada (Jones Tommy & Doane William, 2019).

Este proceso es computacionalmente complejo, por lo que se deben seleccionar parámetros

a priori para generar la inferencia, tales como, el rango de temas en el que se desea evaluar el modelo, el número de interacciones que se requieren y el parámetro de parada del modelo. Para las interacciones del modelo se usa el muestreo de Gibbs, que tiene como objetivo generar una cadena de Márkov que tenga la distribución posterior objetivo como su distribución estacionaria. Es decir, después de varias iteraciones de recorrer la cadena, el muestreo de la distribución debería de converger para estar cerca del muestreo posterior deseado (Steyvers & Griffiths, 2007). No hay forma de saber cuántas iteraciones se requieren para alcanzar la distribución deseada, sin embargo, en la practica el muestro de Gibbs es bastante potente computacionalmente y presenta un mejor rendimiento (Darling, 2011).

Para la generación de la inferencia y del modelo LDA, se debe ingresar el valor del hyperparámetro α , teniendo en cuenta que un valor alto genera documentos con una mayor densidad temática, lo que significa que un documento contenga una combinación del mayor número de tópicos; mientras que un valor bajo, conduce a documentos más esparcidos, donde cada documento contenga unos pocos tópicos dominantes, o incluso solo uno.(Griffiths & Steyvers, 2004). Varios autores han definido algunas reglas para determinar el valor de alfa, como:

- Griffiths, que usa un valor de $\alpha = (0.1, \frac{50}{K})$ (Griffiths & Steyvers, 2004).
- Asunción, propuso un $\alpha = (0.1, 0.1)$ (Asunción et al., 2012).
- Rehůřek and Sojka, indicaron que, para corpus muy extenso es conveniente un $\alpha = (\frac{1}{K}, \frac{1}{K})$ (Řehůřek & Sojka, 2010).

Por defecto el paquete usa un valor de 0.1 para el cálculo de la inferencia, pero permite al usuario colocar un valor, dadas las premisas consideradas anteriormente. Para el cálculo del proceso de inferencia se usa el paquete textmineR (T. Jones et al., 2021).

```

observeEvent(input$Run.model1,{

  ptm <- proc.time()
  stpCohe <- unlist(strsplit(input$OtherKCoherence,","))
  stpCohe <- as.numeric(trimws(stpCohe))
  seqk <- c(seq(from=input$num1,to=input$num2,by=input$num3),stpCohe)# Candidate number of topics k
  iterations <- input$num4 # Parameters control Gibbs sampling
  burnin <- input$num5 # Parameters control Gibbs sampling
  alpha <- input$num6 # Parameters control
  cores <- parallel::detectCores()
  #dtm <- values$dtmF
  if(input$load == "import"){
    dtm <- values$dtmFxl }
  else if(input$load == "load"){
    dtm <- values$dtmF }

  values$coherence_list <- textmineR::TmParallelApply(X = seqk ,
                                                    FUN = function(k){

    m <- textmineR::FitLdaModel(dtm= dtm ,
                               k = k,|
                               iterations =iterations ,
                               burnin = burnin,
                               alpha = alpha,
                               beta = colSums(dtm) / sum(dtm) * 100,
                               optimize_alpha = TRUE,
                               calc_likelihood = TRUE,
                               calc_coherence = TRUE,
                               calc_r2 = FALSE,
                               cpus = cores)

    m$k <- k
    m
  },export= ls(), # c("nih_sample_dtm"), # export only needed for Windows machines
  cpus = cores)

```

Figura 6.6. Sección de Código para la Inferencia de K.

6.2.1.4. OBTENCIÓN DEL MODELO LDA

Una vez conocido el valor del número de tópicos K, con la función FitLDAModel del paquete TextmineR de R (T. Jones et al., 2021), se genera el modelo LDA. El proceso parte de la bolsa de palabras de cada documento, es decir, no se tiene en cuenta el orden en que aparecen las palabras, sino la frecuencia con las que aparecen en el documento.

A partir del modelo generativo, revisado en el punto 2.4.2.2, y considerando la terminología explicada, se procede a obtener el modelo. Para lo cual, se requiere conocer el valor de ciertos parámetros, tales como: el valor del número de tópicos a partir de la coherencia del análisis de inferencia previo revisado en el punto anterior (k), los valores de las iteraciones y el periodo de quemado (Burn-it) de la inferencia basado en el muestreo de Gibbs, el valor del hiperparámetro alpha.

De la ecuación 18 y de las probabilidades detalladas en la misma, y basados en que los documentos fueron generados por medio del modelo generativo, el LDA nos permite

calcular la matriz theta y phi, las cuales serán objeto de transformación y análisis de acuerdo con lo revisado en el apartado 4.3.1.

El principio general del proceso de inferencia a partir del muestreo de Gibbs (Darling, 2011; Steyvers & Griffiths, 2007), se basa en:

1. Inicialización, Se selecciona aleatoriamente un tema para cada palabra en cada documento de una distribución multinomial.

2. Muestreo de Gibbs,

a. Para i iteraciones

b. Para el documento d en documentos.

i. Para cada palabra en el documento d :

1. Asigne un tópico a la palabra actual en función de la probabilidad del este, dado el tema de todas las demás palabras mostrada en la ecuación (18).

En resumen, LDA nos permite encontrar parámetros desconocidos (latentes) en el corpus, tales como, el número de tópicos (k), la mezcla de tópicos del documento, la distribución de palabras de cada tópico, y la asignación del tema de cada palabra en cada documento para así determinar la combinación de los tópicos de cada documento.

```
k <- input$num25
iter <- input$num26
burnin <- input$num27
alpha <- input$num28

cpus <- parallel::detectCores()

if(input$load == "import"){
  dtm <- values$dtmFxl
}
else if(input$load == "load"){
  dtm <- values$dtmF
}
values$model <- textmineR::FitLdaModel(dtm = dtm, # parameter
k = k, # Number of topics k
iterations = iter, # parameter
burnin = burnin, #parameter
alpha = alpha, # parameter
beta = colSums(dtm)/sum(dtm)*100,
optimize_alpha = TRUE, # parameter
calc_likelihood = TRUE,
calc_coherence = TRUE,
calc_r2 = FALSE,
cpus = cpus)

beepr::beep(2)
remove_modal_spinner()
})
```

Figura 6.7. Código del LDABlplots para obtener el LDA del corpus.

Los tópicos requieren un etiquetado que debe ser validado por parte de un experto en el campo de la investigación sobre el cual se realiza el análisis, ya que las etiquetas dadas por el modelo LDA, podrían tener un etiquetado y resultado no válido (Asmussen & Møller, 2019c).

6.2.1.5. REPRESENTACIONES BIPLLOT

A partir de las obtenciones de la matriz theta y phi del LDA y de la transformación realizada de θ , abordados en el punto 4.3.1. LDABiplots nos permite agrupar esta matriz de acuerdo con ciertas características de los documentos, para poder ser representadas estas características comunes en una visualización Biplot.

El paquete genera visualización de los Biplot Clásicos de Gabriel (Gabriel, 1971), analizados en el apartado 3.2.1, y del HJ-Biplot tratado en el punto 3.2.2 propuesto por Galindo (Galindo-Villardón, 1986). Para lo cual se deben seleccionar ciertos parámetros que permitirán la generación de la representación deseada.

Entre los parámetros que se debe seleccionar figura el tipo de centrado y escalado de la matriz de covarianzas. Y así poder obtener los resultados del Biplot, como:

- Los Valores Propios (Eigenvalues), que representan la importancia relativa de cada eje principal en la explicación de la variabilidad de los datos originales. Un valor propio alto para un determinado eje principal indica que ese eje captura una proporción significativa de la variabilidad total de los datos.
- Proporción de Varianza explicada (Variance explained), la cual se calcula de la división de cada valor propio por la suma total de los valores propios, se refiere a la cantidad de variabilidad de los datos originalmente presente en las variables que se representan en el Biplot.

Los ejes con valores propios más altos capturan la mayor parte de la variabilidad de los datos originales y, por lo tanto, explican una proporción mayor de la varianza total (Greenacre, 2012). Adicionalmente el paquete nos da las coordenadas de las variables y de los individuos.

Para la generación de los Biplot, a partir de lo enunciado, se generaron las funciones en código de R de cada una de las visualizaciones en el LDABiplots, observadas en las siguientes graficas.

```
HJBiplot <- function(X, Transform.Data = 'scale'){  
  # List of objects that the function returns  
  hjb <-  
    list(  
      eigenvalues = NULL,  
      explvar = NULL,  
      loadings = NULL,  
      coord_ind = NULL,  
      coord_var = NULL  
    )  
}
```

Figura 6.8. Función HJ-Biplot del LDABiplots

```
GHBiplot <- function(X, Transform.Data = 'scale'){  
  # List of objects that the function returns  
  ghb <-  
    list(  
      eigenvalues = NULL,  
      explvar = NULL,  
      loadings = NULL,  
      coord_ind = NULL,  
      coord_var = NULL  
    )  
}
```

Figura 6.9. Función GH-Biplot del LDABiplots

```
JKBiplot <- function(X, Transform.Data = 'scale'){  
  # List of objects that the function returns  
  jkb <-  
    list(  
      eigenvalues = NULL,  
      explvar = NULL,  
      loadings = NULL,  
      coord_ind = NULL,  
      coord_var = NULL  
    )  
}
```

Figura 6.10. Función JK-Biplot del LDABiplots


```

bip <- dplyr::select(bip, -"topic")
row.names(bip) <- names
if (input$selectypebiplot == 'HJ_Biplot') {
  values$Biplot <- HJBiplot(bip, Transform.Data = biptr())
}
else if (input$selectypebiplot == 'JK_Biplot') {
  values$Biplot <- JKBiplot(bip, Transform.Data = biptr())
}
else if (input$selectypebiplot == 'GH_Biplot') {
  values$Biplot <- GHBiplot(bip, Transform.Data = biptr())
}
})

output$eigen <- DT::renderDT({
  req(values$Biplot)
  dat <- values$Biplot
  dat <- data.frame(dat$eigenvalues)
  DT::datatable(data = dat,
    colnames=c("value"),
    extensions = 'Buttons',
    options = list(dom = 'Bfrtip',
      buttons = c('pageLength',
        'copy',
        'csv',
        'excel',
        'pdf',
        'print'),
      pagelength = 5,
      #lengthMenu = list(c(5,10,20000,-1),
        c('5', '10', 'All'))
    )
})

```

Figura 6.11. Sección de Código de generación de Biplot del paquete.

6.2.1.6. REPRESENTACIÓN HJ-BIPLLOT DEL ENTORNO PESTEL

La finalidad de representar en un HJ-Biplot los términos que tengan relación con el macroentorno del ambiente, es la de evaluar como una organización, sector u objeto de análisis puede verse afectado por el entorno externo. Fahey (Fahey & Narayanan, 1968) detalló seis categorías revisadas en el capítulo 4.3 de este proyecto.

Para este objetivo, se va a usar la técnica de diccionarios lexicográficos, que recopila y presenta de manera sistemática las palabras o términos de un idioma, y que pueden proporcionar información detallada sobre el significado, uso, etimología y otras características. Un ejemplo de estos son los diccionarios lexicográficos usados para el análisis de sentimiento, el cual categoriza palabras como positivas, negativas o neutras (Ramesh et al., 2015).

Tomando en consideración lo mencionado, LDABiplots genera agrupaciones mediante 6 clústeres (uno para cada factor). Este proceso se realizó con el uso de un léxico que identifica palabras correspondientes a cada uno de los factores del PESTEL. El listado se ha formado con 16166 términos inherentes a los factores detallados por Fahey y Perera (Fahey & Narayanan, 1968; Perera, 2020), tomados de diferentes diccionarios de índices, como:

- **Índice Político**, tomado del Diccionario Constitucional Chileno (García et al., 2016), y del Índice Político del Sistema Argentino de Información Jurídica (SAIJ, 2016).
- **Índice Económico**, términos recogidos del Diccionario de términos económicos (Sepúlveda, 2004)
- **Índice Social**, listado adaptado del Diccionario de trabajo social (Ander-Egg, 2000), Diccionario de urbanismo y ordenación (Zoido N et al., 2000) y diccionario de términos de Psicología (Cosacov, 2007).
- **Índice Tecnológico**, basado en los términos del Diccionario técnico de Beigbeder (Beigbeder, 2006) y del manual de términos tecnológicos de Sánchez (Sánchez et al., 2018).
- **Índice Ambiental**, adaptado del diccionario de términos ambientales de Camacho (Camacho B & Ariosa R, 2000)
- **Índice Legal**, términos recopilados del diccionario Jurídico elemental de Cabanelas (Cabanelas, 2003) y del Diccionario jurídico de Guillien (Guillien & Vincent, 2021).

Algunos términos fueron tomados del glosario de términos financieros, contables, administrativos, económicos y legales de Vidales (Vidales, 2003). El léxico generado PESTEL, se puede encontrar dentro del propio paquete de R, o en el siguiente enlace de GitHub:

https://github.com/Pilacuan-Bonete-Luis/lexico_PESTEL/blob/main/PESTEL

	Palabra	Puntuacion	Word
1	absolutismo	political	absolutism
2	abstencion	political	abstention
3	abstencionactiva	political	abstentionactive
4	abstencionismo	political	abstentionism
5	accion	political	action
6	accion colectiva	political	collective action
7	accion directa	political	direct action
8	accion politica	political	political action
9	acracia	political	acracia
10	actitud politica	political	political attitude
11	activismo	political	activism
12	activismo franquiciado	political	franchisee activism
13	actuacion politica	political	politic performance
14	acuerdo politico	political	political agreement
15	aculturacion politica	political	politic acculturation
16	adelanto	political	advancement
17	adhocracia	political	adhocracy

Showing 1 to 18 of 16,166 entries, 3 total columns

Figura 6.12. Sección del Léxico PESTEL

LDABiplots, a partir del léxico PESTEL anterior, crea un listado de factores que serán la base de la identificación de cada palabra contenida en los tópicos generados. La matriz phi será representada por medio de un HJ-Biplot, añadiendo como clústeres el léxico de factores externos, permitiendo observar las relaciones que tienen estas palabras con los tópicos generados y las contribuciones de cada factor PESTEL con los diferentes temas analizados.

```
##### Creando dataframe con los cluster de PESTEL#####
## a partir de la data phi original

word_phi <- as.data.frame(phi[,1])
names(word_phi)[1]="Palabra"

pestel_word<- word_phi %>%
  inner_join(pestel, ., by = "Palabra")

Cluster_pestel<-left_join(word_phi,pestel_word, by="Palabra")
Cluster_pestel<-Cluster_pestel%>% distinct(Palabra, .keep_all = T)

##### Plot Cluster PESTEL Biplot#####
#Paquete Factorextra, genera graficos a partir de PCA (tipo prcomp)

fviz_pca_biplot(hj2, label="var", habillage=as.factor(Cluster_pestel$Puntuacion)) +
  labs(color=NULL) + ggtitle("PESTEL") +
  theme(text = element_text(size = 17),
        panel.background = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(colour = "black"),
        legend.key = element_rect(fill = "white"))
```

Figura 6.13. Código de generación de clúster PESTEL en el HJ-Biplot

CAPÍTULO VII

APLICACIÓN PRÁCTICA LDABILOTS

7.1. INSTALACIÓN DEL PAQUETE LDABILOTS

El entorno de desarrollo integrado (IDE) utilizado principalmente para programar en el lenguaje de R, denominado RStudio, proporciona un conjunto de herramientas diseñadas específicamente para facilitar la escritura, depuración y ejecución de códigos (Posit, 2023). En este entorno se debe de instalar el paquete LDABiplots, escribiendo el código de la figura (7.1).

```
1 install.packages("LDABiplots")
2 library(LDABiplots)
3 runLDABiplots()
```

Figura 7.1. Códigos de Instalación de LDABiplots en RStudio.

Al ejecutar estos comandos se mostrará la pantalla principal del LDABiplots, en la cual se presenta el menú siguiente. También se encuentran unas banderas que permiten seleccionar el idioma deseado, para que el usuario pueda leer la introducción del paquete.

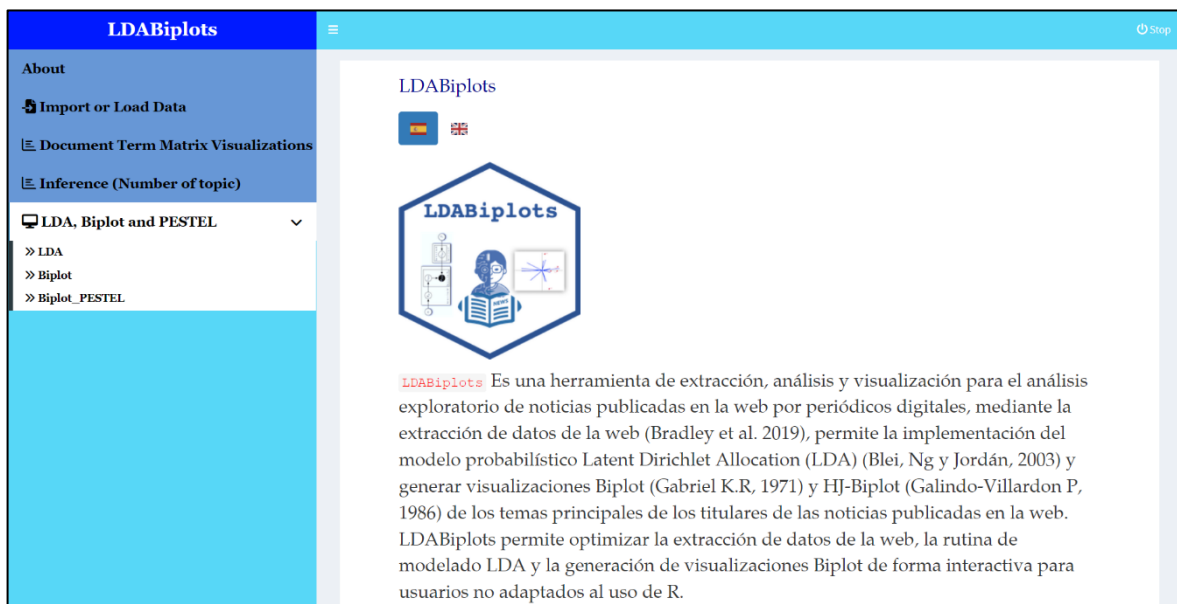


Figura 7.2. Pantalla de Menú del LDABiplots

7.2. MENÚ IMPORTACIÓN DE DATOS

Para mostrar el funcionamiento de la metodología anteriormente expuesta, se procederá a procesar los datos extraídos del 1 de enero del 2019 al 27 de mayo del 2022, con los

términos de búsqueda “Covid”, “Coronavirus”, de los tres diarios de mayor circulación de España “El país”, “El Mundo” y “20minutos”. Esta extracción se realizó utilizando un raspado web de la estructura HTML de las mencionadas páginas web. El raspado se genera de las etiquetas <h2>, <h1>, <a>, entre otras, de los cuerpos <body> contenidos en los documentos <html> que componen la página web de cada uno de los diarios.

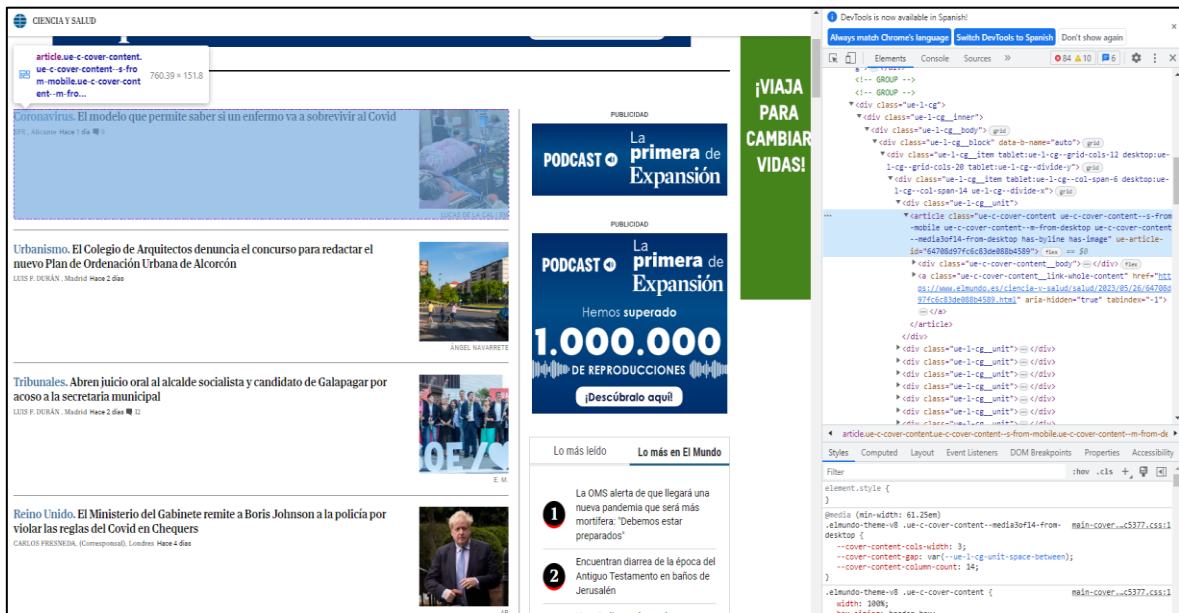


Figura 7.3. Página web El Mundo. Web Scraping de Etiqueta <a>

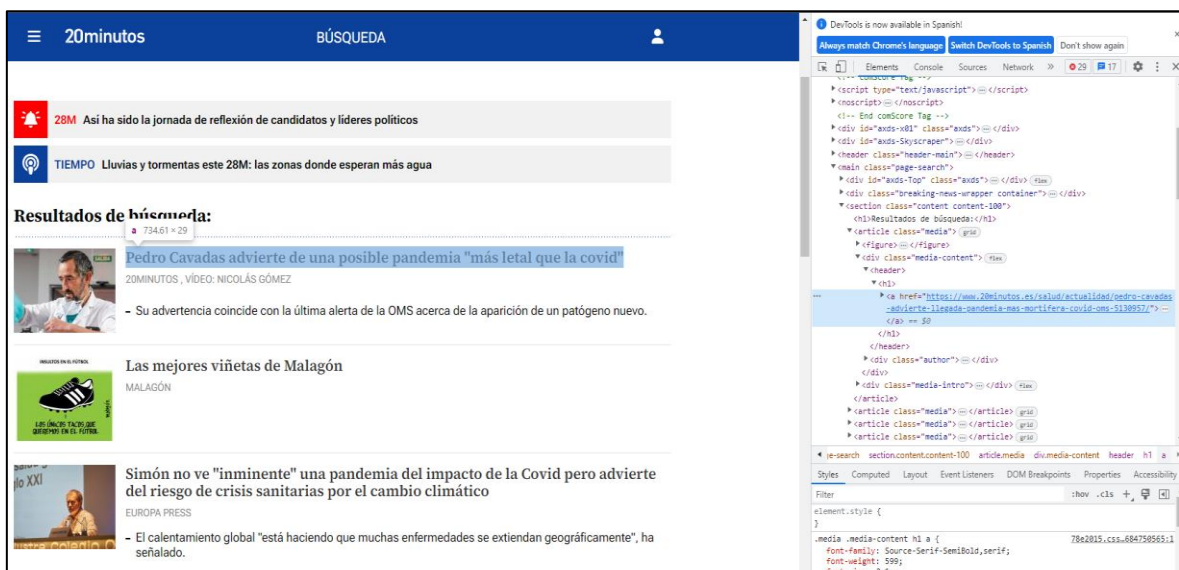


Figura 7.4. Página web 20Minutos. Web Scraping de Etiqueta <a>

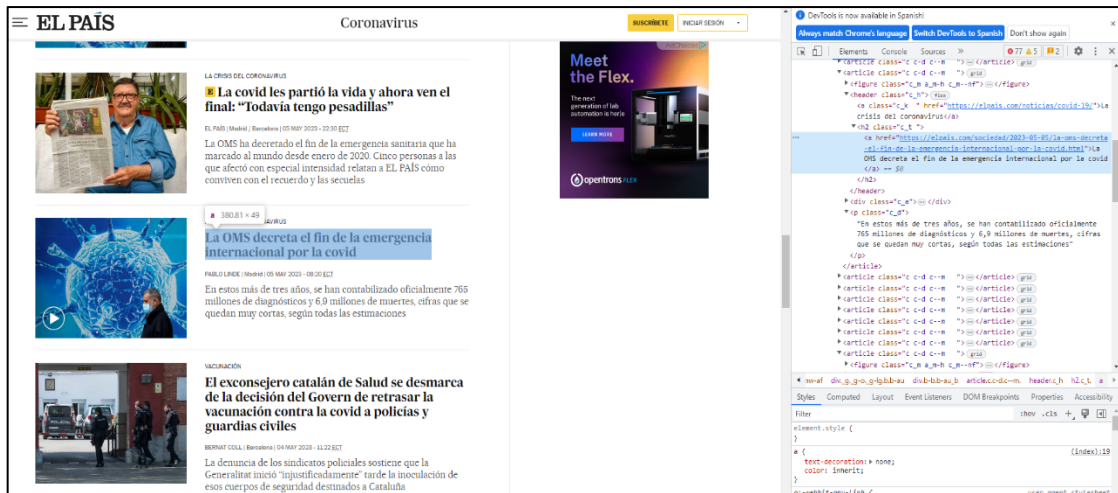


Figura 7.5. Página web El País. Web Scraping de Etiqueta <h2>

Los datos extraídos fueron preprocesados en Excel, para ser estructurados de acuerdo con la figura 6.3, y ser ingresados en la interfaz gráfica del LDABiplots. Se obtuvo así una base de las 48112 noticias extraídas.



Figura 7.6. Importación de datos desde archivo xlsx

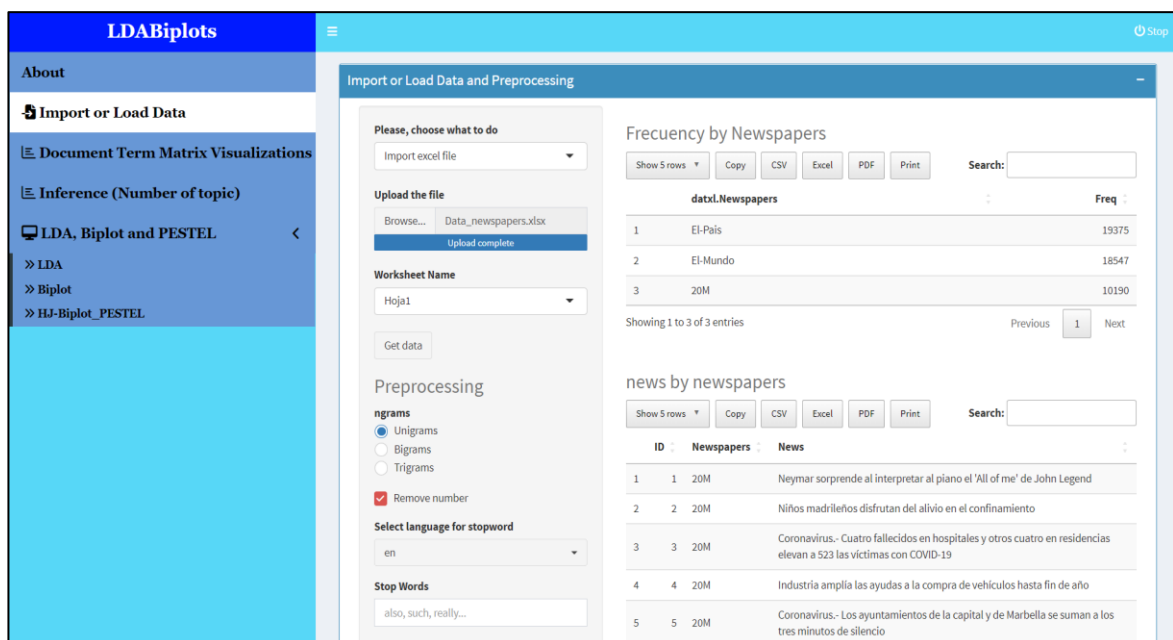


Figura 7.7. Información de Datos textuales extraídos.

Para ejecutar el procesamiento en la interfaz, se debe seleccionar el tipo de tratamiento que se le dará a los datos textuales desde el submenú Preprocesamiento, para lo cual se selecciona: la separación del texto a analizarse en unigramas, la remoción de números, el español como lenguaje para eliminar las palabras vacías, adicionando “covid” y “coronavirus” en el listado, y no se seleccionó el proceso de stemming para nuestro análisis ya que al cortar a la raíz se pierde información conectada con el léxico PESTEL, y se marcó un sparsity de 0.995. Obteniendo así una matriz DTM de 213 términos.

Figura 7.8. Parámetros para obtener DTM en el LDABiplots.

	document	term
Originalxl	48112	39258
Finalxl	48112	213

Figura 7.9. DTM obtenida en el LDABiplots.

que sea lo suficientemente grande. Para el ejemplo se han seleccionado 1000 iteraciones y un Burn-in de 5.

- Parámetro Alpha, se debe seleccionar un valor del hiperparametro Alpha, considerando lo revisado en el apartado 6.2.1.3. Por defecto LDABiplots usa el valor de 0.1 para el cálculo. Una vez seleccionados los parámetros para la inferencia se corre el modelo haciendo click en el botón Run, luego de unos minutos se observará el gráfico de coherencia de tópicos. Este proceso demanda soporte computacional, por lo que el mismo puede tardar cierto tiempo, dependiendo de las características del equipo utilizado por el usuario.

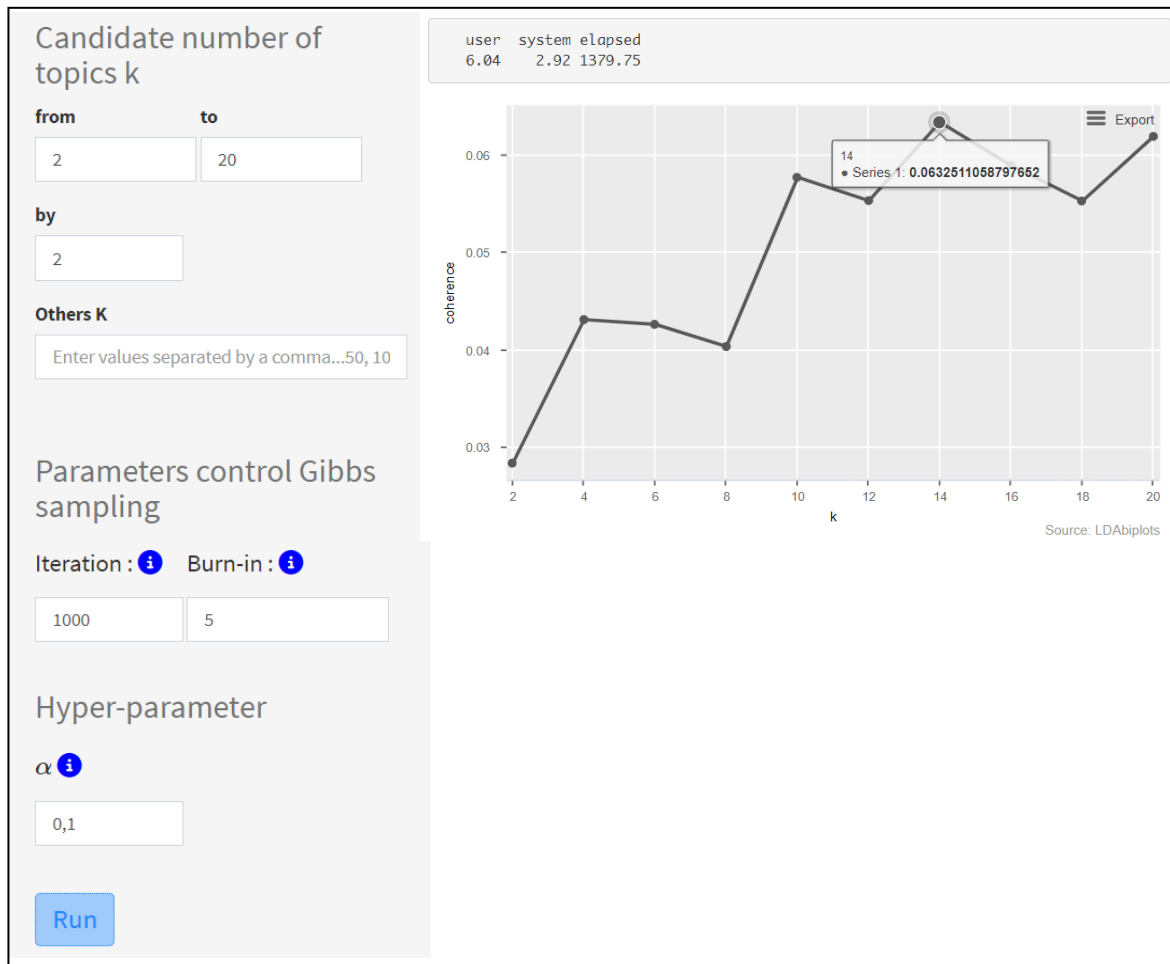


Figura 7.13. Inferencia de K para el modelo LDA en el LDABiplots

El proceso de selección del K óptimo sugerido va a depender del usuario, ya que cuando se presentan varios K con valores de coherencia muy similares, es decisión del usuario seleccionar uno. En la figura 7.13 tenemos valores de K muy cercanos, como 14 y 18. Para el ejemplo, y considerando el tamaño del corpus, seleccionamos un valor de K óptimo recomendado por el proceso de coherencia de 14 tópicos.

7.5. MENÚ LDA Y BILOT

Una vez definidos el número de temas o tópicos K en el proceso de inferencia, se procede a generar el modelo LDA y las representaciones de los resultados, para lo cual el paquete cuenta con un menú donde se van a realizar los análisis LDA, Biplot y PESTEL, como se observa en la figura 7.14.

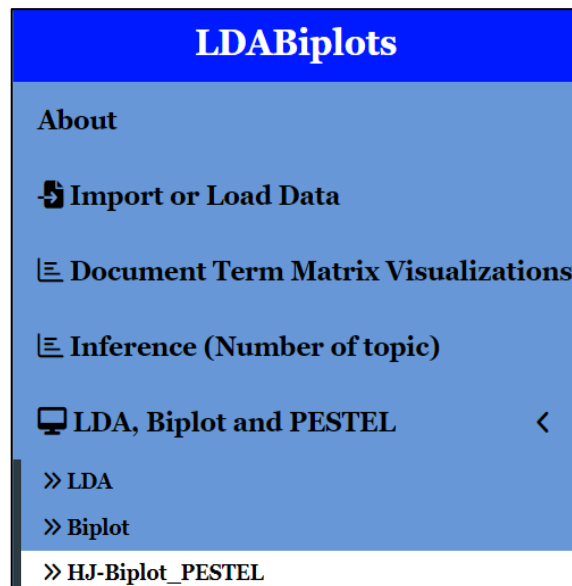


Figura 7.14. Menú de análisis LDA, Biplot, PESTEL

7.5.1. MENÚ PROCESO LDA

Partiendo de lo analizado en el apartado 6.2.1.4, se deben de seleccionar varios parámetros en el menú LDA para poder generar el modelo, los datos requeridos son:

- K recomendado, obtenido del proceso de inferencia el cual es de 14 tópicos.
- Iteraciones del Muestreo de Gibbs, al igual que el proceso de inferencia se propone un valor de 1000.
- Burn-in, que elimina las primeras muestras del LDA, ya que son muy poco probables de que estén se den (B. Zhang et al., 2016). Hemos seleccionado un Burn-in de 5 para nuestro ejemplo.
- El Valor de Alpha, para lo cual se revisan las reglas y se recomienda seleccionar un valor menor que 0.1. Siguiendo las recomendaciones de Rehůřek and Sojka (Řehůřek & Sojka, 2010) calculamos a partir de $1/14$, obteniendo un valor de 0.07.

Una vez marcados los valores de los parámetros, se debe generar el modelo haciendo clic en el botón de Run del menú, obteniendo después de varios minutos las matrices theta y phi resultantes, las cuales serán representadas con métodos clásicos y con Biplot.

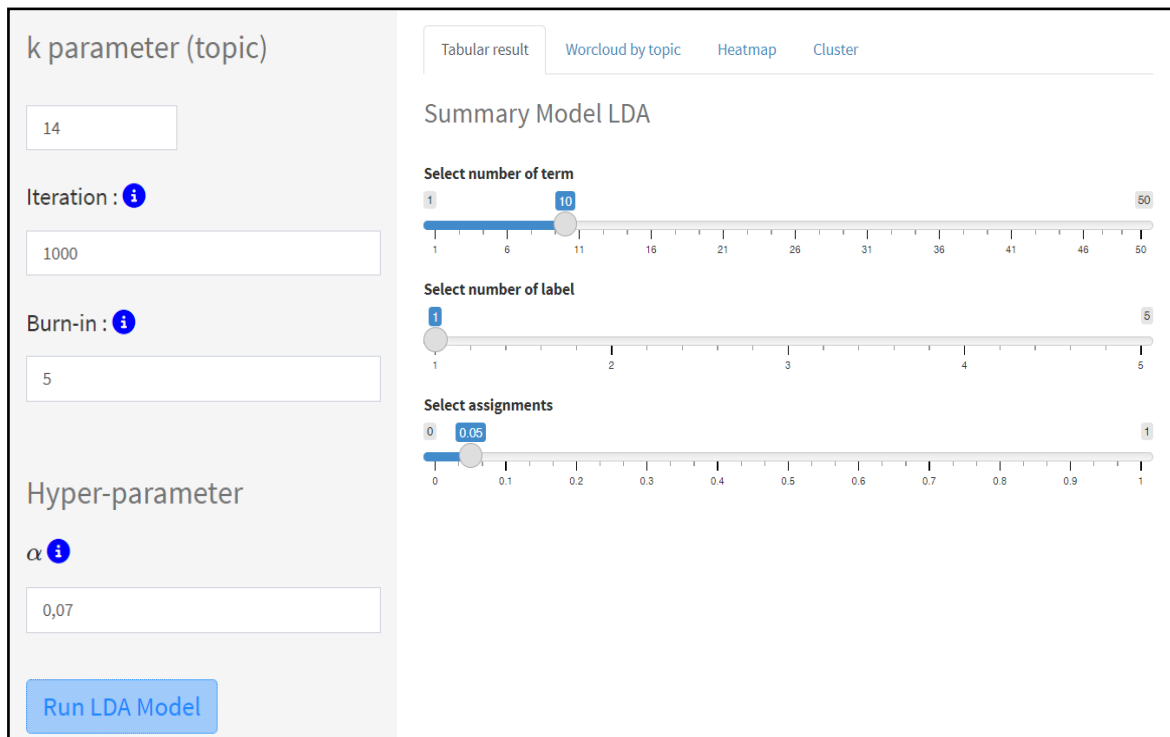


Figura 7.15. Parámetros para obtener el LDA.

Los resultados del modelo se pueden analizar en el LDABiplots de forma tabular y gráfica. Dentro de los resultados tabulares tenemos:

- Summary del Modelo, que es una tabla resumen, que muestra el tópico con su etiqueta recomendada, y los valores de coherencia y prevalencia, así como los términos de cada uno de los tópicos. Se recomienda la intervención de un experto del área considerada para el etiquetado correcto de los tópicos.

topic	label_1	coherence	prevalence	top_terms	
t_1	t_1	vacuna	0.09700	8.63200	vacuna, dosis, anos, vacunacion, tercera
t_2	t_2	pandemia	0.07900	5.96600	vacunas, vacuna, personas, pfizer, dosis
t_3	t_3	pandemia	0.03700	6.87200	dos, meses, pandemia, tres, tras
t_4	t_4	madrid	0.11000	8.82200	madrid, restricciones, queda, toque, comunidad
t_5	t_5	pandemia	0.02100	6.71100	millones, gobierno, euros, pandemia, crisis

Showing 1 to 5 of 14 entries Previous 1 2 3 Next

Figura 7.16. Matriz de Resumen del Modelo LDA.

Las etiquetas dadas por el LDABiplots son recomendadas, como se indicó anteriormente se requiere de un experto del área analizada para etiquetarlas correctamente, en la tabla 4 se aprecia la etiqueta de los tópicos propuestos en este análisis.

Tópicos	Etiquetas	Tópicos	Etiquetas
t_1	Casos Positivos	t_8	Dosis Vacunas
t_2	Nuevos Contagios	t_9	Pandemia en USA
t_3	Nueva Variante	t_10	Pandemia Mundial
t_4	Tercera Dosis	t_11	Toque de Queda
t_5	Comunidad Madrid	t_12	Gobierno Madrid
t_6	Vacunas Población	t_13	Inversión del Gobierno
t_7	Medidas Pedro Sánchez	t_14	Confinamiento en Casa

Tabla 7.1. Etiqueta de Tópicos obtenidos en el LDA

- La Matriz Theta, la cual muestra la probabilidad de que cada noticia de los periódicos analizados pertenezca a cada tema o tópico generado. La sumatoria de estas probabilidades es igual a 1. Esta matriz será la usada para las representaciones Biplot a partir de la agrupación de todos los documentos que conforman un periódico, proceso revisado en el apartado 6.2.1.6.

Theta Matrix

Show 5 rows ▾ Copy CSV Excel PDF Print

Search:

	Newspaper	ID news	Topic	theta
	All	All	All	All
1	20M	1	t_1	0.07143
2	20M	2	t_1	0.07143
3	20M	3	t_1	0.01759
4	20M	4	t_1	0.02349
5	20M	5	t_1	0.07143

Showing 1 to 5 of 673,568 entries

Previous **1** 2 3 4 5 ... 134,714 Next

Figura 7.17. Matriz Theta del LDA obtenida en el LDABiplots.

- La Matriz Phi, la cual muestra la probabilidad de que cada palabra pertenezca a cada tema o tópicos generados. La sumatoria de estas probabilidades es igual a 1. Esta matriz será usada para las representaciones PESTEL - Biplot a partir del etiquetado de las palabras con el léxico PESTEL, proceso revisado en el apartado 6.2.1.6.

Phi Matrix

Show 5 rows ▾ Copy CSV Excel PDF Print

Search:

	topic	term	phi
	All	All	All
1	1	alumnos	0.00018
2	2	alumnos	0.00004
3	3	alumnos	0.00661
4	4	alumnos	0.00009
5	5	alumnos	0.00018

Showing 1 to 5 of 2,982 entries

Previous **1** 2 3 4 5 ... 597 Next

Figura 7.18. Matriz PHI del LDA obtenida en el LDABiplots.

Así mismo, el paquete nos permite obtener representaciones graficas clásicas, entre las que se han incorporado:

- Nube de Palabras por tópicos (Wordcloud by topic), el cual permite observar la probabilidad de las palabras con respecto al tópico de acuerdo con el tamaño de estas, en un gráfico del tipo nube de palabras. Se debe seleccionar el tópico deseado.



Figura 7.19. Nube de Palabras por tópico.

- Mapa de Calor (Heatmap), el cual permite representar el valor medio de probabilidad de todas las noticias que pertenecen a cada diario. Este valor medio obtenido a partir de la matriz theta, es la probabilidad media de que cada diario contribuya a cada uno de los tópicos. En el grafico los colores más fuertes cercanos al rojo indican que los valores de probabilidad de los diarios en ese tópico son más altos, al contrario que para los valores más fuertes cercanos al azul, donde los valores de probabilidad de que las noticias de un diario pertenezcan a ese tópico son bajos. Se puede observar como el diario de noticias 20M tienen una alta probabilidad en relación al tópico 7 (medidas Pedro Sánchez) y 10 (pandemia Mundial), y se puede ver como el mismo diario tiene valores bajos con respecto al tópico 8 (dosis vacunas).

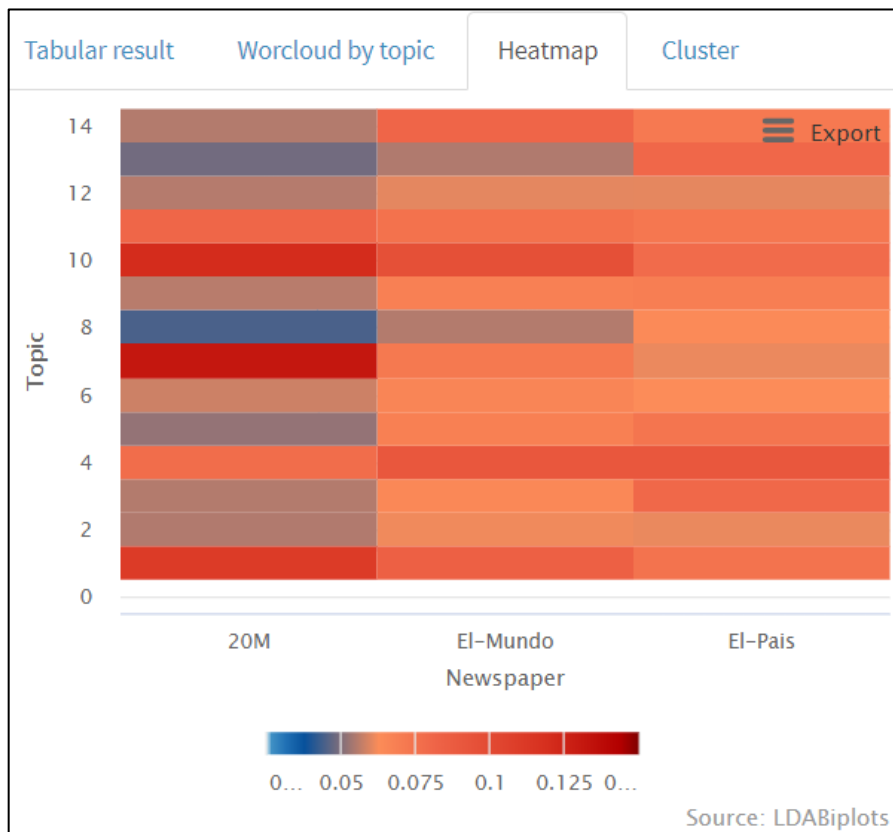


Figura 7.20. Mapa de Calor de la Probabilidad de que los Diarios pertenezcan a cada t3pico

- Cl3ster, en esta pestaña se pueden observar graficas de los t3picos agrupados por su similitud, para esto se debe seleccionar:
 - o El m3todo de agrupaci3n (Agglomeration Method), entre los que tenemos: Complete, el cual calcula la distancia m3xima entre los puntos m3s alejados de dos grupos para medir la similitud entre ellos (El-Hamdouchi & Willett, 1989). Single, que calcula la distancia m3nima entre los puntos m3s cercanos de dos grupos para medir su similitud (El-Hamdouchi & Willett, 1989). Ward.D, que calcula la varianza entre los grupos para minimizar la

varianza total dentro de todos los grupos combinados (Ward & Hook, 1963). Ward.D2, es una variante del ward.D que utiliza la suma de los cuadrados de las diferencias para calcular la varianza y medir la similitud entre grupos (Murtagh & Legendre, 2014). Average, el cual calcula la distancia promedio entre todos los puntos de dos grupos para medir su similitud (Sokal, 1963). Mcquitty; este utiliza un enfoque de ponderación basado en la diferencia de tamaño entre los grupos, donde los grupos más pequeños tienen un mayor peso en el cálculo de la similitud entre grupos (Mcquitty, 1966). Median, calcula la distancia entre las medianas de dos grupos para medir su similitud (Bradley et al., 1996; Sarle et al., 1991). Centroid, se basa en el cálculo de la distancia entre los centroides de dos grupos para medir su similitud (Müllner, 2011; Sarle et al., 1991).

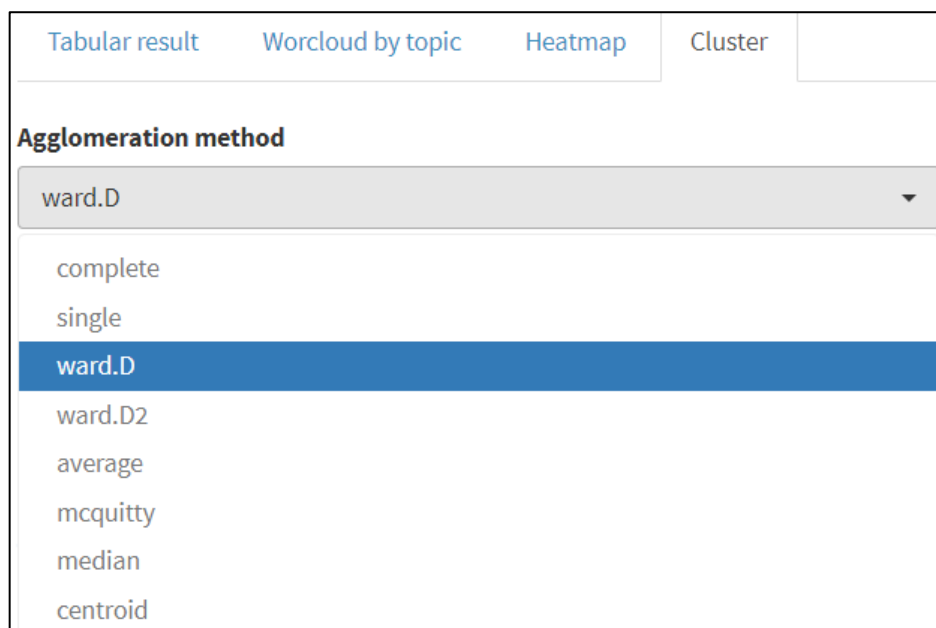


Figura 7.21. Selección de Método de distancia para los clústeres de los tópicos del LDA.

- El número de clústers a generar.
- El tipo de representación de los clústeres, entre las opciones que ofrece el paquete tenemos: Rectangular (Rectangle), que muestra los grupos o clústeres en forma de rectángulos, donde cada rectángulo representa un clúster y su tamaño puede ser proporcional al número de elementos en el mismo (Jain et al., 1999). Circular, donde cada clúster se representa como una sección del círculo y las relaciones de jerarquía se visualizan mediante enlaces que conectan los clústeres en función de su similitud o distancia (Rohlf, 1970). Filogénico (Phylogenic), donde los clústeres se representan como nodos de un árbol, y las relaciones de similitud o distancia entre los clústeres se visualizan mediante las ramas del árbol (Saitou & Nei, 1987).

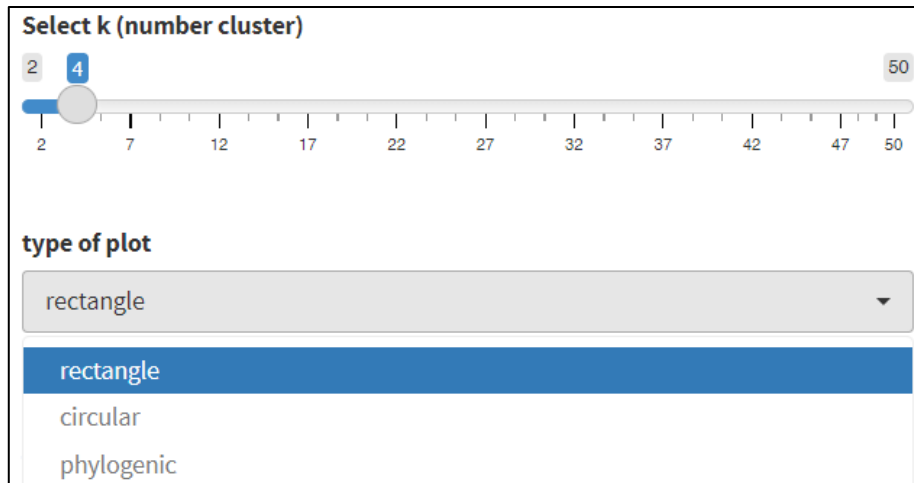


Figura 7.22. Menú de selección de tipo de Clúster.

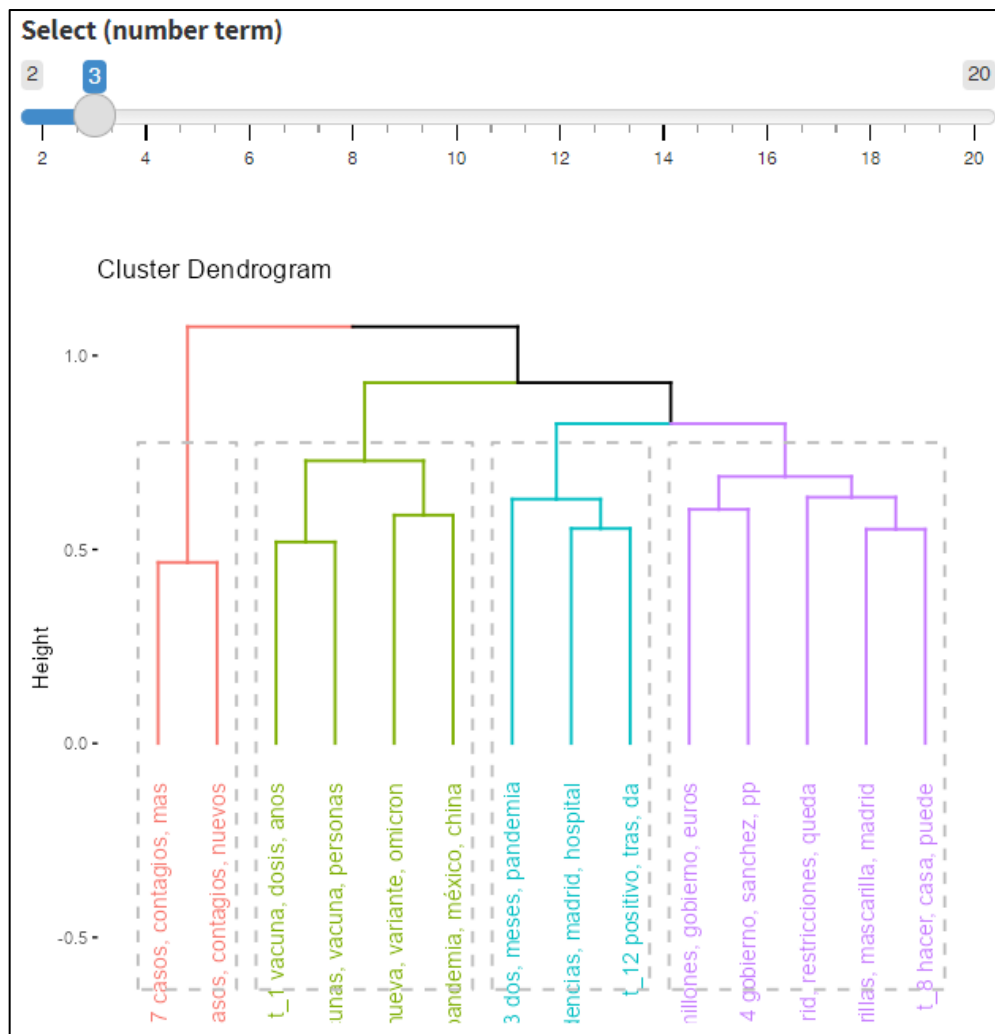


Figura 7.23. Clúster rectangular de los tópicos del LDA.

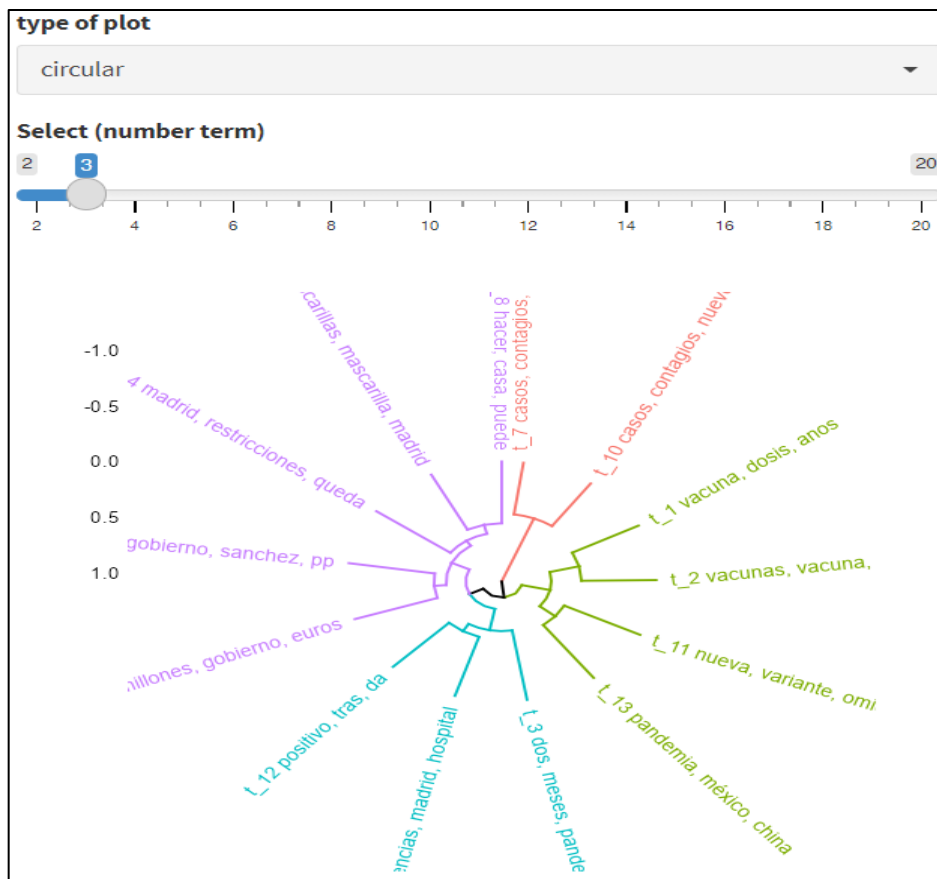


Figura 7.24. Clúster circular de los tópicos del LDA.

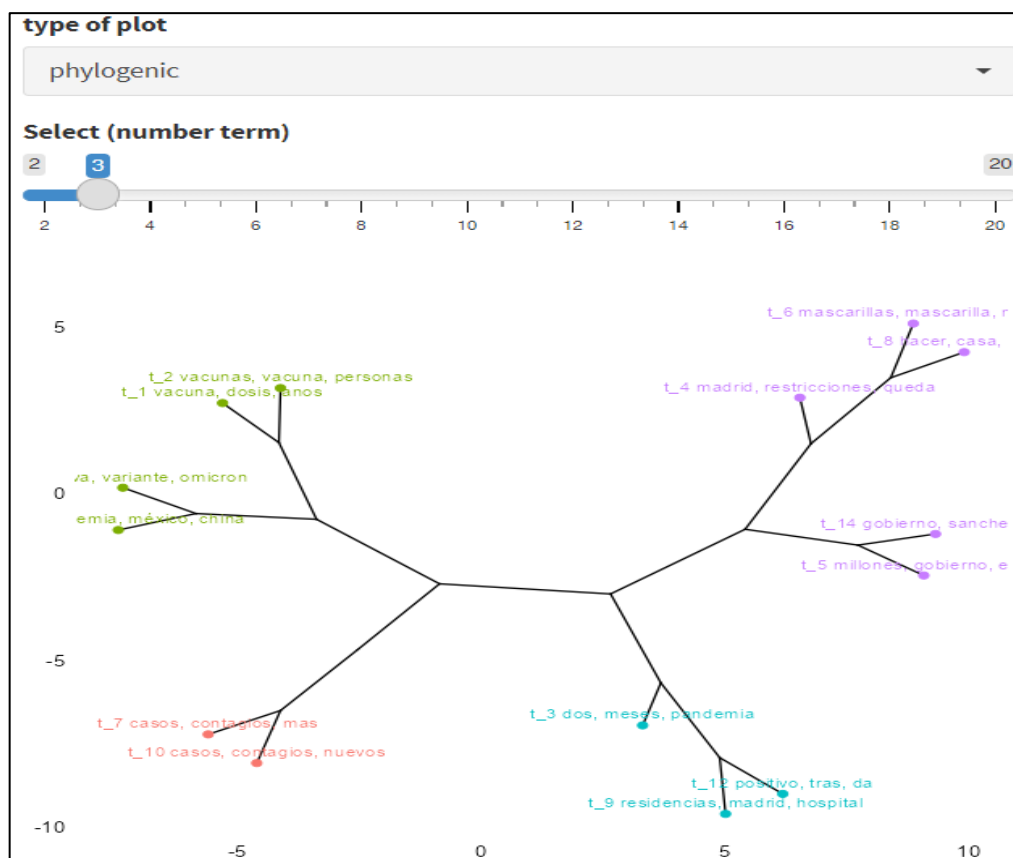


Figura 7.25. Clúster filogenético de los tópicos del LDA.

Se puede observar cómo los 4 clústeres generados agrupan a los tópicos, como por ejemplo se puede observar que un clúster asocia políticas del gobierno en España, donde se encuentran los tópicos 4, 6, 8, 14 y 5; en otro clúster se observa el tópico 7 y 10 donde se tocan nuevos casos de contagio; otro clúster que agrupan los tópicos 1, 2, 11 y 13 que se asocian con vacunas y nuevas variantes, y un último clúster con los tópicos 12, 3, 9 que asocia la pandemia en Madrid.

7.5.2. MENÚ REPRESENTACIONES BIPLLOT

LDABiplots nos permite mostrar los resultados de la matriz theta y phi del LDA obtenida a través de representaciones Biplot. La matriz theta será tratada de acuerdo con el proceso explicado en el apartado 5.3.1. La matriz phi se usará para integrar la minería de texto con las técnicas multivariantes permitiendo analizar el entorno PESTEL.

Para realizar las representaciones Biplot, se debe de generar primeramente el LDA, y en el menú seleccionar entre el HJ_Biplot, JK_Biplot, GH_Biplot, y luego escoger como preprocesado la matriz, entre centrado, escalado, centrado_escalado y ninguna, como método de centrado y escalado de la matriz de covarianzas. Una vez marcadas las opciones se hace correr el proceso.

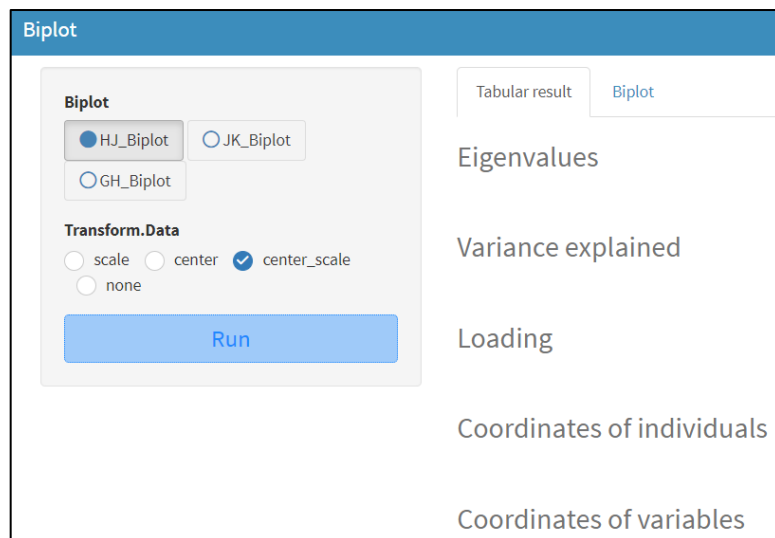


Figura 7.26. Menú Representaciones Biplot.

Los resultados que nos da el programa seleccionando el HJ-Biplot son:

- Valores Propios (Eigenvalues), los resultados se presentan en la figura 7.27.

Eigenvalues	
Dim 1	1.89245
Dim 2	0.91641
Dim 3	0.19114

Figura 7.27. Valores Propios del HJ-Biplot de Matriz Theta del LDA.

- Varianza Explicada (Variance Explained) por los primeros componentes principales obtenidos, los resultados se muestran en la figura 7.28.

Variance explained	
Dim 1	63.08
Dim 2	30.55
Dim 3	6.37

Figura 7.28. Varianza explicada por el HJ-Biplot de Matriz Theta del LDA.

- Cargas (Loading) de los componentes principales, mostrados en la figura 7.29.

	Dim.1	Dim.2	Dim.3
X20M	-0.55977	-0.61663	-0.55356
El.Mundo	-0.68934	-0.02421	0.72404
El.Pais	-0.45987	0.78688	-0.41151

Figura 7.29. Valores de Cargas obtenidos en el HJ-Biplot de Matriz Theta del LDA.

- Coordenadas de individuos (Coordinates of individuals), en este caso las coordenadas de los tópicos que se representan en el HJ-Biplot.

Coordinates of individuals

Show 10 rows ▼ Copy CSV Excel PDF Print

Search:

	Dim.1	Dim.2	Dim.3
t_1	-1.64282	-0.5739	-0.17277
t_10	-2.6327	-0.49771	0.14262
t_11	-0.44367	-0.10411	-0.0183
t_12	1.48895	-0.60746	0.1811
t_13	0.8836	1.29038	-0.89762
t_14	-0.22721	0.35444	0.85763
t_2	1.40407	-0.55667	0.23502
t_3	0.25471	1.07594	-0.43732
t_4	-2.15696	1.50821	0.11364
t_5	0.41316	0.61534	0.11544
t_6	0.96612	-0.4241	0.35357
t_7	-0.65941	-2.19899	-0.63788
t_8	1.78093	-0.04637	-0.06258
t_9	0.57122	0.165	0.22744

Showing 1 to 14 of 14 entries Previous 1 Next

Figura 7.30. Coordenadas de los tópicos en el HJ-Biplot.

- Coordenadas de Variables (Coordinates of variables), donde se observan las coordenadas de los vectores que representan a cada característica en común, en este ejemplo los 3 diarios de noticias de mayor circulación de España.

Coordinates of variables			
Show 10 rows ▾	Copy	CSV	Excel
	PDF	Print	
	Search: <input type="text"/>		
	Dim.1	Dim.2	Dim.3
X20M	-2.77646	-2.12835	-0.87259
El.Mundo	-3.41912	-0.08357	1.14133
El.Pais	-2.28095	2.71596	-0.64869

Showing 1 to 3 of 3 entries Previous **1** Next

Figura 7.31. Coordenadas de los Diarios en el HJ-Biplot.

- Representación Gráfica generada de acuerdo con los parámetros seleccionados, este grafico puede ser modificado en su forma, con las diferentes opciones que brinda el paquete, entre las que tenemos:

- Modificación del fondo o tema
- Los 2 ejes para mostrarse en el grafico
- Los colores de los marcadores columnas
- Los colores de los marcadores filas
- El tamaño de los marcadores
- El tamaño de las etiquetas de ambos marcadores



Figura 7.32. Representación gráfica HJ-Biplot de matriz theta del LDA.

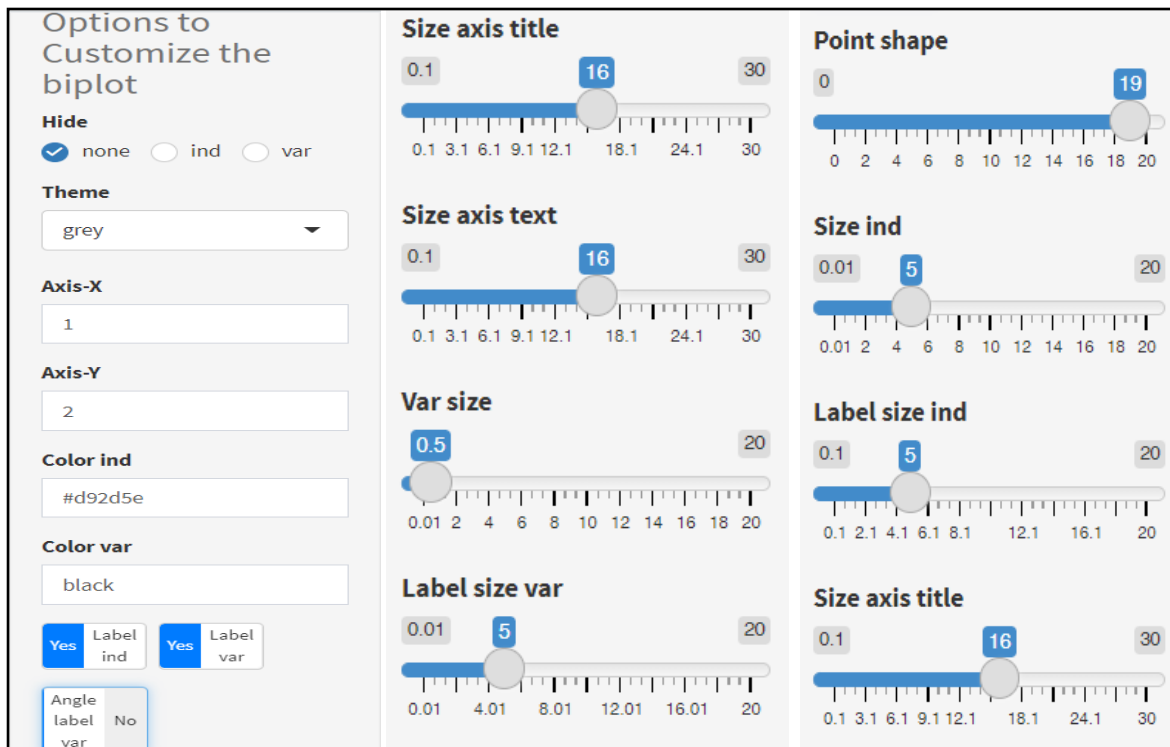


Figura 7.33. Menú de opciones para representaciones gráficas Biplot.

Todas las tablas y representaciones graficas pueden descargarse en diferentes formatos.

En la figura 7.32 se observa las contribuciones que tienen los tópicos con cada uno de los diarios representados en un HJ-Biplot, como por ejemplo se observa como el tópico 4 (tercera dosis) presenta una mayor relación con el diario de noticias EL País, o como el tópico 1 (casos positivos) tiene una mayor relación con los diarios 20M y El Mundo. En la representación HJ-Biplot de los datos obtenidos de la matriz traspuesta de theta, se puede conocer las relaciones de cada uno de los tópicos con los diarios de noticias, con una alta calidad de representación.

7.5.3. MENÚ HJ-BIPLLOT_PESTEL

Este menú permite generar la visualización de la matriz de probabilidades de las palabras del corpus a cada uno de los tópicos (matriz phi) obtenida en el LDA de los datos textuales obtenidos del raspado web (webscraping). Para generar la representación HJ-Biplot_PESTEL se debe de hacer clic en el botón Run PESTEL.

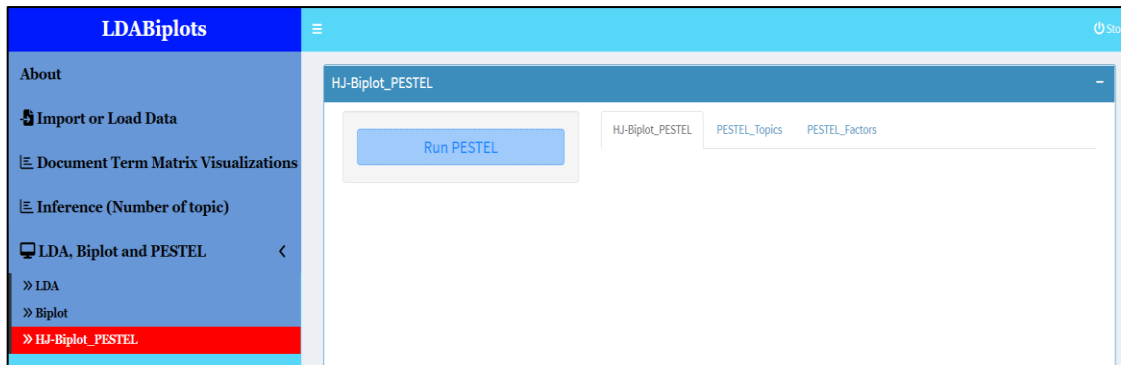


Figura 7.34. Menú HJ-Biplot_PESTEL de la Matriz Phi del LDA

Las representaciones graficas obtenidas son:

- HJ-Biplot de palabras etiquetadas de acuerdo con el léxico PESTEL

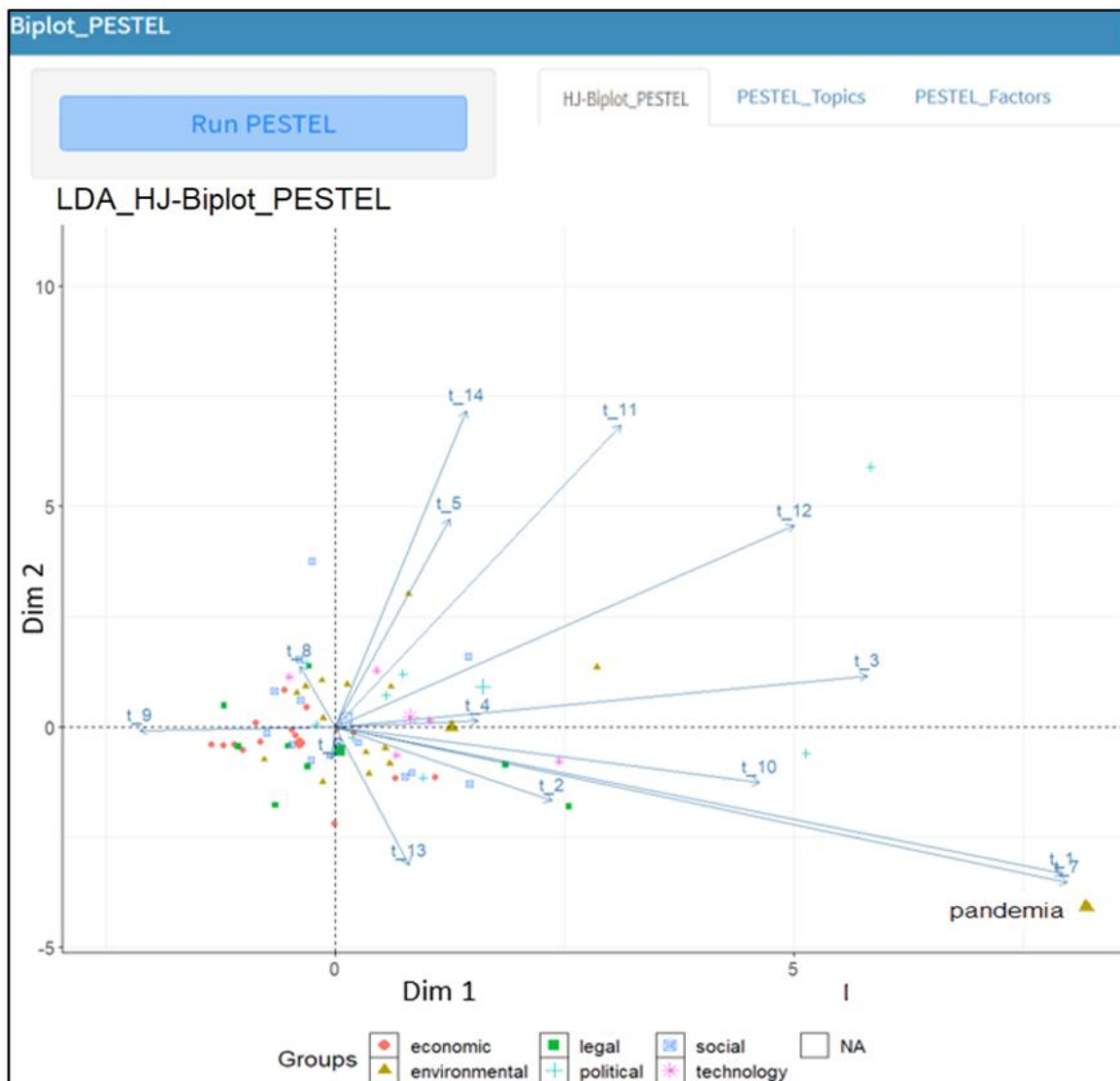


Figura 7.35. Representación LDA_HJ-Biplot_PESTEL de las palabras con respecto a los tópicos obtenidos.

Esta representación que integra los métodos revisados en este trabajo permite observar las relaciones de cada palabra con los tópicos obtenidos en el LDA etiquetando las mismas con el léxico PESTEL, por ejemplo, el término *pandemia* etiquetada con el factor ambiental presenta una relación alta con los tópicos 1 (casos positivos) y 7 (medidas de Pedro Sánchez). Esta integración es de gran utilidad para la toma de decisiones en diferentes ámbitos organizacionales.

Las contribuciones de los marcadores columnas de esta representación HJ-Biplot que son los tópicos se muestran en la figura 7.36, y en la figura 7.37 una sección de las contribuciones de las palabras.

	Dim 1	Dim 2	Dim 3
t_1	57.79	10.21	0.13
t_2	5.13	2.56	0.08
t_3	30.95	1.23	0.48
t_4	2.24	0.02	69.03
t_5	1.43	20.49	6.61
t_6	0.00	0.49	62.67
t_7	58.56	11.43	0.17
t_8	0.14	1.77	3.40
t_9	4.17	0.01	3.01
t_10	19.69	1.46	0.01
t_11	8.90	43.13	1.70
t_12	23.02	19.17	1.62
t_13	0.60	9.08	1.07
t_14	1.86	47.38	0.47

Figura 7.36. Contribuciones de los Tópicos en el LDA_HJ-Biplot_PESTEL

	Dim 1	Dim 2	Dim 3		Dim 1	Dim 2	Dim 3
pandemia	83.34	7.33	0.04	nuevos_casos	33.97	3.39	12.07
crisis	68.39	0.94	1.55	baja	33.00	8.48	12.89
galicia	63.46	1.23	1.30	nuevo	30.90	0.58	7.08
cantabria	58.42	1.92	16.98	horas	30.53	0.65	12.02
euskadi	53.20	0.71	17.69	maskarillas	29.49	44.51	0.35
ultimas_horas	53.08	6.54	15.34	semanas	29.18	2.08	14.99
españa	51.97	17.75	0.89	dia	28.88	5.13	5.02
ultimas	51.41	6.17	15.27	vacunarse	28.09	0.75	42.24
datos	50.39	1.76	15.74	poblacion	27.97	4.62	37.56
supera	50.16	8.98	10.47	gobierno	27.93	28.48	2.32
casos_activos	49.96	5.89	15.13	mundo	26.72	13.43	0.73
cifra	49.67	6.67	16.24	registra	26.63	2.65	10.56
notifica	47.34	5.42	14.77	fallecidos	26.15	1.66	9.38
nuevos_contagios	46.21	5.23	14.59	dosis_vacuna	23.86	5.83	14.58
activos	45.20	5.06	14.42	vacunados	23.48	3.76	12.02
sigue	43.72	6.67	18.16	china	23.34	16.97	0.39
rioja	43.69	6.55	1.13	menores	22.55	6.16	49.69
suma	41.42	4.54	13.87	confinamiento	22.37	5.00	1.13
siete	41.00	4.96	1.24	gran	21.71	24.41	4.21
frente	40.23	1.99	0.44	positivos	20.39	1.57	1.32
mantiene	40.07	1.25	28.78	año	20.14	21.95	6.73
baleares	35.61	5.37	16.19	muertes	19.59	2.84	9.78
virus	35.18	3.61	0.00	pfizer	19.20	5.18	24.94
aumento	34.98	0.53	23.41	muerte	19.09	14.05	6.99
abre	34.51	0.00	12.56	partir	18.78	13.02	0.52
municipios	34.19	0.83	20.17	muertos	18.49	1.13	7.24
				millones	18.14	0.00	1.04

Figura 7.37. Parte de las contribuciones de las Palabras en el LDA_HJ-Biplot_PESTEL

- Gráfico de Barras, donde se observa la proporción de palabras que conforman cada uno de los tópicos generados.

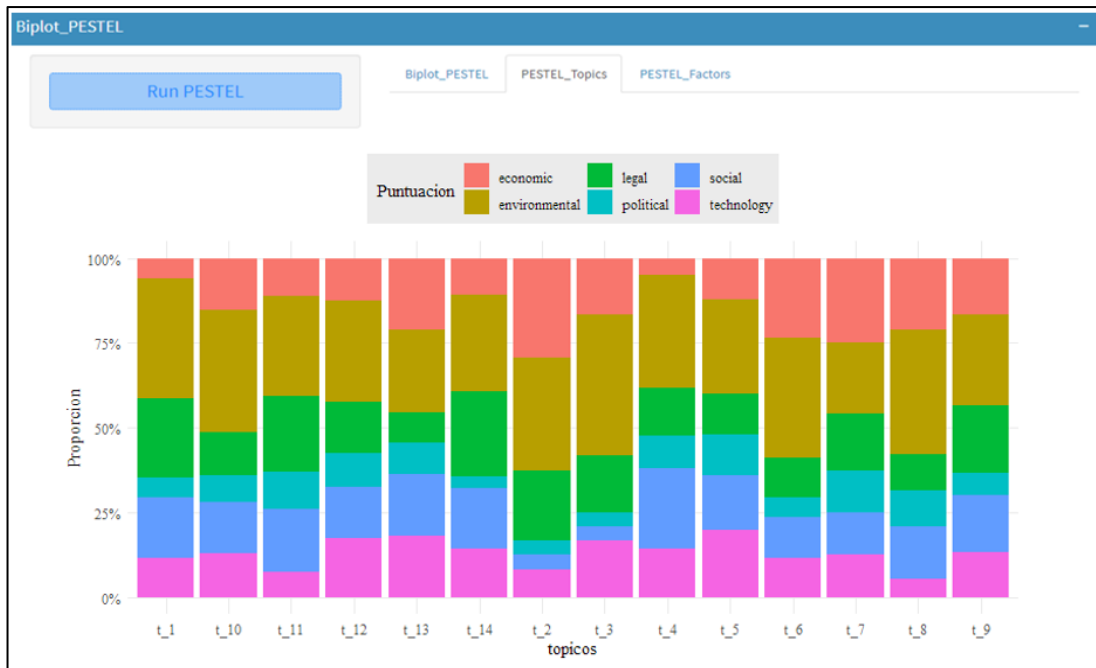


Figura 7.38. Proporción de contenido PESTEL por Tópicos.

En esta grafica se aprecia la proporción de los términos etiquetados por el léxico PESTEL que están contenidos en cada uno de los tópicos, esto se obtiene a partir de la matriz de resumen donde se agrupan las palabras con valores altos de probabilidad para cada tópico. Cabe indicar que una palabra puede estar contenido en diferentes tópicos o en todos pero con diferentes valores de probabilidad, para este análisis grafico solo se considera a los términos que contengan los valores más altos y esto se puede parametrizar a partir del menú del LDA.

Se puede observar que todos los tópicos tienen una proporción similar de contenido ambiental en todos los tópicos, así también se observa que el tópico 4 (tercera dosis) tiene una proporción mínima de factores económicos, pero proporciones altas de factores ambientales y sociales.

Así mismo, se puede observar cómo los tópicos 5 (comunidad Madrid) y 7 (medidas de Pedro Sánchez) presentan proporciones mas altas en factores políticos.

- Gráfico de Barras Horizontales de cada uno de los Factores analizados en el PESTEL, mostrando hasta un máximo de 7 términos por cada tópico.



Figura 7.39. Términos Políticos por Tópicos.

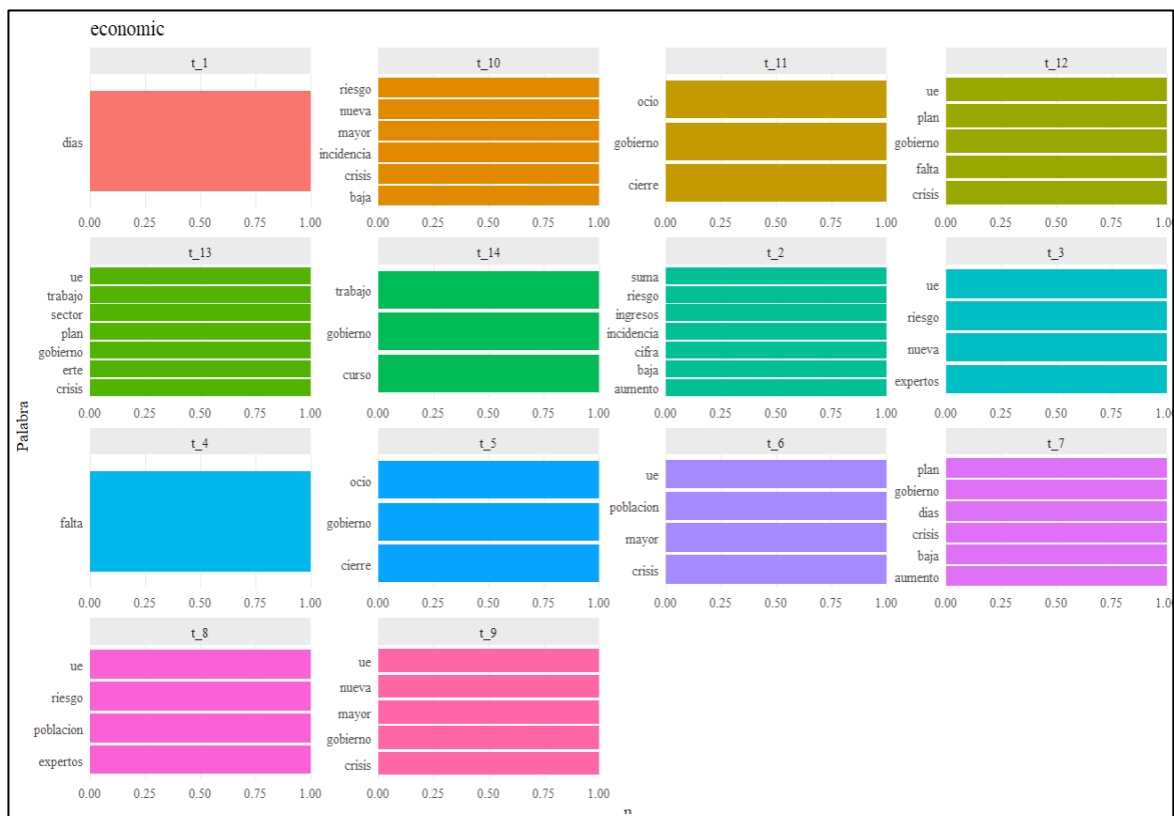


Figura 7.40. Términos Económicos por Tópicos

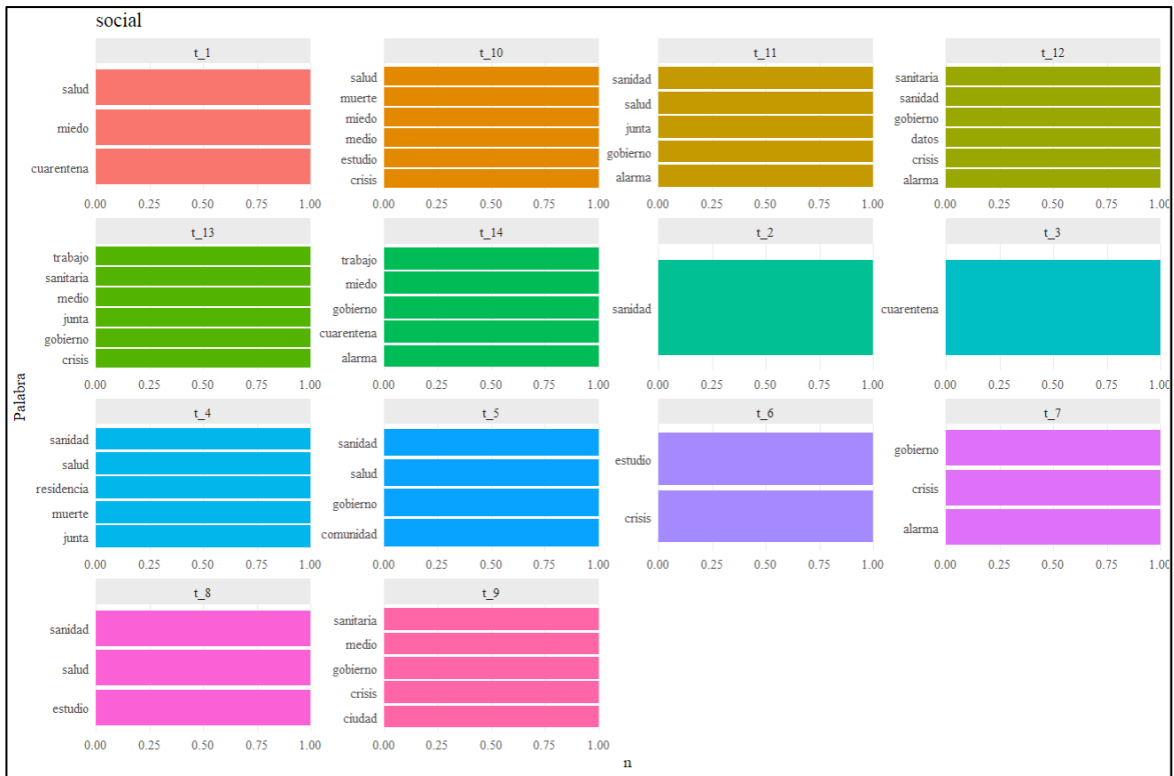


Figura 7.41. Términos Sociales por Tópicos

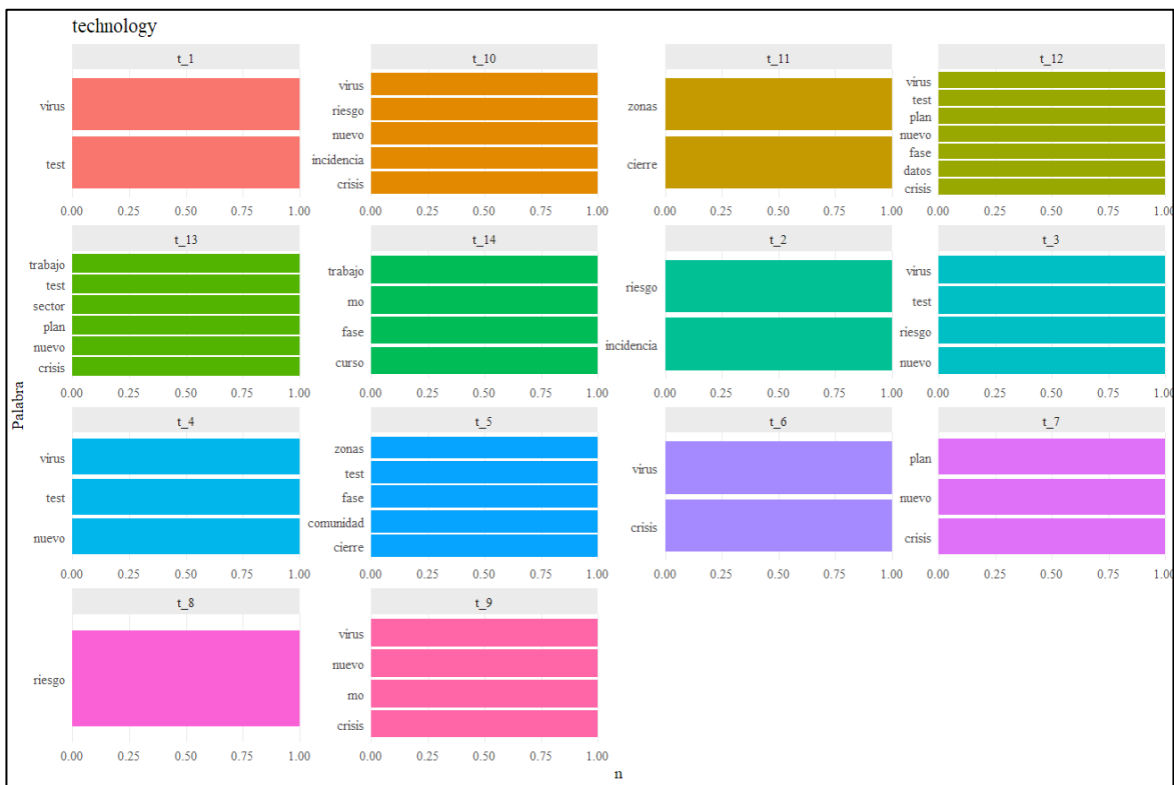


Figura 7.42. Términos Tecnológicos por Tópicos.



Figura 7.43. Términos Ambientales por Tópicos.

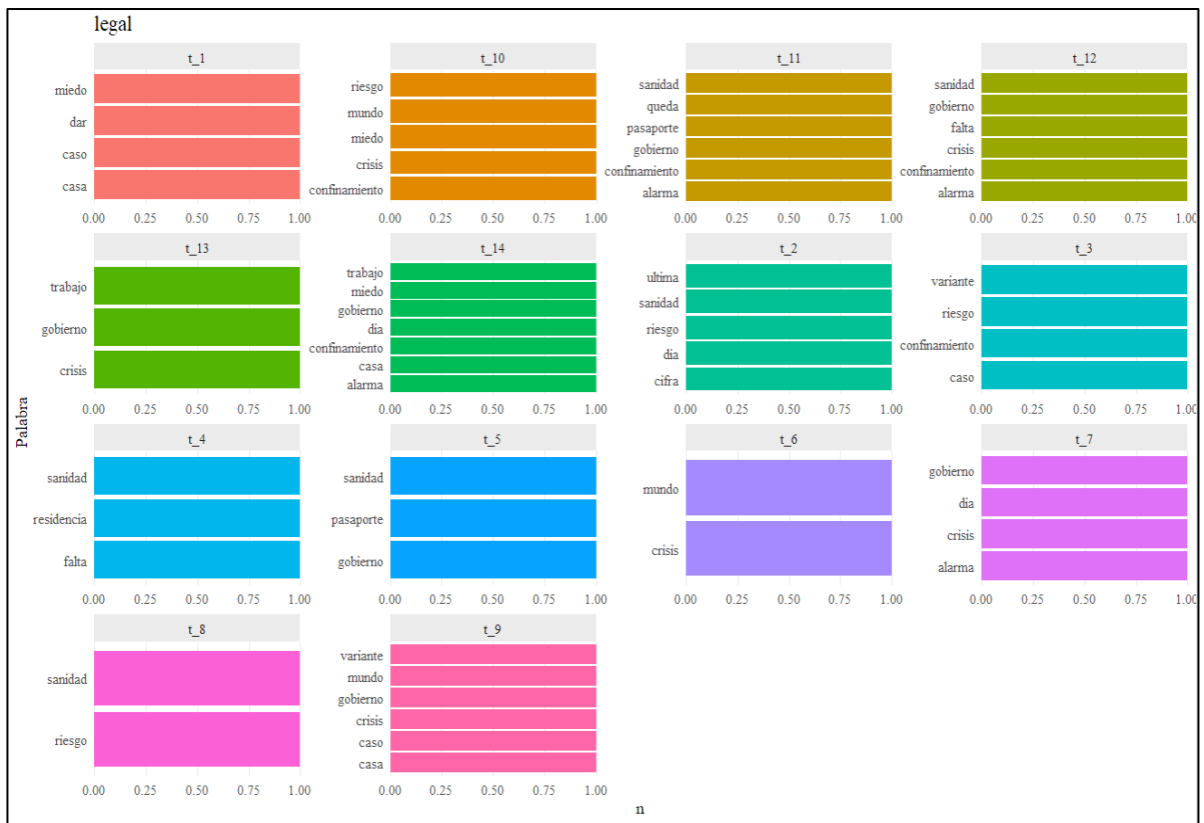


Figura 7.44. Términos Legales por Tópicos.

Los gráficos 7.39 a 7.44 nos muestran el contenido de cada palabra etiquetada por cada uno de los factores del macroentorno que se encuentra contenido en cada uno de los tópicos.

Todos los análisis estudiados en el presente trabajo han permitido desarrollar un software de acceso libre para investigadores y comunidad en general, que permite integrar la minería de texto, con la técnica multivariante HJ-Biplot aplicándola en el análisis del macroentorno PESTEL. Esta herramienta denominada LDABiplots buscar realizar análisis del macroentorno PESTEL a partir de publicaciones web de los diarios de noticias, los resultados buscan ser el punto de partida para las organizaciones para el desarrollo de estrategias a corto y largo plazo.

En el LDABiplots el tiempo de procesamiento estará entre 0.5 a 10 horas, esto va a depender de las características del equipo de cómputo usado, así como del tamaño del corpus a ser extraído y de la conectividad web. Sin embargo, este tiempo es menor al desarrollado por un proceso de análisis PESTEL convencional, donde de acuerdo con Perera 2020 (Perera, 2020), se requiere de todo un proceso para aplicación de este análisis el cual puede tardar varios días.

CONCLUSIONES

CONCLUSIONES.

1. La revisión bibliográfica puso de manifiesto que la minería de datos textuales (MDT) y el análisis estadístico de datos textuales (AEDT) son técnicas de investigación activas a nivel mundial que se aplican cada vez más frecuente en diferentes ámbitos, como análisis de redes sociales, de literatura científica, de textos digitales de repositorios y en este trabajo lo hemos usado para el análisis de noticias de los periódicos digitales.
2. En este trabajo se revisaron algunas metodologías para el tratamiento estadístico de datos textuales, como el análisis semántico latente (LSA), el análisis probabilístico semántico latente (PLSA) y la asignación latente de Dirichlet (LDA). Este último usado especialmente para la minería de datos textuales, análisis de sentimientos y modelado de temas ocultos para descubrimiento de nuevo conocimiento y base de la metodología propuesta.
3. Se desarrollo una metodología que integra herramientas de la minería de datos textuales, como el raspado web (webscraping), el procesamiento y estructuración de datos en tablas léxicas, el modelado de tópicos; con técnicas multivariantes con las representaciones Biplot, en especial énfasis en el HJ-Biplot para el análisis del entorno externo organizacional PESTEL.
4. Se ha desarrollado un aplicativo interactivo LDABiplots, bajo un lenguaje de programación abierto que integra la metodología propuesta, ofreciendo múltiples funciones para la minería de datos textuales. [confirmar la autoría del programa]
5. Se ha desarrollado un léxico PESTEL, que permite el etiquetado de las palabras para representar su contribución con cada tópico generado en el LDA, generando así una representación LDA_HJ-Biplot_PESTEL.
6. La metodología propuesta se aplicó a 48112 noticias extraídas de tres páginas HTML de diarios web de España, encontrándose 14 temas ocultos en los titulares de estas noticias, donde el factor ambiental se manifiesta en mayor proporción en todos los tópicos, en el t_2 etiquetado como nuevos contagios se ha encontrado que pandemia es la palabra presente que mayor contribución tiene con este tópico.

7. El HJ-Biplot es una alternativa metodológica que permite el análisis de los factores externos organizacionales PESTEL, ya que se pueden observar las contribuciones de la probabilidad de cada palabra con respecto a cada uno de los tópicos generados en el proceso LDA.

8. Existen pocos paquetes que generan de manera interactiva los modelos LDA. A diferencia de los ya existentes el paquete LDABiplots integra desde la extracción de datos textuales de la web, hasta la incorporación de las representaciones Biplot para el análisis de las matrices de probabilidad obtenidas en el modelado de tópicos, e integra en el HJ-Biplot el análisis del entorno organizacional PESTEL

CONTRIBUCIONES CIENTÍFICAS



Article

HJ-Biplot as a Tool to Give an Extra Analytical Boost for the Latent Dirichlet Assignment (LDA) Model: With an Application to Digital News Analysis about COVID-19

Luis Pilacuan-Bonete ^{1,2,*}, Purificación Galindo-Villardón ^{1,3,4} and Francisco Delgado-Álvarez ¹

¹ Department of Statistics, University of Salamanca, 37008 Salamanca, Spain; pgalindo@usal.es (P.G.-V.); jdelgado@usales (F.D.-Á.)

² Faculty of Industrial Engineering, Universidad de Guayaquil, Guayaquil 090514, Ecuador

³ Escuela Superior Politécnica del Litoral, Escuela Superior Politécnica del Litoral (ESPOL), Centro de Estudios e Investigaciones Estadísticas, Campus Gustavo Galindo, Km. 30.5 Via Perimetral, Guayaquil P.O. Box 09-01-5863, Ecuador

⁴ Centro de Gestión de Estudios Estadísticos, Universidad Estatal de Milagro (UNEMI), Ciudadela Universitaria Km. 1.5 vía al Km 26, Guayas 091050, Ecuador

* Correspondence: luis.pilacuanb@usal.es; Tel.: +593-981105994

Abstract: This work objective is to generate an HJ-biplot representation for the content analysis obtained by latent Dirichlet assignment (LDA) of the headlines of three Spanish newspapers in their web versions referring to the topic of the pandemic caused by the SARS-CoV-2 virus (COVID-19) with more than 500 million affected and almost six million deaths to date. The HJ-biplot is used to give an extra analytical boost to the model, it is an easy-to-interpret multivariate technique which does not require in-depth knowledge of statistics, allows capturing the relationship between the topics about the COVID-19 news and the three digital newspapers, and it compares them with LDAvis and heatmap representations, the HJ-biplot provides a better representation and visualization, allowing us to analyze the relationship between each newspaper analyzed (column markers represented by vectors) and the 14 topics obtained from the LDA model (row markers represented by points) represented in the plane with the greatest informative capacity. It is concluded that the newspapers El Mundo and 20 M present greater homogeneity between the topics published during the pandemic, while El País presents topics that are less related to the other two newspapers, highlighting topics such as t_12 (Government_Madrid) and t_13 (Government_millions).

Keywords: SARS-CoV-2; COVID-19; HJ-biplot; latent Dirichlet assignment; LDA

MSC: 62H35



Citation: Pilacuan-Bonete, L.; Galindo-Villardón, P.; Delgado-Álvarez, F. HJ-Biplot as a Tool to Give an Extra Analytical Boost for the Latent Dirichlet Assignment (LDA) Model: With an Application to Digital News Analysis about COVID-19. *Mathematics* **2022**, *10*, 2529. <https://doi.org/10.3390/math10142529>

Academic Editor: Andrea De Gaetano

Received: 1 June 2022

Accepted: 24 June 2022

Published: 20 July 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Humanity is suffering from a pandemic caused by the SARS-CoV-2 virus with more than 500 million people affected and almost six million deaths to date. This tragic situation is causing opinions and information related to SARS-CoV-2 (COVID-19) to be written in all available media: print and digital press, social networks, web pages, forums, etc. A technological resource available for the analysis of this web information is the textual analysis of content, which is being used regularly in all types of environments, including web environments. Much research has been carried out on this type of analysis, for different applications, such as studying publications on social networks [1], analyzing the marketing management of companies [2], establishing worker profiles from LinkedIn [3], analysis of scientific literature [4], and analyzing effects on health workers through Twitter posts [5], among others.

As of 27 May 2022, there were 528,431,653 confirmed cases of the SARS-CoV-2 virus registered in the world according to the World Health Organization [6]. This disease was

first identified in hospitalized patients in Wuhan, China, in December 2019 [7] and is associated with symptoms of severe pneumonia, causing fever, cough, and respiratory failure [8], being able to cause the death of the infected patient. The COVID-19 disease was declared a pandemic by the WHO on 11 March 2020, and is affecting not only human health but also the world economy [9], generating high interest in people searching for information in credible media such as newspapers. In Spain, the first case of infection by the virus was registered on the island of La Gomera on 31 January 2020 [10] and as of 27 May of the same year, 12,326,264 people had been infected [6].

According to the Association for Media Research (AIMC), the three newspapers in Spain with the highest number of daily readers are El País, El Mundo, and 20 Minutos [11] with more than 570k readers per day each. The news published by the newspapers can influence the public opinion of the readers [12]. Multiple studies have analyzed this influence on the readers of the news published in various newspapers, such as in politics [13], consumption [14], and discrimination against criminals [15]. The world is currently experiencing dark times due to the pandemic caused by SARS-CoV-2, also known as the 'coronavirus', making the study of the publications generated on this subject especially relevant.

For the textual analysis of the information available in cyberspace, different techniques are applied to consist of obtaining information from the web, known as 'web scraping', which allows the extraction of part or all the data from web pages, written in different formats such as XML and HTML, among others [16]. These extracted data can be subjected to different transformation processes of semantic and syntactic information, through natural language processing (NLP) techniques for the formulation of text corpus [17].

The information thus extracted is subjected to different textual analysis techniques to analyze this text corpus, such as term frequency analysis (TF), which allows statistical interpretation of the specificity of the term and its application in retrieval [18]. A quantitative corpus is generated that describes the words as variables and the n documents extracted as individuals, to represent the textual fragments as a linear equation. These corpuses are structured following different semantic analysis techniques. Among these techniques are those of latent semantic analysis (LSA), which uses singular value decomposition (SVD) for the segmentation of the corpus matrix, applying the statistical basis of the co-occurrence of words in the corpus and ignoring the grammatical structure [19]. Another methodology used is the latent Dirichlet allocation (LDA), consisting of a three-level hierarchical Bayesian probabilistic model, which allows a collection of observations to be explained as a whole and which shows the similarity between the extracted data [20], another technique is the use of machine learning algorithms, applied in textual studies with MTL multitasking learning models [21], or text sentiment analysis applications based on the synthetic minority over-sampling technique (SMOTE) [22], among others applications.

The text corpus, methods such as the LDAvis [23], a very extended alternative for the representation of the topics, which allows display on a web page using an interactive graphic (scalable vector graphics, SVG). Representation employing a heatmap type has also been proposed [24] allows us to visualize a rearrangement by some set of values, usually the mean of the rows or columns. Another proposed technique is the HJ-biplot method formulated by Galindo [25] in which the rows and columns of a data matrix are represented in the same system of factorial axes; examples can be cited from the works developed to represent bibliometric studies [26], to represent the quality of life discussion groups [27], or to classify the investiture speeches of Spanish rulers [28].

Based on everything mentioned above, the present study aims to generate an HJ-biplot representation of the distribution matrix of the topics on the documents, resulting from an LDA analysis of the news generated regarding the SARS-CoV-2 pandemic in the three generalist Spanish newspapers with the highest number of readers, thus allowing visualization of the relationship for each topic of the news of the newspapers concerning COVID-19.

2. Materials and Methods

For the content analysis of the three most read newspapers in Spain, text mining tools and techniques are used which allow for the generation—through statistical methods—a visualization of the data [29]. Parallelization techniques have been applied in data processing to obtain higher computational performance [30]. A standard text mining process starts from the integration of raw information, coming from different data sources, which is cleaned to eliminate inconsistencies and duplicates that generate noise in the analysis [31]. With the data transformed into a homogeneous format, and through text mining filtering and aggregation techniques, analyses can be carried out where the most interesting existing patterns are identified.

Next, the various techniques applied in each of the processes followed in the analysis will be presented, detailed in Figure 1. These techniques were applied in pre-existing modules for the different analyses in an open-source software R version 3.6.3 [32].

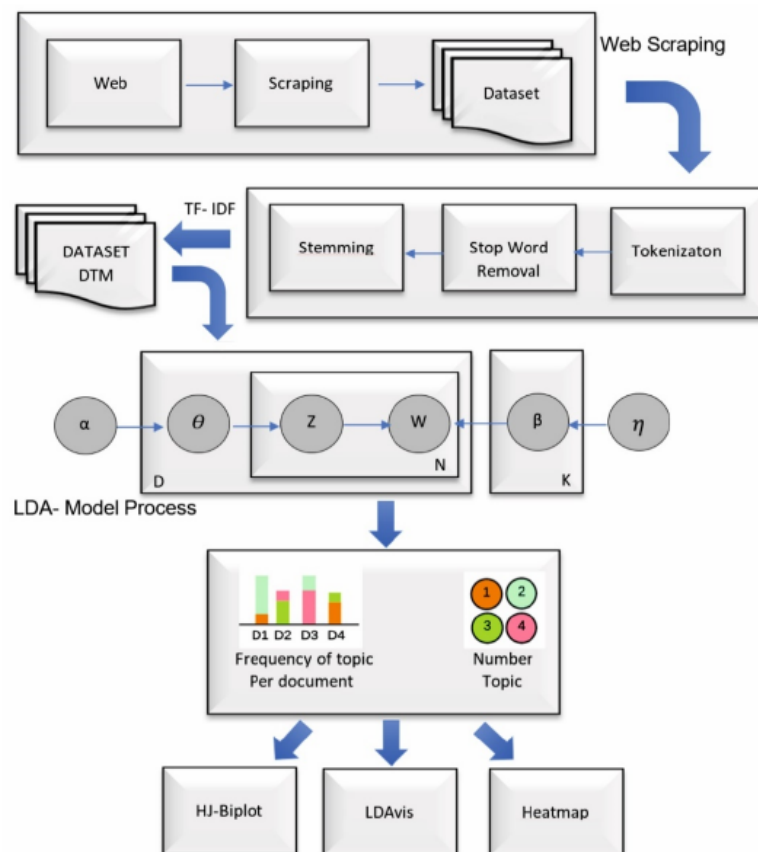


Figure 1. Applied methodology architecture.

2.1. Web Data Extraction

Data extraction techniques from the web have had high growth in recent times. Due to the massification of information on the World Wide Web (WWW), this has become an important global database. Web scraping, used for web content mining, obtains information from the content of web pages, with two basic objectives: extract information to improve search engines and information retrieval fields [33] and analyze and explore information to

gain useful content knowledge [34]. In this content mining technique, opinions, feelings, and emotions are extracted from the text to understand the context of web content [35].

These text analysis processes are widely used in a wide range of applications: there are several studies where these techniques have been applied in the extraction of information from Facebook [36], Twitter [37], web pages [38] in general, and bibliographic sites to obtain academic information, such as expert recommendations [39], among other sites with different study purposes [40]. Rekik [41] concluded that it is important to evaluate the quality of information on websites; for this, firstly, the criteria to carry out the evaluation are defined and regrouped semantically; subsequently, useful information is extracted from them to create a set of data criteria with which to obtain more specific information.

Ferrara [41] detailed how complex it can be to extract information from the web, especially when it comes to unstructured data, or depending on the languages of each page and even the browsers used; but he also described the multiple applications that these techniques have in different fields, taking into account the computational cost of these methodologies.

For the present study, as already indicated, the R software environment will be used, which provides multiple supports for web mining [42], and has different packages and functions that handle data extraction tasks, such as Rcrawler [43]. However, the user must manually manage the content that they want to extract from the URLs—that is, data cannot be obtained automatically—so other types of tools are also required if the extraction process needs to be automated.

In addition, the R package called Rvest (version 0.3.5) has been used, which allows information to be extracted from URLs that are in HTML or XML format [44]. This package, depending on the tasks to be carried out, can be combined with others to incorporate other functionalities. Specifically, in the present study, it was combined with the R packages dplyr (version 0.8.5) and the Base package (version 3.5.0), to extract the unstructured data and subject it to a cleaning and transformation process in character-type text format, to later be submitted to the final content analysis.

2.2. Term Frequency

The methodology called term frequency in the document, or TF (term frequency), found its way into almost all terminology weighting schemes. According to Jones's postulate [18], terms can be said to be words or possibly phrases or word-words; assuming that there are N documents in a collection and that the term t_i appears in n_i of them, then the proposed measure, defined as a weight, will be applied to the term t_i , and is described in Equation (1), also known as the inverse document frequency (or IDF), this formulation being one of the most used (Robertson, 2004).

$$\text{idf}(t_i) = \log \frac{N}{n_i} \quad (1)$$

The assignment of unique weighted terms properly produces retrieval results superior to those that can be obtained with other text techniques used, depending on the term weighting system chosen [45,46]. In this study, the weighting of terms called TF-IDF is considered, which is a metric where the TF provides a direct estimate of the probability of occurrence of a term, normalized by the total frequency of the document [47] and in which this indicator is multiplied for the IDF, which in turn can be interpreted as the amount of information, given as the log of the inverse probability [18].

Taking an array of terms as input, the R package named texmineR (version 3.0.4) obtains using the TermDocFreq function [48] a data matrix with columns for term frequency, document frequency, and weighted inverse document frequency [49]. This package allows applying lemmatization and elimination of those words that the researcher considers 'noisy'—such as adjectives, articles, or other such words commonly called 'stop words'.

2.3. Latent Dirichlet Assignment (LDA)

The latent Dirichlet allocation (LDA) is a generative probabilistic model of a corpus, the basic idea being that a random mix of latent topics is represented, where each topic is characterized by a distribution over words [20]. LDA is a Bayesian variant of probabilistic latent semantic analysis (PLSA), whose predecessor is latent semantic analysis (LSA), LDA is based on a set of assumptions, which states that words in a text are interchangeable between documents and that documents they are represented as a string of individual words that make up the document.

The probabilistic generative process is defined by Blei and Lafferty [50] and is represented in Figure 2. To give a better understanding of the LDA graph, the observed data are the words of a document, and the hidden variables represent the structure of the latent topics. The interaction between the observed documents and the structure of the topics is manifested in the generative process associated with LDA. The generative process will be rewritten, but only the steps for generating the entire document collection will be presented.

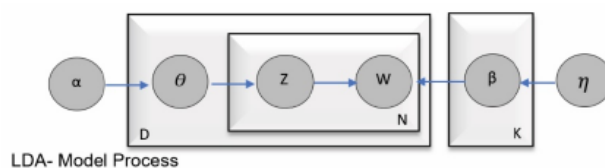


Figure 2. Probabilistic generative process LDA.

The LDA generative process assumes the documents come, is described as:

1. For each topic K
 - I. Draw a distribution over the words (i.e., vocabulary V), $\beta_k \sim Dir_V(\eta)$
2. For each document D :
 - I. Draw a distribution over topics (i.e., ratio of the topic to document) $\theta_d \sim Dir(\alpha)$
 - II. For each word w within document D :
 - i. Draw a topic assignment, $z_{d,n} \sim Mult(\theta_d)$ (i.e., topic assignment per word)
 - ii. draw a word $w_{d,n} \sim Mult(\beta_{z_{d,n}})$

Where each topic k comes from a Dirichlet distribution $\beta_k \sim Dir_V(\eta)$ and is a multinomial distribution over the vocabulary, each document D is represented as a distribution of topics and originates from a $\theta_d \sim Dir(\alpha)$. The Dirichlet parameter η defines the probability of words within topics, and α the probability of topics within documents. The joint distribution of all hidden variables β_k, θ_d (document topic ratios within D), $z_{d,n}$ (word topic assignments), and observed variables $w_{d,n}$ (words in documents), is described in Equation (2):

$$P(\beta_k, \theta_D, Z_D, W_D) = \prod_{k=1}^K P(\beta_k | \eta) \prod_{d=1}^D P(\theta_d | \alpha) \prod_{n=1}^N P(Z_{d,n} | \theta_d) P(W_{d,n} | \beta_k) \quad (2)$$

In LDA, all documents share the same set of topics, but each document shows the corresponding topic in different proportions. This process, analyzed in-depth by Blei [20], is computationally intractable but it is the key to LDA, so approximation methods must be applied both for quantitative calculations; such as the generalization of predictions and documents, and for exploratory tasks. To obtain the LDA model with the data under study, the R textmineR package (version 3.0.4) was used, through the FitLdaModel function, which allows us to fit a Dirichlet latent assignment topic model through Gibbs sampling, which is used to obtain the variational inference approach [51]. This model is composed of three matrices:

theta (θ): distribution of topics on the documents;

phi (Φ): distribution of words on topics;
 Gamma (γ): distribution of topics by words;
 Equation (3) describes the calculation of the posterior probabilities of the distributions of the observations of seeing the corpus observed in each one of the topics.

$$P(\beta_k, \theta_D, Z_D | W_D) = \frac{P(\beta_k, \theta_D, Z_D, W_D)}{P(W_D)} \tag{3}$$

Gibbs sampling provides direct estimates of the topic assignment Z for each word, from the estimates θ is described in Equation (4) and Φ is described in Equation (5) of the within-document topic distributions and word-topic distributions, respectively.

where:
 C^{WK} is a matrix of dimension $W(\text{words}) \times K(\text{topics})$, where C_{ij}^{WK} is the number of times word i is assigned to topic j .
 C^{DK} is a matrix of dimension $D(\text{Documents}) \times K$, where C_{dj}^{DK} is the number of times topic j is assigned to some keyword in document d .
 α y β are hyperparameters, which act as constraints on the model.

$$\theta_j^{(d)} = \frac{C_{dj}^{DK} + \alpha}{\sum_{k=1}^K C_{dk}^{DK} + K \alpha} \tag{4}$$

$$\Phi_i^{(j)} = \frac{C_{ij}^{WK} + \beta}{\sum_{k=1}^W C_{kj}^{WK} + W \beta} \tag{5}$$

Both expressions according to the interpretations of Steyvers and Griffiths [51] correspond to the predictive distributions of sampling a new i -th word from the j -th topic/topic and sampling a new word (not yet observed) in document d from the j -th topic/topic. Once the model has been created, its goodness of fit is evaluated using the well-known coefficient of determination R^2 applied to topic models. This figure of merit is interpreted in the usual way, as the proportion of variability in the data explained by the model [52]. The textmineR package also allows calculating several indicators on the LDA model, such as the probabilistic coherence [53], which can be interpreted as an estimate of the comprehensibility of a topic by a human.

2.4. HJ-Biplot

Biplots, proposed by Gabriel [54], are graphical representations of multivariate data, allowing the visualization of three or more variables, like a scatter plot showing the joint distribution of two variables. The HJ-biplot is a multidimensional data technique proposed as an alternative to improve the classical biplots introduced by Gabriel, the GH-biplot achieves a high-quality representation of the variables (column marker), while the JK-biplot achieves a high-quality in the ranges of the individuals (row marker). An alternative to optimize biplot methods described by Galindo [25] proposed a multivariate technique called HJ-biplot.

The HJ-biplot [25] is a multivariate graphical representation of a matrix X using markers j_1, \dots, j_f for its rows and h_1, \dots, h_c for its columns, chosen so that both markers can be superimposed in the same reference system with maximum representation quality. It is an evolution of the biplots formulated by Gabriel [54] and both are based on the decomposition singular values [55] of the starting matrix X and the subsequent definition of a lower rank approximation for the same [56]. This representation is described in Equation (6), where the HJ-biplot [25] is defined as

$$\begin{matrix} X = UDX^T & J = U D \\ & H = V D \end{matrix} \tag{6}$$

where X is the data matrix, U is the matrix orthogonal of data columns containing the eigenvectors of XX^T , V is the matrix orthogonal of data whose columns contain the eigenvectors of $X^T X$, and D is the diagonal matrix containing the eigenvalues of X . The row markers are matched to the rows of the JK biplot markers ($J = UD$); in turn, the column markers of the HJ biplot match the marked columns of the GH biplot ($H = VD$), considering the matrix x centered.

In our study, the HJ -biplot has been applied for the graphical representation of the theta and phi data matrix obtained from the LDA analysis and thus visualize the distribution of the words w , with the Topics K and with the documents D . For the analysis of the data of the present study, the R Package `GGBiplotGUI` 1.0.9 was used [57], which allows different types of biplot representations to be made through a graphical interface. This package has been used in different publications for the analysis of different types of data—including environmental, genetic, and agronomic data [58].

2.5. LDavis and Heatmap

To obtain an overview of the topics and the differences between them, as well as to facilitate a graphic review of the words most associated with each topic individually, an alternative used is `LDavis`, which is an interactive web-based visualization of the estimated topics by the latent Dirichlet assignment which is created by a combination of R and D3 using the popular D3 JavaScript library [59].

A heatmap is a graphic representation of data where the individual values contained in a matrix are represented as colors, both static and interactive; normally, the rows and columns are reordered by the averages obtained, or according to the restrictions imposed by the user of the package of `r`. The function is provided natively in R. It produces a high-quality matrix and offers statistical tools to normalize the input data, run clustering algorithms, and visualize the result with dendrograms.

Both representation methods are available in R packages. For the application of the `LDavis` method, the `LDavis` package (version 0.3.2) and the `ComplexHeatmap` package (version 2.12.0) were used to compare the results with the obtained in the HJ -biplot.

3. Results

A web scraping technique has been applied, using the `Rvest` de R package, to the pages dedicated to the coronavirus in the three newspapers understudy and published in the following URLs that correspond to the headlines of the news related to the coronavirus: COVID-19 in Spain, published from 1 January 2019 to 27 May 2022: '<https://elpais.com/noticias/coronavirus/>', '<https://www.20minutos.es/busqueda/1/?q=covid+coronavirus/>', '<https://www.elmundo.es/e/co/coronavirus.html>' (accessed on 27 May 2022).

With the collected data, a matrix of 3 columns and 48,112 rows was built, the first column contains an identifier of each document, the second column contains the name of the newspapers, and the third is the headlines of the news extracted from the website of each newspaper. Table 1 shows the number of web news headlines obtained by each newspaper.

Table 1. Number of headlines for each newspaper.

ID	Newspapers	Frequency
1	El-Pais	19,375
2	El-Mundo	18,547
3	20 M	10,190

The `CreateDtm` function, from the `textmineR` package, was applied to create the matrix of document terms (news headlines for each newspaper). Stop words usual in Spanish, among which the word 'Coronavirus' was also included, since it is considered that it would generate noise in the headlines, due to its possibly high frequency of appearance, and punctuation marks were also removed, to separate the words from the titles. The `TermDocFreq` function was applied to this matrix, obtaining a `dgCMatrix` DTM with

247,137 words in the columns and 48,112 documents in the rows. Using the package functions, a term frequency matrix is generated with the respective IDF weight of each term. To facilitate a graphical analysis in the present study, cleaning of the words whose frequency is less than 900 repetitions or appearances in the entire corpus of the generated text is carried out, finally obtaining a matrix with only 22 terms of the initially generated corpus. Figure 3 shows part of the most used words to create the LDA model.

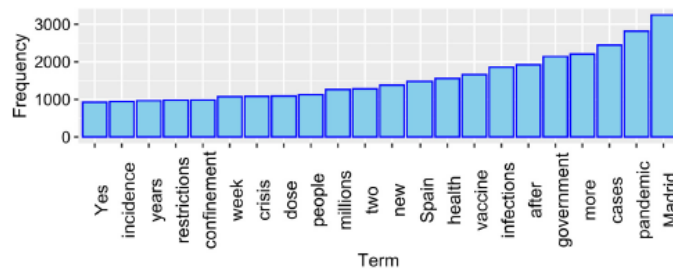


Figure 3. Terms with a higher frequency of matrix DTM.

To generate the LDA model with the FitLdaModel function of the TextmineR package, the optimal number of K topics was first determined, according to the coherence explained by the terms found in each topic. When analyzing the coherence through Figure 4, it is found that 14 are the topics that represent the greatest coherence of the model of 20 possible topics initially evaluated. With the value obtained, an LDA model is generated, restricting it to 14 topics, to obtain the theta Θ , phi Φ , and gamma γ matrices.

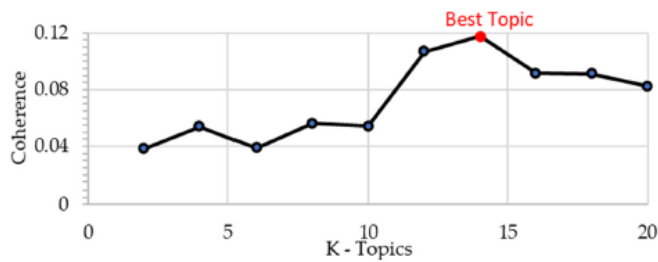


Figure 4. Coherence of topics.

3.1. Results LDAvis and Heatmap

For the analysis of the results obtained in the LDA, it is also represented by means of a heatmap which represents the possibility that a topic K belongs to a newspaper by means of a heatmap, as shown in Figure 5. It is observed as topics t_9 (USA_Pandemic) and t_6 (Vaccines_people) present a higher average frequency in the 20 M newspaper, the topic t_9 also shows to be relevant for the newspaper El Mundo, while in El País it is shown that almost all the topics present a proportion in this figure, the calculation of the averages of the documents of each newspaper with the topics was used, and these values are those represented in Figure 5.

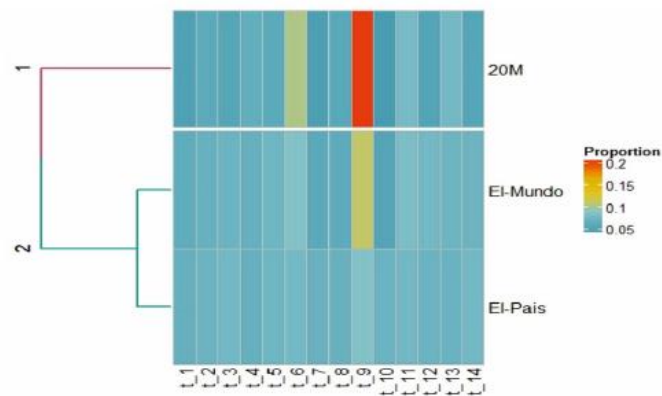


Figure 5. Heatmap of the obtained LDA.

LDAvis allows the visualization of the topics obtained from the Model LDA generated in the study on a web page using an interactive graphic (SVG, scalable vector graphics). With this method, the topics are represented with a circle, the greater the diameter of the circle, the greater the proportion of words in this topic of the model. These circles are represented positioned in a multidimensional scaling plane (MDS), where the distance map between topics is a visualization of these in a two-dimensional space, where the circles are plotted using a multidimensional scaling algorithm based on the words that they contain, so the closer topics have more words in common. The web application allows you to interact with the graph so that when you select the topic you can see the words that make up the selected topic, ordered in decreasing order according to the frequency with which they appear in each topic. As an example of this functionality, Figure 6 shows topic 2 the words “vaccine”, “dose”, “years”, “third”, and “older” are positioned among the most representative of the generated topic.

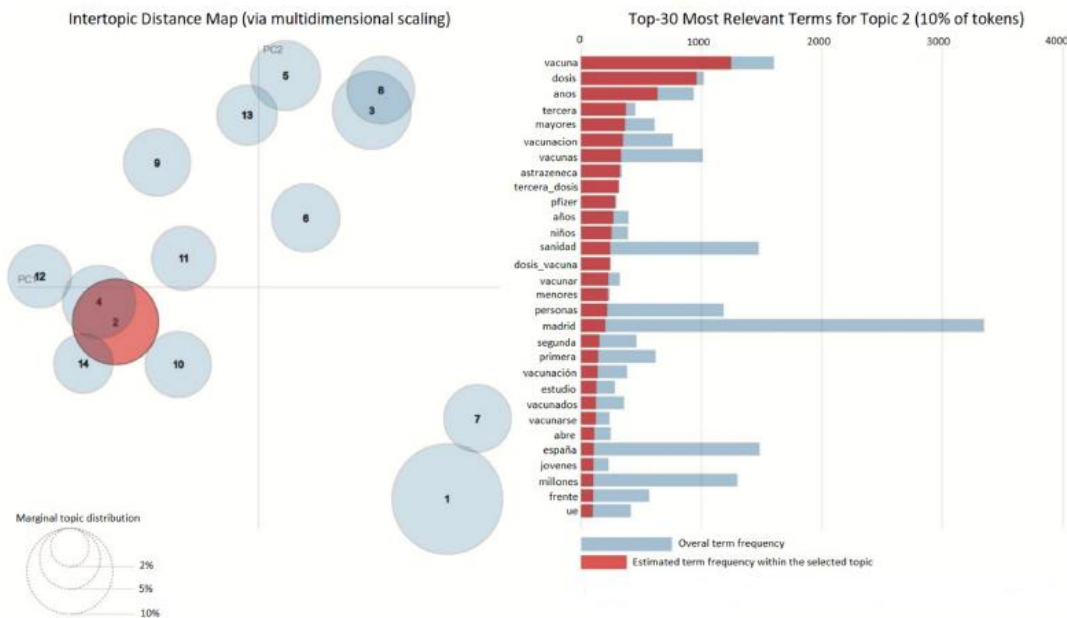


Figure 6. LDAvis representation of topics.

3.2. Results HJ-Biplot

To obtain the HJ-biplot representation of the LDA model obtained, the three matrices θ , Φ , and γ were generated using the summary function of the Base package of *r*, in which the 14 topics are characterized, obtaining the theta matrices θ (48,112 Documents \times 14 Topics), phi Φ (226 words \times 14 topics), and gamma γ (14 topics \times 226 words). Appendix A shows the matrix of topics *K* with the terms *w*, from the corpus of documents *D*, with the coherence explained by each topic in the model.

For the representation of the topics in the newspapers analyzed by means of the HJ-biplot, a matrix is generated from the matrix θ , obtained in the LDA, since it classifies the topics according to the probability that each of these belongs to each analyzed document. The transposed matrix of theta is generated, to this matrix θ^T , the average of each document \bar{d} is calculated (this equation applies from *i* document to *n* document for each newspaper), which in this matrix represents the possibility that each document *d* belongs to each topic *K*, the average is calculated for each set of newspapers, *D*, thus obtaining the possibility that each topic *K* belongs to the respective newspapers analyzed, this process is described in Figure 7 and Equation (7). That is a representation of the topics *K* is made for each one of the documents *D* (in this case *D* being the newspapers), forming a new matrix *X*.

$$\bar{d}_i = \frac{d_{i1} + d_{i2} \dots + d_{in}}{D} \tag{7}$$

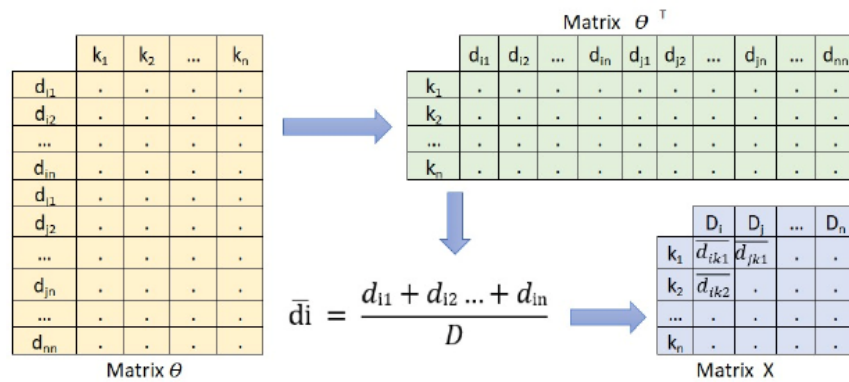


Figure 7. Process of obtaining the X matrix for HJ-biplot.

The interpretation of the HJ-biplot obtained in the present study is shown in Figure 8. Rules like those used in multidimensional scaling (MDS), correspondence analysis, factor analysis, and classical biplot are used. Thus, the length of the markers corresponding to the columns (vectors) approximate the standard deviation of the topics; the cosines of the angles between the markers (vectors) column approximate the correlations between news newspapers. To understand HJ-biplot, let us consider the order of the orthogonal projections of the row markers (points) onto a column marker (vector) approximates the order of the row elements (centers) in that column (the same property holds for the projection of the markers column in the direction defined by a row marker). Acute angles are associated with newspapers with a positive correlation (20 M and El Mundo), whereas obtuse angles indicate negative correlation and right angles indicate variables unrelated (20 M and El Pais, for example). Likewise, the cosines of the angles between the topic markers and the axes (principal components) approximate the correlations between the two. The greater the projection of a point on a vector, the more the center deviates from the mean of the daily news. The distances among row markers are interpreted as an inverse function of their similarities, in such a way that closer markers (topics) are more similar.

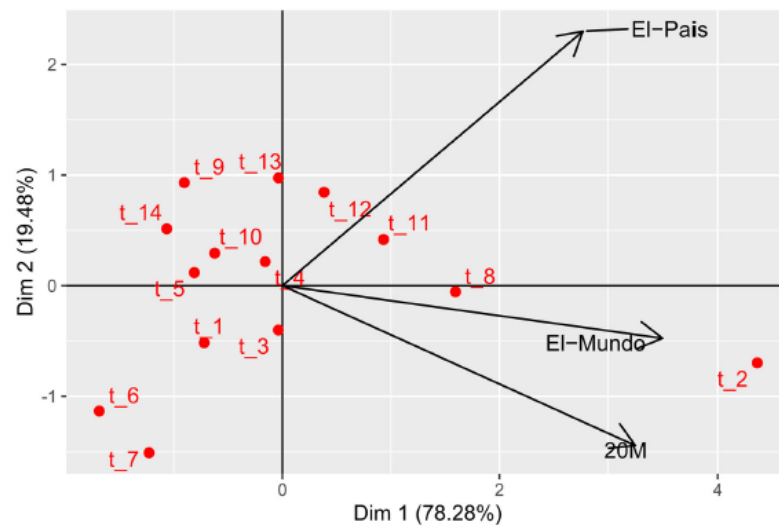


Figure 8. HJ-biplot of topics obtained in LDA.

In the first place, it is observed that—overall—there is a good quality of representation in this first factorial plane, with 97.76% of inertia absorbed or explained. In Table 2, the names of the 14 topics obtained in the LDA model are observed. The topic *t*₁₂ “Government_Madrid” is explained in the headlines of the newspaper El País since the projection of this topic on the vector that represents the newspaper “El País” is much higher than the projection of the same topic. The other two vectors/markers represent El Mundo and 20M, in which this projection is almost null. Although the most explained topic in El Mundo would correspond to *t*₂ “New_Contagions” which also has an explanation in the other two newspapers, although less than in the first. Topics *t*₈ “Dose_Vaccine” and *t*₁₁ “Curfew” are mainly related to the newspapers El País, El Mundo, and 20M. The relationship between newspapers is also observed in this representation. According to the topics covered in the headlines related to COVID, El Mundo and 20 Minutos are perceived as having a high relationship with each other in the topics covered since the angle between the vectors that represent both of markers is smaller. However, El País presents other topics that it delves into in greater detail, since it does not present a good correlation with the other two newspapers, as it presents an angle of practically 90° between the vectors. It is observed how certain topics are related only to some news newspapers; but in general, the three have most of the most similar topics in their publications.

Table 2. Label topics.

Topics	Label Topics	Topics	Label Topics
<i>t</i> ₁	Positive_Cases	<i>t</i> ₈	Dose_Vaccine
<i>t</i> ₂	New_Contagions	<i>t</i> ₉	USA_Pandemic
<i>t</i> ₃	New_Variant	<i>t</i> ₁₀	Wave_Pandemic
<i>t</i> ₄	Third_dose	<i>t</i> ₁₁	Curfew
<i>t</i> ₅	Community_Madrid	<i>t</i> ₁₂	Government_Madrid
<i>t</i> ₆	Vaccines_people	<i>t</i> ₁₃	Government_Millions
<i>t</i> ₇	Measures_Pedro_Sanchez	<i>t</i> ₁₄	House_confinement

4. Discussion

The LDAvis represents the topics in a two-dimensional space without considering the periodicals; in addition, this package visualizes the topics as circles in the two-dimensional plane whose centers are determined by calculating the Jensen–Shannon divergence [60]

between the topics and then using a multidimensional scale to project the distances between subjects in two dimensions. The overall prevalence of each theme is coded using the areas of the circles. Therefore, the results differ from the LDA obtained when executing the `FitLdaModel` function of the `TextminerR` package, where the Gibbs sampling is considered, as well as the values given to the hyperparameters for obtaining the topics and the grouping of the words according to the given weight. By the frequency in each topic, this can also be controlled by the number of iterations assigned to obtain the model.

In the representation of the LDA model obtained by means of a heatmap, the averages of the possibility that each document of a newspaper belongs to a certain topic are observed in a grouped manner, these average values obtained by each newspaper are shown as the possibility that each topical has a higher value of the frequency of belonging to a particular newspaper. Additionally, it can be parameterized so that the newspapers are displayed in a grouped manner, the distance considered in this method is the Euclidean [61], and in this representation, the row markers (topics) are given greater representativeness.

Biplot techniques are based on the same principles on which most dimensionality reduction factorial techniques are based. The fundamental difference is that a joint representation of rows and columns is incorporated, unlike principal component analysis which reduces the column data to a smaller number of components seeking to explain as much of the total variance in the variables as possible by calculating the components as linear combinations of the original variables [62], or unlike analysis factorial of correspondence that is used when there is a significant association between the categorical variables studied, representing the rows and columns of the contingency table in two reduced vector spaces, to later superimpose them and obtain the joint representation of both [63].

The HJ-biplot method, which is presented as an alternative for the representation of the results obtained in an LDA content modeling, can obtain a high-quality representation simultaneously in the row markers (topics) and column markers (newspapers), enabling the study of the correlation between documents and visualizing the topics according to the corpus of the document with which it has the greatest representativeness. The distance between the row markers (topics) enables the identification of clusters of individuals with similar profiles. Any hierarchical or non-hierarchical clustering technique can be used to help identify relevant clusters.

5. Conclusions

The topic model is an unsupervised method applied to text mining. In this study, it was applied to news from digital newspapers, where the HJ-biplot is presented as a new option to visualize newspapers and topics with the highest quality of representation, which is not possible with other traditional biplot models. Two comparative methods of representation of the LDA model were used: the heatmap represents the topics with better quality and the LDAv does not consider the newspapers within its multidimensional scaling representation, which does not allow exploration of the possible relationships that the topic has with the newspapers. Therefore, it is not possible to observe which topic contributes to a newspaper, unlike with the HJ-biplot and the heatmap which allow this analysis, but in a different way between them.

It is recommended that the effect of applying different methods of selecting and extracting data from the web be explored, as well as applying other methods associated with LDA to obtain the topics, and even applying machine learning methods for the representation in the HJ-biplot of the topics and digital newspapers.

Author Contributions: Conceptualization, L.P.-B., P.G.-V. and F.D.-Á.; methodology, L.P.-B.; validation, L.P.-B., P.G.-V. and F.D.-Á.; formal analysis, L.P.-B.; investigation, L.P.-B.; data curation, L.P.-B.; writing—original draft preparation, L.P.-B.; writing—review and editing, L.P.-B., P.G.-V. and F.D.-Á.; supervision, P.G.-V. and F.D.-Á. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not required.

Informed Consent Statement: Not required.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/Pilacuan-Bonete-Luis/Data_HJ-Biplot_newspapers/blob/main/Data_newspapers.xlsx (accessed on 27 May 2022).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

In Table A1: each topic is observed with its respective group of words obtained in the LDA model, as well as the coherence which gives us an idea of how coherent a model is in terms of the distribution of its topics. The more different the words of the topics are among themselves, the less related the topics will be and the better coherence the model will have. As for prevalence, which is a measure of how sensitive the model is when data that are not part of it are added before, the lower the prevalence value is the better the model is.

Table A1. Topics, label topics, coherence, prevalence, and top term of the model LDA.

Topics	Label Topics	Coherence	Prevalence	Top Terms
t_1	Positive_Cases	0.046	6.533	positive, gives, dies, months, years, days, three, quarantine, give, after, hospital, test, first, four, Madrid, six, five, UCI, pandemic, days, home, ago, case, life, trump, virus, anus, seven, fear, years, weeks, people, health, Barcelona, mask, patients, USA, returns, PCR, less
t_2	New_Contagions	0.138	11.359	cases, new, infections, incidence, health, deaths, registered, deceased, hours, Spain, new_cases, week, dead, positive, sum, notifies, UCI, Balearic, islands, low, new_infections, day, Madrid, last, figure, last_hours, continues, exceeds, Cantabria, Euskadi, pandemic, almost, hospitalized, last, income, increase, four, Spain, risk, three, decrease
t_3	New_Variant	0.166	7.178	new, variant, omicron, kingdom, united, united_kingdom, span, Europe, virus, normality, who, quarantine, Johnson, first, Spain, case, restrictions, confinement, pandemic, country, vaccinated, asi, Germany, alert, risk, puts, EU, vaccines, France, Italy, new, test, vaccine, wave, cases, return, tourism, united, returns, experts
t_4	Third_dose	0.021	6.919	residences, Madrid, hospital, outbreak, health, residence, elderly, hospitals, elderly, health, patients, centers, infected, test, positive, people, health, three, new, workers, UCI, virus, Valencia, masks, missing, PCR, pandemic, four, Generalitat, less, dead, leave, health, board, Catalonia, Barcelona, get vaccinated, Ayuso, ask, death
t_5	Community_Madrid	0.205	6.359	community, Madrid, Valenciano, valencia_community, madrid_community, phase, Monday, test, de-escalation, health, week, closure, Generalitat, leave, restrictions, cvirus, government, passport, vaccination, Catalonia, Ayuso, health, residences, PCR, may, masks, Barcelona, requests, infections, measures, people, mask, pandemic, centers, vaccination, hospitals, bars, areas, leisure, Spain

Table A1. Cont.

Topics	Label Topics	Coherence	Prevalence	Top Terms
t_6	Vaccines_people	0.013	5.8	vaccines, people, pandemic, millions, Spain, complete, vaccination, world, population, like this, year, less, rioja, almost, front, half, first, exceeds, vaccine, greater, EU, Galicia, ago, country, Europe, months, Spain, study, virus, six, crisis, summer, three, great, children, dose, USA, vaccinated, years, five
t_7	Measures_Pedro_Sanchez	0.146	6.272	Sanchez, cases, active, Pedro, active_cases, Pedro_Sanchez, day, Galicia, rioja, infections, alarm, days, deceased, decrease, increase, municipalities, new, new, Sanchez, government, positive, new_contagions, week, pp, hospitalized, Spain, requests, exceeds, less, announces, citizens, low, cvirus, crisis, continues, centers, front, maintains, pandemic, plan
t_8	Dose_Vaccine	0.121	8.65	vaccine, dose, years, third, older, vaccination, vaccines, AstraZeneca, third_dose, Pfizer, years, children, health, vaccine_dose, vaccinate, minors, people, Madrid, second, first, vaccination, study, vaccinated, get vaccinated, open, Spain, young people, millions, front, EU, population, residences, week, risk, less, USA, leave, health, experts, months
t_9	USA_Pandemic	0.114	6.2	pandemic, China, Mexico, USA, USA, uu, city, who, world, trump, vaccine, virus, united, new, new, alert, crisis, vaccination, first, great, vaccines, Johnson, year, millions, Europe, case, USA, how, outbreak, health, variant, major, ask, medium, announce, normality, home, health, government, EU
t_10	Wave_Pandemic	0.066	6.51	pandemic, wave, first, second, time, risk, Spain, virus, greater, worse, Spain, contagion, infections, world, middle, year, ICU, life, crisis, death, third, year, fear, new, confinement, new, alert, low, Europe, less, incidence, front, hospitals, asi, month, study, week, health, health, China
t_11	Curfew	0.276	7.89	restrictions, Madrid, curfew, touch, touch, stay, measures, closure, new, confinement, people, government, Catalonia, bars, leisure, request, municipalities, passport, areas, infections, Barcelona, week, Catalonia, Andalusia, Christmas, close, alarm, health, communities, mask, curb, Generalitat, meeting, avoid, start, Monday, three, healthcare, weigh, six, maintain
t_12	Government_Madrid	0.027	7.316	government, Madrid, Ayuso, ask, pp, crisis, measures, alarm, masks, pandemic, Sanchez, citizens, plan, front, ask, Sanchez, confinement, communities, de-escalation, health, avoid, test, says, new, virus, now, health, mask, EU, phase, lack, residences, sanitary, return, weigh, announce, Christmas, sanitary, put, data

Table A1. Cont.

Topics	Label Topics	Coherence	Prevalence	Top Terms
t_13	Government_Millions	0.023	6.925	millions, government, pandemic, euros, crisis, companies, erte, aid, workers, tourism, sector, plan, Spain, masks, Barcelona, year, summer, less, announce, almost, request, Generalitat, Madrid, year, test, health, board, work, month, middle, half, front, first, may, EU, health, new, Sanchez, leave, Spain
t_14	House_confinement	0.004	6.09	home, confinement, Madrid, quarantine, children, mask, how, students, day, homecoming, course, street, masks, alarm, first, today, Spain, work, pandemic, so, Monday, leave, may, government, less, Catalonia, Barcelona, de-escalation, now, children, days, phase, restrictions, life, avoid, fear, France, close, day, normality

References

- He, W.; Zha, S.; Li, L. Social Media Competitive Analysis and Text Mining: A Case Study in the Pizza Industry. *Int. J. Inf. Manag.* **2013**, *33*, 464–472. [CrossRef]
- Alalwan, A.A.; Rana, N.P.; Dwivedi, Y.K.; Algharabat, R. Social Media in Marketing: A Review and Analysis of the Existing Literature. *Telemat. Inform.* **2017**, *34*, 1177–1190. [CrossRef]
- Pejic-Bach, M.; Bertoncel, T.; Meško, M.; Krstić, Ž. Text Mining of Industry 4.0 Job Advertisements. *Int. J. Inf. Manag.* **2020**, *50*, 416–431. [CrossRef]
- De la Hoz-M, J.; Fernández-Gómez, M.J.; Mendes, S. LDAShiny: An R Package for Exploratory Review of Scientific Literature Based on a Bayesian Probabilistic Model and Machine Learning Tools. *Mathematics* **2021**, *9*, 1671. [CrossRef]
- Slobodin, O.; Plohotnikov, I.; Cohen, I.-C.; Elyashar, A.; Cohen, O.; Puzis, R. Global and Local Trends Affecting the Experience of US and UK Healthcare Professionals during COVID-19: Twitter Text Analysis. *Int. J. Environ. Res. Public Health* **2022**, *19*, 6895. [CrossRef]
- WHO. *COVID-19 Weekly Epidemiological Update*; WHO: Geneva, Switzerland, 2022.
- Zhu, N.; Zhang, D.; Wang, W.; Li, X.; Yang, B.; Song, J.; Zhao, X.; Huang, B.; Shi, W.; Lu, R.; et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N. Engl. J. Med.* **2020**, *382*, 727–733. [CrossRef]
- Brüssow, H. The Novel Coronavirus—A Snapshot of Current Knowledge. *Microb. Biotechnol.* **2020**, *13*, 607–612. [CrossRef]
- McKibbin, W.J.; Fernando, R. The Global Macroeconomic Impacts of COVID-19: Seven Scenarios. *SSRN Electron. J.* **2020**, *20*, 1–30. [CrossRef]
- 20Minutos. ¿Cuál Fue El Primer Caso de Coronavirus en España y en La Península? Available online: <https://www.20minutos.es/noticia/4186871/0/coronavirus-primer-caso-espana-peninsula/> (accessed on 15 April 2020).
- Estudio General de Medios Ranking de Diarios. Available online: <http://reporting.aimc.es/index.html#/main/diarios> (accessed on 16 April 2020).
- Mutz, D.C.; Soss, J. Reading Public Opinion: The Influence of News Coverage on Perceptions of Public Sentiment. *Public Opin. Q.* **1997**, *61*, 431. [CrossRef]
- Hoffman, L.H.; Glynn, C.J.; Huges, M.E.; Sietman, R.B.; Thomson, T. The Role of Communication in Public Opinion Processes: Understanding the Impacts of Intrapersonal, Media, and Social Filters. *Int. J. Public Opin. Res.* **2007**, *19*, 287–312. [CrossRef]
- Peretti, P.O.; Lucas, C. Newspaper Advertising Influences on Consumers' Behavior by Socioeconomic Status of Customers. *Psychol. Rep.* **1975**, *37*, 693–694. [CrossRef]
- Thornton, J.A.; Wahl, O.F. Impact of a Newspaper Article on Attitudes toward Mental Illness. *J. Community Psychol.* **1996**, *24*, 17–25. [CrossRef]
- Baumgartner, R.; Gatterbauer, W.; Gottlob, G. Web Data Extraction System. *Encycl. Database Syst.* **2009**, *1*, 3465–3471.
- Collobert, R.; Weston, J.; Com, J.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P. Natural Language Processing (Almost) from Scratch. *J. Mach. Learn. Res.* **2011**, *12*, 2493–2537. [CrossRef]
- Jones, K.S. A Statistical Interpretation of Term Specificity and Its Application in Retrieval. *J. Doc.* **1972**, *28*, 11–21. [CrossRef]
- Deerwester, S.; Harshman, R.; Susan, T.; George, W.; Thomas, K. Indexing by Latent Semantic Analysis. *J. Am. Soc. Inf. Sci.* **1990**, *41*, 391–407. [CrossRef]
- Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022. [CrossRef]
- Aldjanabi, W.; Dahou, A.; Al-Qaness, M.A.A.; Elaziz, M.A.; Helmi, A.M.; Damaševičius, R. Arabic Offensive and Hate Speech Detection Using a Cross-Corpora Multi-Task Learning Model. *Informatics* **2021**, *8*, 69. [CrossRef]
- Hadwan, M.; Al-Sarem, M.; Saeed, E.; Al-Hagery, M.A. An Improved Sentiment Classification Approach for Measuring User Satisfaction toward Governmental Services' Mobile Apps Using Machine Learning Methods with Feature Engineering and SMOTE Technique. *Appl. Sci.* **2022**, *12*, 5547. [CrossRef]

23. Sievert, C.; Shirley, K.E. LDAvis: A Method for Visualizing and Interpreting Topics. In Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, Baltimore, MD, USA, 27 June 2014; Association for Computational Linguistics: Stroudsburg, PA, USA, 2014; pp. 63–70.
24. Raivo Kolde. cran.r-project.org/package=pheatmap— Pheatmap: Pretty heatmaps. R Package Heatmap version 2.12.00. 2019. Available online: <https://cran.r-project.org/web/packages/pheatmap/index.html/> (accessed on 15 April 2022).
25. Galindo-Villardón, P. Una Alternativa de Representación Simultánea: HJ-Biplot (An Alternative of Simultaneous Representation: HJ-Biplot). *Questio* 1986, 10, 13–23.
26. Díaz-Faes, A.A.; González-Albo, B.; Galindo, M.P.; Bordons, M. HJ-Biplot Como Herramienta de Inspección de Matrices de Datos Bibliométricos. *Revista Española Documentación Científica* 2013, 36, e001. [CrossRef]
27. Julia, D.C.; Galindo, P.V.; Villardón, M.P.G. Grupos de Discusión y HJ-Biplot: Una Nueva Forma de Análisis Textual. *Revista Ibérica Sistemas Tecnologías Información* 2014, E2, 19–35. [CrossRef]
28. Zulaima, O.M. *Contribuciones al Análisis de Datos Textuales*; Universidad de Salamanca: Salamanca, Spain, 2006.
29. Han, J.; Kamber, M.; Pei, J. *Data Mining: Concepts and Techniques*, 3rd ed.; Elsevier Inc.: Amsterdam, The Netherlands, 2012; ISBN 9780123814791.
30. Fayyad, U.; Stolorz, P. Data Mining and KDD: Promise and Challenges. *Futur. Gener. Comput. Syst.* 1997, 13, 99–115. [CrossRef]
31. Alyahyan, E.; Düşteğör, D. Predicting Academic Success in Higher Education: Literature Review and Best Practices. *Int. J. Educ. Technol. High. Educ.* 2020, 17, 3. [CrossRef]
32. The R Foundation R 2020. Available online: <https://www.r-project.org/> (accessed on 1 May 2021).
33. Manning, C.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*; Cambridge University Press: Cambridge, UK, 2008; ISBN 9780511809071.
34. Markov, Z.; Larose, D.T. *Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage*; John Wiley & Sons: Hoboken, NJ, USA, 2007; ISBN 0470108088.
35. Kamath, S.S.; Bagalkotkar, A.; Khandelwal, A.; Pandey, S.; Poomima, K. Sentiment Analysis Based Approaches for Understanding User Context in Web Content. In Proceedings of the 2013 International Conference on Communication Systems and Network Technologies, CSNT 2013, Gwalior, India, 6–8 April 2013; pp. 607–611.
36. Catanese, S.A.; De Meo, P.; Ferrara, E.; Fiumara, G.; Provetti, A. Crawling Facebook for Social Network Analysis Purposes. In Proceedings of the International Conference on Web Intelligence, Mining and Semantics, Sogndal, Norway, 25–27 May 2011; ACM Press: New York, NY, USA, 2011; p. 1.
37. Chandler, J.D.; Salvador, R.; Kim, Y. Language, Brand and Speech Acts on Twitter. *J. Prod. Brand Manag.* 2018, 27, 375–384. [CrossRef]
38. Plake, C.; Schiemann, T.; Pankalla, M.; Hakenberg, J.; Leser, U. ALIBABA: PubMed as a Graph. *Bioinformatics* 2006, 22, 2444–2445. [CrossRef]
39. Xie, X.; Fu, Y.; Jin, H.; Zhao, Y.; Cao, W. A Novel Text Mining Approach for Scholar Information Extraction from Web Content in Chinese. *Futur. Gener. Comput. Syst.* 2019, 111, 859–872. [CrossRef]
40. Schedlbauer, J.; Raptis, G.; Ludwig, B. Medical Informatics Labor Market Analysis Using Web Crawling, Web Scraping, and Text Mining. *Int. J. Med. Inform.* 2021, 150, 104453. [CrossRef]
41. Rekik, R.; Kallel, I.; Casillas, J.; Alimi, A.M. Assessing Web Sites Quality: A Systematic Literature Review by Text and Association Rules Mining. *Int. J. Inf. Manag.* 2018, 38, 201–216. [CrossRef]
42. Zhao, Y. *R and Data Mining: Examples and Case Studies*; Academic Press: Cambridge, MA, USA; Elsevier: Amsterdam, The Netherlands, 2012; ISBN 9780123969637.
43. Khalil, S.; Fakir, M. RCrawler: An R Package for Parallel Web Crawling and Scraping. *SoftwareX* 2017, 6, 98–106. [CrossRef]
44. Wickham Hadley Easily Harvest (Scrape) Web Pages 2019. Available online: <https://rvest.tidyverse.org/> (accessed on 1 May 2021).
45. Salton, G.; Buckley, C. Term-Weighting Approaches in Automatic Text Retrieval. *Inf. Process. Manag.* 1988, 24, 513–523. [CrossRef]
46. Aizawa, A. An Information-Theoretic Perspective of Tf-Idf Measures. *Inf. Process. Manag.* 2003, 39, 45–65. [CrossRef]
47. Luhn, H.P. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM J. Res. Dev.* 1957, 1, 309–317. [CrossRef]
48. Thomas, J. Función TermDocFreq | RDocumentation 2019. Available online: <https://www.rdocumentation.org/packages/textmineR/versions/3.0.4/topics/TermDocFreq> (accessed on 1 May 2021).
49. Tommy, J.; William, D. Functions for Text Mining and Topic Modeling 2019. Available online: <https://www.rtextminer.com/> (accessed on 1 May 2021).
50. Blei, D.M.; Lafferty, J.D. Topic Models. In *Text Mining: Classification, Clustering, and Applications*; Taylor & Francis Group, Ed.; Chapman and Hall/CRC: New York, NY, USA, 2009; pp. 71–82. ISBN 9780429191985.
51. Steyvers, M.; Griffiths, T. Probabilistic Topic Models. In *Handbook of Latent Semantic Analysis*; Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, W., Eds.; Lawrence Erlbaum: Mahwah, NJ, USA, 2006; pp. 427–448. ISBN 1135603286.
52. Jones, T. A Coefficient of Determination for Probabilistic Topic Models. *arXiv* 2019, arXiv:1911.11061. [CrossRef]
53. Rosner, F.; Hinneburg, A.; Röder, M.; Nettling, M.; Both, A. Evaluating Topic Coherence Measures. *arXiv* 2014, arXiv:1403.6397. [CrossRef]

54. Gabriel, K.R. The Biplot Graphic Display of Matrices with Application to Principal Component Analysis. *Biometrika* **1971**, *58*, 453–467. [CrossRef]
55. Eckart, C.; Young, G. The Approximation of One Matrix by Another of Lower Rank. *Psychometrika* **1936**, *1*, 211–218. [CrossRef]
56. Eckart, C.; Young, G. A Principal Axis Transformation for Non-Hermitian Matrices. *Bull. Am. Math. Soc.* **1939**, *45*, 118–121. [CrossRef]
57. Frutos, E.; Galindo, M.P. cran.r-project.org/package=GGEbiplotGUI. GGEbiplotGUI 2016. Available online: <https://cran.r-project.org/web/packages/GGEbiplotGUI/index.html> (accessed on 1 May 2021).
58. Frutos, E.; Galindo, M.P.; Leiva, V. An Interactive Biplot Implementation in R for Modeling Genotype-by-Environment Interaction. *Stoch. Environ. Res. Risk Assess.* **2014**, *28*, 1629–1641. [CrossRef]
59. Bostock, M.; Ogievetsky, V.; Heer, J. D3 Data-Driven Documents. *IEEE Trans. Vis. Comput. Graph.* **2011**, *17*, 2301–2309. [CrossRef]
60. Lin, J. Divergence Measures Based on the Shannon Entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151. [CrossRef]
61. Zuguang, G. Packages ComplexHeatmap. 2021. Available online: <https://www.bioconductor.org/packages/release/bioc/html/ComplexHeatmap.html> (accessed on 1 May 2021).
62. Pearson, K. LIII. On Lines and Planes of Closest Fit to Systems of Points in Space. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1901**, *6*, 559–572. [CrossRef]
63. Benzécri, J.-P. *L'analyse Des Données. Tomo I: La Taxonomie*; Dunod: Paris, France, 1973; Volume 2, ISBN 2040071539.

BIBLIOGRAFÍA

REFERENCIAS.

- Adamson, G. W., & Boreham, J. (1974). The use of an association measure based on character structure to identify semantically related pairs of words and document titles. *Information Storage and Retrieval*, 10(7-8), 253-260. [https://doi.org/10.1016/0020-0271\(74\)90020-5](https://doi.org/10.1016/0020-0271(74)90020-5)
- Anandarajan, M., Hill, C., & Nolan, T. (2019). *Practical Text Analytics Maximizing the Value of Text Data Advances in Analytics and Data Science (Vol. 2)*. Springer. <http://www.springer.com/series/15876>
- Ander-Egg, E. (2000). *Diccionario de trabajo social*.
- Ao, Z., Horváth, G., Sheng, C., Song, Y., & Sun, Y. (2023). Skill requirements in job advertisements: A comparison of skill-categorization methods based on wage regressions. *Information Processing & Management*, 60(2), 103185. <https://doi.org/10.1016/J.IPM.2022.103185>
- Arco, L. (2008). Agrupamiento basado en la intermediación diferencial y su valoración utilizando la teoría de los conjuntos aproximados [Universidad «Martha Abreu» las Villas]. <http://dspace.uclv.edu.cu:8089/handle/123456789/12470>
- Asmussen, C. B., & Møller, C. (2019a). Smart literature review: a practical topic modelling approach to exploratory literature review. *Journal of Big Data*, 6(1), 1-18. <https://doi.org/10.1186/S40537-019-0255-7/TABLES/6>
- Asmussen, C. B., & Møller, C. (2019b). Smart literature review: a practical topic modelling approach to exploratory literature review. *Journal of Big Data*, 6(1), 1-18. <https://doi.org/10.1186/S40537-019-0255-7/TABLES/6>
- Asmussen, C. B., & Møller, C. (2019c). Smart literature review: a practical topic modelling approach to exploratory literature review. *Journal of Big Data*, 6(1), 1-18. <https://doi.org/10.1186/S40537-019-0255-7/TABLES/6>
- Asuncion, A., Welling, M., Smyth, P., & Teh, Y. W. (2012). On Smoothing and Inference for Topic Models. *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, UAI 2009, 27-34. <https://doi.org/https://doi.org/10.48550/arXiv.1205.2662>
- Attias, H. (1999). A Variational Bayesian Framework for Graphical Models. *Advances in Neural Information Processing Systems*, 12.
- Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. ACM Press. <https://www.researchgate.net/profile/Ricardo-Baeza->

Yates/publication/2352627_Modern_Information_Retrieval/links/54c128a80cf2d03405c4eb60/Modern-Information-Retrieval.pdf

- Ballesteros, V. I. (2022). Análisis multivariante de los estilos de aprendizaje, estilos de pensamiento e inteligencia emocional, en estudiantes de nivel medio y superior. Universidad de Salamanca.
- Bauer, L. (1983). *English Word-Formation* (B. Comrie & C. J. Fillmore, Eds.). Cambridge University Press.
- Beigbeder, F. (2006). *Diccionario técnico: inglés-español, español-inglés* (2.^a ed.). Ediciones Díaz de Santos.
- Benoit, K., Muhr, D., & Watanabe, K. (2021). stopwords (2.3). CRAN de R. <https://cran.r-project.org/web/packages/stopwords/index.html>
- Benzécri, J.-P. (1964). *Cours de linguistique mathématique*.
- Benzécri, J.-P. (1973). L'analyse des données. En *L'Analyse des Correspondances*.
- Berry, M. W., Drmač, Z., & Jessup, E. R. (1999a). Matrices, vector spaces, and information retrieval. *SIAM Review*, 41(2). <https://doi.org/10.1137/S0036144598347035>
- Berry, M. W., Drmač, Z., & Jessup, E. R. (1999b). Matrices, Vector Spaces, and Information Retrieval. *Society for Industrial and Applied Mathematics*, 41(2), 335-362. <http://www.siam.org/journals/sirev/41-2/34703.html>
- Berry, M. W., Dumais, S. T., & O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. *SIAM Review*, 37(4). <https://doi.org/10.1137/1037127>
- Berry, M. W., & Kogan, J. (2010). *Text Mining: Applications and Theory* (First).
- Billheimer, D., Booker, A. J., Condliff, M. K., Greaves, M. T., Holt, F. B., Kao, A. S.-W., Pierce, D. J., Potteet, S. R., & Wu, Y.-J. (2003). Method and system for text mining using multidimensional subspaces (Patent N.º US 6,611,825 B1). United States Patent.
- Blei, D., Carin, L., & Dunson, D. (2010). Probabilistic topic models. *IEEE Signal Processing Magazine*, 27(6), 55-65. <https://doi.org/10.1109/MSP.2010.938079>
- Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77-84. <https://doi.org/10.1145/2133806.2133826>
- Blei, D. M., & Lafferty, J. D. (2005). Correlated Topic Models. *Advances in Neural Information Processing Systems*, 18. www.jstor.org
- Blei, D. M., & Lafferty, J. D. (2009). Topic Models. En *taylor & Francis Group* (Ed.), *Text mining: Classification, Clustering, And Applications* (1 st Editi, pp. 71-82). Chapman and Hall/CRC. <https://doi.org/https://doi.org/10.1201/9781420059458>

- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2001). Latent Dirichlet Allocation. *Advances in Neural Information Processing Systems*, 14.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2002). Latent Dirichlet allocation. *Advances in Neural Information Processing Systems*, 14(1-2), 601-608. <https://doi.org/10.1162/jmlr.2003.3.4-5.993>
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(4-5), 993-1022. <https://doi.org/10.1016/b978-0-12-411519-4.00006-9>
- Bouchet-Valat, M. (2020). Snowball Stemmers Based on the C «libstemmer». <https://cran.r-project.org/web/packages/SnowballC/SnowballC.pdf>
- Boyes, R. (2020). shinylda. GitHub. <https://github.com/rboyes/shinylda>
- Bradley, P., Mangasarian, O., & Street, W. (1996). Clustering via Concave Minimization. *Advances in Neural Information Processing Systems*, 9.
- Caballero-Julia, D., & Campillo, P. (2021). Epistemological Considerations of Text Mining: Implications for Systematic Literature Review. *Mathematics* 2021, Vol. 9, Page 1865, 9(16), 1865. <https://doi.org/10.3390/MATH9161865>
- Caballero-Julia, D., Vicente-Galindo, P., & Galindo-Villardón, P. (2014). Grupos de discusión y HJ-Biplot: Una nueva forma de análisis textual. *RISTI - Revista Ibérica de Sistemas e Tecnologías de Información*, 2, 19-35. <https://doi.org/10.17013/RISTI.E2.19-35>
- Cabanelas, G. (2003). *Diccionario Jurídico Elemental*. Heliasta.
- Camacho B, A., & Ariosa R, L. (2000). *Diccionario de términos Ambientales* (E. Hernández, Ed.; 1.^a ed.). Ediciones Acuario.
- Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, 72(7-9), 1775-1781. <https://doi.org/10.1016/J.NEUCOM.2008.06.011>
- Cavieres Abarca, A., Fredes Mena, S., Ramires Novoa, A., Castillo Guerrero, R., & Gómez Fuentes, H. (2010). Tesoros y Web Semántica: diseño metodológico para estructurar contenidos Web mediante SKOS-Core. *Serie Bibliotecología y Gestión de Información*, 57.
- Chang, J. (2015). lda: Métodos de muestreo de Gibbs contraídos para modelos temáticos (1.4.2). CRAN de R. <https://cran.r-project.org/web/packages/lda/index.html>

- Chang, W., Cheng, J., Allaire, Sievert, C., Schloerke, B., Xie, Y., & Allen, J. (2022). shiny: Web Application Framework for R. Comprehensive R Archive Network (CRAN). <https://CRAN.R-project.org/package=shiny>
- Chellappandi, P., & Vijayakumar, C. S. (2018). Bibliometrics, Scientometrics, Webometrics/Cybermetrics, Informetrics and Altmetrics -- An Emerging Field in Library and Information Science Research. *Shanlax International Journal of Education*, 7(1), 5-8. <https://doi.org/10.5281/zenodo.2529398>
- Choueka, Y. (1980). Computerized full-text retrieval systems and research in the humanities: The responsa project. *Computers and the Humanities*, 14(3), 153-169. <https://doi.org/10.1007/BF02403764/METRICS>
- Clark, S. (2014). Vector Space Models of Lexical Meaning. En S. Lappin & C. Fox (Eds.), *Handbook of Contemporary Semantics* (second, pp. 1-43).
- Collins, A. M., & Loftus, E. F. (1975). A Spreading-Activation Theory of Semantic Processing. *Psychological Review*, 82(6), 407-428.
- Cosacov, E. (2007). *Diccionario de términos técnicos de la Psicología* (3.^a ed.). Editorial Brujas.
- Cubilla Montilla, M. I. (2019). *Contribuciones al Análisis Biplot basadas en Soluciones Factoriales Disjuntas y en Soluciones Sparse*. Universidad de Salamanca.
- Darling, W. (2011). A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 642-647.
- De La Hoz Maestre, J. (2020). *Revisión exploratoria de literatura científica en acuicultura: Análisis de tendencia utilizando un modelo probabilístico bayesiano y herramientas de machine learning*. Universidad de Salamanca.
- De La Hoz-M, J., Fernández-Gómez, M., & Méndez, S. (2021). CRAN - Paquete LDAShiny (0.9.3). CRAN de R. <https://cran.r-project.org/web/packages/LDAShiny/index.html>
- de Sousa, G. C., & Castañeda-Ayarza, J. A. (2022). PESTEL analysis and the macro-environmental factors that influence the development of the electric and hybrid vehicles industry in Brazil. *Case Studies on Transport Policy*, 10(1), 686-699. <https://doi.org/10.1016/J.CSTP.2022.01.030>
- Deerwester, S., Susan, T. D., George, W. F., Thomas, K. L., & Richard, H. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391-407. [https://doi.org/https://doi.org/10.1002/\(SICI\)1097-4571\(199009\)41:6<391::AID-ASII>3.0.CO;2-9](https://doi.org/https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASII>3.0.CO;2-9)

- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1), 1-22. <https://doi.org/10.1111/J.2517-6161.1977.TB01600.X>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171-4186. <https://doi.org/https://doi.org/10.48550/arXiv.1810.04805>
- Díez, J. A., & Moulines, C. U. (1997). *Fundamentos de filosofía de la ciencia* (1.^a ed., Vol. 1). Editorial Ariel.
- Dong, A., & Agogino, A. M. (1997). Text analysis for constructing design representations. *Artificial Intelligence in Engineering*, 11(2), 65-75. [https://doi.org/10.1016/S0954-1810\(96\)00036-2](https://doi.org/10.1016/S0954-1810(96)00036-2)
- Doyle, B. L. (1961). Semantic road maps for literature searchers. *Journal of the ACM*, 8(4), 553-578. <https://doi.org/10.2964/JSIK.18-5-7>
- Dumais, S. T. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, & Computers*, 23(2), 229-236.
- Dumais, S. T. (1992). Enhancing Performance in Latent Semantic Indexing (LSI) Retrieval.
- Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., & Harshman, R. (1988). Using latent semantic analysis to improve access to textual information. *Conference on Human Factors in Computing Systems - Proceedings, Part F130202*, 281-285. <https://doi.org/10.1145/57167.57214>
- Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika* 1936 1:3, 1(3), 211-218. <https://doi.org/10.1007/BF02288367>
- Egido, J. (2020). *dynBiplotGUI* (1.1.6; pp. 1-15). CRAN de R.
- El-Hamdouchi, A., & Willett, P. (1989). Comparison of Hierarchic Agglomerative Clustering Methods for Document Retrieval. *The Computer Journal*, 32(3), 220-227. <https://doi.org/10.1093/COMJNL/32.3.220>
- Fahey, L., & Narayanan, V. K. (1968). Macroenvironmental analysis for strategic management. En (No Title). St. Paul Minn. West.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17(3), 37-37. <https://doi.org/10.1609/AIMAG.V17I3.1230>

- Feldman, R., & Dagan, I. (1995). Knowledge Discovery in Textual Databases (KDT). Proceeding of the first International Conference on Knowledge Discovery and data mining, 112-117. <https://www.researchgate.net/publication/2781984>
- Fernandes, J. P. (2019). Developing viable, adjustable strategies for planning and management—A methodological approach. *Land Use Policy*, 82, 563-572. <https://doi.org/10.1016/J.LANDUSEPOL.2018.12.044>
- Fernandes, J. P. A., & Guiomar, N. (2016). Environmental Ethics: Driving Factors Beneath Behavior, Discourse and Decision-Making. *Journal of Agricultural and Environmental Ethics*, 29(3), 507-540. <https://doi.org/10.1007/S10806-016-9607-X/TABLES/1>
- Foltz, P. W. (1996). Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments, & Computers*, 28(2), 197-202. <https://doi.org/https://doi.org/10.3758/BF03204765>
- Fridolin, W. (2022). CRAN Task View: Natural Language Processing. <https://cran.r-project.org/web/views/NaturalLanguageProcessing.html>
- Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3), 453-467. <https://doi.org/10.1093/biomet/58.3.453>
- Gabriel, K. R. (1981). Biplot display of multivariate matrices for inspection of data and diagnosis. En V. Barnett (Ed.), *Interpreting Multivariate Data* (pp. 147-173). John Wiley & Sons.
- Gabriel, K. R., & Odoroff, C. L. (1990). Biplots in biomedical research. *Statistics in Medicine*, 9(5), 469-485. <https://doi.org/10.1002/SIM.4780090502>
- Galindo-Villardón, P. (1986). Una alternativa de representación simultánea: HJ-Biplot (An alternative of simultaneous representation: HJ-Biplot). *Questío*, 10(1), 13-23. <https://diarium.usal.es/pgalindo/files/2012/07/Questiio.pdf>
- Galindo-Villardón, P., & Cuadras, C. M. (1986). Una extensión del método Biplot y su relación con otras técnicas. *Publicaciones de Bioestadística y Biomatemática*, 17.
- Galindo-Villardón, P., & Egado, J. (2009). Estudio Multivariante de las Características del Turista Español que Visita México. En S. F. Juárez & M. M. Ojeda (Eds.), *Memoria II Encuentro de Biometría y la V Reunión de la Región Centroamericana y del Caribe de la Sociedad Internacional de Biometría* (pp. 145-150).
- García, G., Contreras, P., & Martínez, V. (2016). Índice del diccionario Constitucional Chileno. En *Diccionario Constitucional Chileno*. Hueders.

- Gayo-Avello, D., Álvarez-Gutiérrez, D., & Gayo-Avello, J. (2004). Naïve algorithms for keyphrase extraction and text summarization from a single document inspired by the protein biosynthesis process. *Lecture Notes in Computer Science*, 3141, 440-455. https://doi.org/10.1007/978-3-540-27835-1_32
- Geman, S., & Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6), 721-741. <https://doi.org/10.1109/TPAMI.1984.4767596>
- Gil Pascual, J. (2021). *Minería de texto con R. Aplicaciones y técnicas estadísticas de apoyo*. UNED.
- Gillespie, A. (2014). *Foundations of Economics* (3.^a ed.). Oxford University Press.
- Golub, G., & Van Loan, C. (1996). *Matrix computations* (3rd ed.). Johns Hopkins University Press. <https://dl.acm.org/doi/book/10.5555/248979>
- Good, I. J. (1965). The Estimation of Probabilities. An essay on modern Bayesian Methods. En *Biométrical Journal* (1era ed., Número 1). M.I.T. Press. <https://doi.org/10.1002/BIMJ.19680100118>
- Gower, J. (1995). A General Theory of Biplots. En W. J. Krzanowski (Ed.), *Recent Advances in Descriptive Multivariate Statistics* (pp. 283-303). Oxford University Press.
- Greenacre, M. J. (2012). Biplots: the joy of singular value decomposition. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(4), 399-406. <https://doi.org/10.1002/WICS.1200>
- Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl_1), 5228-5235. <https://doi.org/10.1073/pnas.0307752101>
- Grimmer, J. (2010). A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases. *Political Analysis*, 18(1), 1-35. <https://doi.org/10.1093/PAN/MPP034>
- Grün, B., Hornik, K., & Blei, D. (2023). *topicmodels* (0.2-14). CRAN de R. <https://cran.r-project.org/web/packages/topicmodels/index.html>
- Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 32(18), 2847-2849. <https://doi.org/10.1093/BIOINFORMATICS/BTW313>
- Guillien, R., & Vincent, J. (2021). *Diccionario jurídico* (2.^a ed.). Editorial Temis.

- Hearst, M. A. (1999). Untangling Text Data Mining. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, 3-10. www.aaii.org/
- Hechavarría Díaz, A., & Pérez Suárez, A. (2006, octubre). La lematización en el preprocesamiento de textos para RI. Evaluación de distintos algoritmos de lematización. IV Congreso de Reconocimiento de Patrones.
- Hofmann, T. (1999). Probabilistic Latent Semantic Indexing. Proceedings of the Twenty Second Annual International SIGIR Conference on Research and Development in Information Retrieval, 289-296.
- Huang, Y. L., & Kuan, C. M. (2021). Economic prediction with the FOMC minutes: An application of text mining. *International Review of Economics & Finance*, 71, 751-761. <https://doi.org/10.1016/J.IREF.2020.09.020>
- Hurtado, L.-F., Pla, F., & Buscaldi, D. (2015). ELiRF-UPV at TASS 2015: Sentiment Analysis in Twitter. En J. Villena-Román (Ed.), *CEUR Workshop Proceedings* (pp. 75-79). CEUR Workshop Proceeding. https://ceur-ws.org/Vol-1397/tass_proceedings_final_version.pdf#page=75
- Jaccard, P. (1912). The Distribution of the Flora in the Alpine Zone. *New Phytologist*, 11(2), 37-50. <https://doi.org/10.1111/J.1469-8137.1912.TB05611.X>
- Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering. *ACM Computing Surveys (CSUR)*, 31(3), 264-323. <https://doi.org/10.1145/331499.331504>
- Jelinek, F. (1976). Continuous Speech Recognition by Statistical Methods. Proceedings of the IEEE, 64(4). <https://doi.org/10.1109/PROC.1976.10159>
- Jessup, E. R., & Martin, J. H. (2005). Taking a New Look at the Latent Semantic Analysis Approach to Information Retrieval.
- Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1), 11-21. <https://doi.org/10.1108/EB026526>
- Jones, T., Doane, W., & Attbom, M. (2021). textmineR: Functions for Text Mining and Topic Modeling (3.0.5). CRAN de R. <https://cran.r-project.org/web/packages/textmineR/index.html>
- Jones Tommy, & Doane William. (2019). Functions for Text Mining and Topic Modeling.
- Joyce, T., & Needham, R. M. (1958). The Thesaurus Approach to Information Retrieval. *American Documentation*, 9(3), 192-197.
- Khalil, S., & Fakir, M. (2017). RCrawler: An R package for parallel web crawling and scraping. *SoftwareX*, 6, 98-106. <https://doi.org/10.1016/j.softx.2017.04.004>

- Kienberger, M., Solino Pazo, M. M., Eva Cramer Rebeccaand, & Borszik, A. (2021). Determination Strategies for Unknown Words in Texts - An Investigation of Language Learner Strategies at three Spanish Universities. *Deutsch Als Fremdsprache-Zeitschrift Zur Theorie Und Praxis Des Fachesdeutsch Als Fremdsprache*, 58(3), 156-168.
- Kintsch, W. (2002). On the notions of theme and topic in psychological process models of text comprehension. En M. Louwerse & W. Van Peer (Eds.), *Thematics: Interdisciplinary studies* (pp. 157-170). <https://doi.org/10.1075/CELCR.3.14KIN>
- Kodratoff, Y. (1999, junio 8). Knowledge Discovery in Texts: A Definition and Applications. 11th International Symposium on Foundations of Intelligent Systems (ISMIS-99). <https://dl.acm.org/doi/abs/10.5555/646358.689959>
- Kostoff, R. N., Toothman, D. R., Eberhart, H. J., & Humenik, J. A. (2001). Text mining using database tomography and bibliometrics: A review. *Technological Forecasting and Social Change*, 68(3), 223-253. [https://doi.org/10.1016/S0040-1625\(01\)00133-0](https://doi.org/10.1016/S0040-1625(01)00133-0)
- Landauer, T. K., & Dumais, S. T. (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, 104(2), 211-240. <https://doi.org/10.1037/0033-295X.104.2.211>
- Landauer, T. K., Foltz, P. W., & Laham, D. (2009). An introduction to latent semantic analysis. *Discourse Processes*, 25(2-3), 259-284. <https://doi.org/https://doi.org/10.1080/01638539809545028>
- Lebart, L., & Salem, A. (1988). *Analyse Statistique Des Donnees Textuelles*. En Bordas. https://biblio.cerist.dz/index.php/hrbdonf5214/ouvrages/000000000000005946660000_00_2.pdf
- Lebart, L., Salem, A., & Becue-Bertaut, M. (2000). *Análisis Estadístico de Textos* (Vol. 2).
- Leonetti, M., & Escandell, M. V. (2004). *Semántica Conceptual / Semántica Procedimental*. Actas del V Congreso del Lingüística General, 1727-1738. <https://hum.unne.edu.ar/biblioteca/apuntes/Apuntes Letras/TEXTOS DIGITALES LINGÜÍSTICA/Semantica conceptual y procedimental.pdf>
- Liddy, E. D. (1998). Natural Language Processing for Information Retrieval and Knowledge Discovery. En P. Cochrane & E. Johnson (Eds.), *34th Clinic of Library Applications of Data Processing* (pp. 137-147).
- Lindley, D. (1964). The Bayesian Analysis of Contingency Tables. En *The Annals of Mathematical Statistics* (4.^a ed., Vol. 35, pp. 1622-1643). <https://doi.org/https://doi.org/10.1214/aoms/1177700386>

- Liu, V., & Curran, J. (2006, abril 1). Web Text Corpus for Natural Language Processing. Conference of the European Chapter of the Association for Computational Linguistics.
- Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, 2(2), 159-165. <https://doi.org/10.1147/RD.22.0159>
- Luhn, H. P. (2010). A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Development*, 1(4), 309-317. <https://doi.org/10.1147/RD.14.0309>
- Markov, Z., & Larose, D. T. (2007). *Data mining the Web : uncovering patterns in Web content, structure, and usage* (John Wiley & Sons, Ed.; ilustrada). Wiley-Interscience.
- Martí, M. A., & Llisterri, J. (2002). Tratamiento del Lenguaje Natural. En *Tratamiento del lenguaje natural. Tecnología de la lengua oral y escrita*. Universidad de Barcelona.
- Martin, D., & Berry, M. (2007). Mathematical foundations behind latent semantic analysis. En T. Landauer (Ed.), *Handbook of latent semantic analysis*, (pp. 35-56). Routledge.
- Martín-Rodero, H., Sanz-Valero, J., & Galindo-Villardón, P. (2018). The methodological quality of systematic reviews indexed in the MEDLINE database A multivariate approach. *Electronic Library*, 36(1), 146-158. <https://doi.org/10.1108/EL-01-2017-0002/FULL/PDF>
- McManus, J., Li, M., & Moitra, D. (2007a). China: a PESTEL analysis. En *China and India, Oportunidades y amenazas para la industria global del software* (pp. 19-35). Chandos Publishing. <https://doi.org/10.1016/B978-1-84334-158-1.50002-5>
- McManus, J., Li, M., & Moitra, D. (2007b). India: a PESTEL analysis. En *China and India, Oportunidades y amenazas para la industria global del software* (pp. 37-56). Chandos Publishing. <https://doi.org/10.1016/B978-1-84334-158-1.50003-7>
- Mcquitty, L. L. (1966). Similarity analysis by reciprocal pairs for discrete and continuous data. *Educational and Psychological Measurement*, 26(4), 825-831. https://doi.org/10.1177/001316446602600402/ASSET/001316446602600402.FP.PN_G_V03
- Mechura, M. (2016). Lemmatization List hash_lemmas (Lexicon). https://search.r-project.org/CRAN/refmans/lexicon/html/hash_lemmas.html
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6), 1087-1092. <https://doi.org/10.1063/1.1699114>
- Metropolis, N., & Ulam, S. (1952). A property of randomness of an arithmetical function (No. AECU-2038; LADC-1177).

- Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing Semantic Coherence in Topic Models. 262-272. <https://doi.org/10.5555/2145432.2145462>
- Miner, G., Elder, J., Nisbet, R. A., Delen, D., Fast, A., & Hill, T. (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications* (Academic Press, Ed.; 1st ed). Elsevier.
- Minka, T. P. (2001). A family of algorithms for approximate Bayesian inference. Massachusetts InstitutLof Technology .
- Mitchell, J., & Lapata, M. (2010). Composition in Distributional Models of Semantics. *Cognitive Science*, 34, 1388-1429. <https://doi.org/10.1111/j.1551-6709.2010.01106.x>
- Müllner, D. (2011). Modern hierarchical, agglomerative clustering algorithms. <https://arxiv.org/abs/1109.2378v1>
- Murtagh, F., & Legendre, P. (2014). Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification*, 31(3), 274-295. <https://doi.org/10.1007/S00357-014-9161-Z/METRICS>
- Nguyen, V.-A., Boyd-Graber, J., & Resnik, P. (2014). Sometimes Average is Best: The Importance of Averaging for Prediction using MCMC Inference in Topic Modeling. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1752-1757.
- Nikita, M., & Chaney, N. (2020). *ldatuning: Tuning of the Latent Dirichlet Allocation Models Parameters* [R package ldatuning version 1.0.2]. Comprehensive R Archive Network (CRAN). <https://CRAN.R-project.org/package=ldatuning>
- OpenAI. (2023). *GPT-4 Technical Report*. <https://doi.org/https://doi.org/10.48550/arXiv.2303.08774>
- Osuna, Z. M. (2006). *Contribuciones al análisis de datos textuales*. Universidad de Salamanca.
- Osuna, Z., Martín, F., & Galindo-Villardón, P. (2004). Análisis estadístico de datos textuales. Aplicación al estudio de las declaraciones del Libertador Simón Bolívar. *Revista Latinoamericana de Estudios del Discurso*, 4(2), 55-62.
- Paatero, P., & Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2), 111-126. <https://doi.org/10.1002/ENV.3170050203>

- Pan, S., Chon, K., & Song, H. (2008). Visualizing tourism trends: A combination of ATLAS.ti and BiPlot. En *Journal of Travel Research* (Vol. 46, Número 3, pp. 339-348). SAGE Publications Ltd. <https://doi.org/10.1177/0047287507308318>
- Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, EMNLP 2002*, 79-86. <https://arxiv.org/abs/cs/0205070v1>
- Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 6(11), 559-572. <https://doi.org/https://doi.org/10.1080/14786440109462720>
- Perera, R. (2020). The PESTLE Analysis . En *The Pestle Analysis* (2.^a ed.). Nerdynaut.
- Pilacuan-Bonete, L., Galindo-Villardón, P., & Delgado-Álvarez, F. (2022). HJ-Biplot as a Tool to Give an Extra Analytical Boost for the Latent Dirichlet Assignment (LDA) Model: With an Application to Digital News Analysis about COVID-19. *Mathematics*, 10(14). <https://doi.org/10.3390/math10142529>
- Pilacuan-Bonete, L., Galindo-Villardón, P., Delgado-Álvarez, F., & De La Hoz Maestre, J. (2022). Package LDABiplots (0.1.2; pp. 1-11). CRAN R. <https://cran.r-project.org/web/packages/LDABiplots/index.html>
- Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3), 130-137. <https://doi.org/10.1108/EB046814/FULL/XML>
- Posit. (2023). RStudio IDE. <https://posit.co/products/open-source/rstudio/>
- Putman, H. (1975). The meaning of «meaning». En *Philosophical Papers, Mind, Language, and Reality* (Vol. 2, pp. 215-271). Cambridge University Press. https://doi.org/10.11517/JJSAI.26.4_334
- Quillian, M. R. (1967). Word concepts: a theory and simulation of some basic semantic capabilities. *Behavioral science*, 12(5), 410-430. <https://doi.org/10.1002/BS.3830120511>
- Quillian, M. R. (1969). The teachable language comprehender: a simulation program and theory of language. *Communications of the ACM*, 12(8), 459-476. <https://doi.org/10.1145/363196.363214>
- R Core Team. (2023). R: The R Project for Statistical Computing. <https://www.r-project.org/>
- R Development Core Team. (2000). *Introducción a R*.
- Rada, R., & Bicknell, E. (1989). Ranking documents with a thesaurus . *Journal of the American Society for Information Science*, 40(5), 304-310.

[https://doi.org/https://doi.org/10.1002/\(SICI\)1097-4571\(198909\)40:5%3C304::AID-ASI2%3E3.0.CO;2-6](https://doi.org/https://doi.org/10.1002/(SICI)1097-4571(198909)40:5%3C304::AID-ASI2%3E3.0.CO;2-6)

- Ramesh, R., Divya, G., Divya, D., Kurian, M., & Vishnuprabha, V. (2015). Big Data Sentiment Analysis using Hadoop. *IJIRST-International Journal for Innovative Research in Science & Technology*, 1(11).
- Raulji, J. K., Saini, J. R., & Ambedkar, B. (2016). Stop-Word Removal Algorithm and its Implementation for Sanskrit Language. *International Journal of Computer Applications*, 150(2), 975-8887. <https://doi.org/10.5958/2249-3220.2015.00015.4>
- Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. En M. Valletta (Ed.), *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks* (pp. 46-50). Universidad de Malta.
- Ren, W., & Han, K. (2014). Sentiment Detection of Web Users Using Probabilistic Latent Semantic Analysis. *Journal of Multimedia*, 9(10), 1194-1200. <https://doi.org/10.4304/jmm.9.10.1194-1200>
- Rinker, T. (2019). Title Lexicons for Text Analysis (1.2.1). <https://github.com/trinker/lexicon>
- Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*, 58(4), 1064-1082. <https://doi.org/10.1111/AJPS.12103>
- Robertson, A. M., & Willett, P. (1998). Applications of N-Grams in Textual Information Systems. *Journal of Documentation*, 54(1), 48-69.
- Rohlf, F. J. (1970). Adaptive Hierarchical Clustering Schemes. *Systematic Biology*, 19(1), 58-82. <https://doi.org/10.1093/SYSBIO/19.1.58>
- SAIJ. (2016). Índice Político del Sistema Argentino de Información Jurídica. Biblioteca Central de la Corte Suprema de Justicia de la Nación. <http://vocabularios.saij.gob.ar/portalthes/index.php?task=fetchTerm&arg=3567&v=3>
- Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4), 406-425. <https://doi.org/10.1093/OXFORDJOURNALS.MOLBEV.A040454>
- Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513-523. [https://doi.org/10.1016/0306-4573\(88\)90021-0](https://doi.org/10.1016/0306-4573(88)90021-0)

- Salton, G., Wong, A., & Yang, C. S. (1975). A Vector Space Model for Automatic Indexing. *Information Retrieval and Language Processing, Communications of the ACM*, 18(11), 613-620. <https://doi.org/https://doi.org/10.1145/361219.361220>
- Salton, G., & Yang, C. S. (1973). On the specification of term values in automatic indexing. *Journal of Documentation*, 29(4), 351-372. <https://doi.org/10.1108/EB026562/FULL/PDF>
- Salton, G., Yang, C. S., & Yu, C. T. (1975). A theory of term importance in automatic text analysis. *Journal of the American Society for Information Science*, 26(1), 33-44. <https://doi.org/10.1002/ASI.4630260106>
- Sánchez, H., Reyes, C., & Mejía Sáenz, K. (2018). *Manual de términos en investigación científica, tecnológica y humanística* (H. Sánchez Carlessi, Ed.; 1.^a ed.). Universidad Ricardo Palma.
- Sarle, W. S., Kaufman, L., & Rousseeuw, P. J. (1991). Finding Groups in Data: An Introduction to Cluster Analysis. *Journal of the American Statistical Association*, 86(415), 830. <https://doi.org/10.2307/2290430>
- Saul, L., & Pereira, F. (1997a). Aggregate and mixed-order Markov models for statistical language processing. *Proceedings of the 2nd International Conference on Empirical Methods in Natural Language Processing*, cmp-lg/9706007. <https://doi.org/10.48550/ARXIV.CMP-LG/9706007>
- Saul, L., & Pereira, F. (1997b, junio 9). Aggregate and mixed-order Markov models for statistical language processing. *Proceedings of the 2nd International Conference on Empirical Methods in Natural Language Processing*.
- Sbalchiero, S., & Eder, M. (2020). Topic modeling, long texts and the best number of topics. Some Problems and solutions. *Quality and Quantity*, 54(4), 1095-1108. <https://doi.org/10.1007/S11135-020-00976-W/TABLES/7>
- Seijo, F. C., Luna, J. M. F., & Guadix, J. F. H. (2011). *Recuperación de Información. Un enfoque práctico y multidisciplinar*. RA-MA Editorial.
- Sepúlveda, C. (2004). *Diccionario de Términos Económicos* (11.^a ed.). Universitaria S.A.
- Severyn, A., & Moschitti, A. (2015). Twitter Sentiment Analysis with deep convolutional neural networks. *SIGIR 2015 - Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 959-962. <https://doi.org/10.1145/2766462.2767830>
- Shayaa, S., Jaafar, N. I., Bahri, S., Sulaiman, A., Seuk Wai, P., Wai Chung, Y., Piprani, A. Z., & Al-Garadi, M. A. (2018). Sentiment analysis of big data: Methods, applications,

- and open challenges. *IEEE*, 6, 37807-37827.
<https://doi.org/10.1109/ACCESS.2018.2851311>
- Shen, Y., & Guo, H. (2022). Research on high-performance English translation based on topic model. *Digital Communications and Networks*.
<https://doi.org/10.1016/J.DCAN.2022.03.015>
- Sidorov, G. (2019). Syntactic n-grams: The concept. En *SpringerBriefs in Computer Science*. https://doi.org/10.1007/978-3-030-14771-6_8
- Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., & Chanona-Hernández, L. (2014). Syntactic N-grams as machine learning features for natural language processing. *Expert Systems with Applications*, 41(3), 853-860.
<https://doi.org/10.1016/j.eswa.2013.08.015>
- Sievert, C., & Shirley, K. E. (2014). LDAvis: A method for visualizing and interpreting topics. *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, 63-70.
- Sinclair, J. M. (1966). Beginning the Study of Lexis. *Memory of J. R. Firth*, 410-430.
- Sinclair, J. M. (1991). *Corpus, concordance, collocation* (J. Sinclair & R. Carter, Eds.; 2.^a ed.). Oxford University Press.
- Sokal, R. R. (1963). The Principles And Practice Of Numerical Taxonomy. *Taxon*, 12(5), 190-199. <https://doi.org/10.2307/1217562>
- Sorensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biologiske Skrifter*, 5(4), 1-34.
- Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1), 11-21.
<https://doi.org/10.1108/00220410410560573>
- Srinivasarao, U., & Sharaff, A. (2022). Email thread sentiment sequence identification using PLSA clustering algorithm. *Expert Systems with Applications*, 193, 116475.
<https://doi.org/10.1016/J.ESWA.2021.116475>
- Steyvers, M., & Griffiths, T. (2007). Probabilistic Topic Models. En T. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (1st Edition, pp. 427-448). Taylor and Francis.
<https://doi.org/10.4324/9780203936399-29>

- Sumba, X., & Bouguila, N. (2020). Improving Classification Using Topic Correlation and Expectation Propagation. *Advances in Artificial Intelligence*, 496-507. https://doi.org/10.1007/978-3-030-47358-7_51/COVER
- Tan, A.-H. (1999, abril). Text Mining: The state of the art and the challenges. Workshop Knowledge Discovery from advanced Databases PAKDDD-99.
- Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476), 1566-1581. <https://doi.org/10.1198/016214506000000302>
- Thakur, V. (2021). Framework for PESTEL dimensions of sustainable healthcare waste management: Learnings from COVID-19 outbreak. *Journal of Cleaner Production*, 287, 125562. <https://doi.org/10.1016/J.JCLEPRO.2020.125562>
- Topal, M. O., Bas, A., & van Heerden, I. (2021, febrero 16). Exploring Transformers in Natural Language Generation: GPT, BERT, and XLNet. Conference on Interdisciplinary Applications of Artificial Intelligence .
- Turney, P., & Pantel, P. (2010, febrero). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37, 141-188. <https://doi.org/DOI:10.1613/jair.2934>
- Vidales, L. (2003). Glosario de términos financieros: términos financieros, contables, administrativos, económicos y legales (1.ª ed.). Plaza y Valdés.
- Viera, A. F. G., & Viera, A. F. G. (2017). Técnicas de aprendizaje de máquina utilizadas para la minería de texto. *Investigación bibliotecológica*, 31(71), 103-126. <https://doi.org/10.22201/IIBI.0187358XP.2017.71.57812>
- Ward, J. H., & Hook, M. E. (1963). Application of an hierarchical grouping procedure to a problem of grouping profiles. *Educational and Psychological Measurement*, 23(1), 69-81. https://doi.org/10.1177/001316446302300107/ASSET/001316446302300107.FP.PN_G_V03
- Weiss, S. M., Indurkha, N., Zhang, T., & Damerau, F. J. (2005). Text mining: Predictive methods for analyzing unstructured information. En *Text Mining: Predictive Methods for Analyzing Unstructured Information* (1 ed). Springer. <https://doi.org/10.1007/978-0-387-34555-0>
- Wickham, H., Bryan, J., & Posit. (2023). CRAN - Package readxl (1.4.2). <https://cran.r-project.org/web/packages/readxl/index.html>

- Wickham, H., Francois, R., Henry, L., & Muller, K. (2023). dplyr (1.1.2). CRAN de R. <https://cran.r-project.org/web/packages/dplyr/index.html>
- Wickham Hadley. (2019). Easily Harvest (Scrape) Web Pages (0.3.5). foundation R. <https://github.com/tidyverse/rvest/issues>
- Yano, T., Smith, N. A., & Wilkerson, J. D. (2012). Textual Predictors of Bill Survival in Congressional Committees. <https://doi.org/10.5555/2382029>
- Yi, X., & Allan, J. (2009). A comparative study of utilizing topic models for information retrieval. *Lecture Notes in Computer Science*, 5478 LNCS, 29-41. https://doi.org/10.1007/978-3-642-00958-7_6
- Zellig S, H. (1952). Discourse Analysis. *Language*, 28(1), 1-30.
- Zellig S, H. (1991). *Theory of Language and Information: A Mathematical Approach*. Oxford University Press UK.
- Zhang, B., Peng, B., & Qiu, J. (2016). High Performance LDA through Collective Model Communication Optimization. *Procedia Computer Science*, 80, 86-97. <https://doi.org/10.1016/J.PROCS.2016.05.300>
- Zhang, J., & Zong, C. (2015). Deep Neural Networks in Machine Translation: An Overview. *IEEE Intelligent Systems*, 30(05), 16-25. <https://doi.org/10.1109/MIS.2015.69>
- Zoido N, F., Vega B, S., Morales M, G., Hernández, R., & González Ruben. (2000). *Diccionario de geografía urbana, urbanismo y ordenación del territorio* (1.^a ed.). Editorial Ariel.