UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Informática

Clasificación automática de empresas en sectores a partir de su descripción de actividad

Trabajo Fin de Grado

Grado en Ciencia de Datos

AUTOR/A: Canovas Vidal, Irene

Tutor/a: Doménech i de Soria, Josep

CURSO ACADÉMICO: 2023/2024

Resumen

La NACE (Nomenclature of Economic Activities) es un sistema que se encarga de clasificar empresas a nivel europeo según su actividad principal y secundaria. Hoy en día, son las propias empresas las que seleccionan manualmente el código que mejor describe su actividad económica, lo que en ocasiones puede ser inexacto. Este Trabajo de Fin de Grado tiene como objetivo proponer y evaluar alternativas para la clasificación automática de las empresas mediante técnicas de aprendizaje automático y procesamiento de datos textuales. Se empleará información obtenida de la base de datos SABI para llevar a cabo este estudio. El TFG busca mejorar la precisión y eficiencia de la clasificación realizada actualmente por la NACE. Se utilizarán las descripciones de actividad que las propias empresas proporcionan al incluirse en la NACE para realizar una nueva clasificación mediante técnicas de aprendizaje automático y procesamiento de texto. Los resultados esperados incluyen una mayor precisión en la asignación de códigos NACE, lo que facilitará la identificación y comparación de empresas dentro de un mismo sector económico.

Palabras clave: NACE, Clasificación automática, SABI, Procesamiento del lenguaje natural, aprendizaje automático.

Abstract

The NACE (Nomenclature of Economic Activities) is a system that classifies companies at European level according to their main and secondary activity. Nowadays, it is the companies themselves that manually select the code that best describes their economic activity, which can sometimes be inaccurate. This Final Degree Project aims to propose and evaluate alternatives for the automatic classification of companies using machine learning techniques and textual data processing. Information obtained from the SABI database will be used to carry out this study. The TFG aims to improve the accuracy and efficiency of the current NACE classification. The activity descriptions provided by the enterprises themselves when they are included in the NACE will be used to make a new classification using machine learning and text processing techniques. The expected results include greater accuracy in the assignment of NACE codes, which will facilitate the identification and comparison of enterprises within the same economic sector.

Keywords: NACE, Automatic classification, SABI, Natural language processing, Machine learning.

Índice de contenidos

1. Introducción 8
   1.1. Resumen 8
   1.2. Motivación 8
   1.3. Objetivos 8
   1.4. Estructura del TFG 9
2. Marco contextual 10
   2.1. La actividad económica de una empresa 10
   2.2. Clasificaciones de actividades económicas 11
      2.2.1. Evolución de la NACE 11
      2.2.2. Estructura y codificación de la NACE 12
      2.2.3. Criterios adoptados para el desarrollo de la NACE y su armonización internacional 12
      2.2.4. Reglas de clasificación de actividades y unidades 13
      2.2.5. Problemáticas de la clasificación NACE 14
   2.3. Trabajos anteriores sobre clasificación de actividad económica a partir de texto 15
   2.4. Recapitulación 16
3. Metodología 18
   3.1. Muestra 18
      3.1.1. Origen de datos 18
      3.1.2. Variables 18
   3.2. Preprocesamiento de datos 19
   3.3. Topic Modeling 22
      3.3.1. Latent Dirichlet Allocation (LDA) 23
      3.3.2. Non-negative Matrix Factorization (NMF) 26
      3.3.3. BERTopic 28
   3.4. Hierarchical Topic Modelling 30
   3.5. Evaluación de la clasificación de actividades 31
      3.5.1. Entropía 32
      3.5.2. Coeficiente de Gini 33
4. Resultados 35
   4.1. Análisis descriptivo 35
   4.2. Entropía por sección NACE 38
      4.2.1. Cálculo de la Entropía para la clasificación realizada por la NACE 38

4.2.2. Cálculo de la Entropía para la clasificación realizada con el modelo LDA 40

4.2.1. Cálculo de la Entropía para la clasificación realizada con el modelo BERTopic 45

4.2.2. Cálculo de la Entropía para la clasificación realizada con el modelo NMF 49

4.2.3. Comparación de resultados 54

5. Conclusiones 57

Bibliografía 60

6. Anexo I. Objetivos de Desarrollo Sostenible 64

7. Anexo II. Nubes de palabras 66

7.1. Nubes de palabras para las secciones NACE 66

7.2. Nubes de palabras para las secciones LDA 71

7.3. Nubes de palabras para las secciones BERTopic 76

7.4. Nubes de palabras para las secciones NMF 80

Índice de figuras

Figura 1: Ejemplo de clasificación NACE de una empresa según las actividades que desarrolla. 14
Figura 2: Ejemplo en formato de árbol de clasificación NACE de una empresa según las actividades que desarrolla. 14
Figura 3: Formato inicial de la base de datos descargada de SABI. 19
Figura 4: Estructura general de clasificación de empresas en sectores según su código primario. 20
Figura 5: Diagrama de cómo funciona el Topic Modelling. 23
Figura 6: Ejemplo de cómo funciona LDA. 24
Figura 7: Diagrama del modelo LDA. 24
Figura 8: Diagrama del modelo NMF. 26
Figura 9: Diagrama del modelo BERTopic. 29
Figura 10: Descripción general del algoritmo HLDA. 31
Figura 11: Ejemplo correlación de temas en PAM. 31
Figura 12: Entropía de la información en un ensayo de Bernoulli X. 32
Figura 13: Recta de igualdad perfecta, curva de Lorenz y área de Gini. 34
Figura 14: Grado de relación del trabajo con los Objetivos de Desarrollo Sostenible ... 64

Índice de gráficos

Gráfico 1: Distribución de Empresas por sección asignado con NACE. 35
Gráfico 2: Nube de palabras para la sección A con NACE. 36
Gráfico 3: Nube de palabras para la sección C con NACE. 36
Gráfico 4: Nube de palabras para la sección I con NACE. 37
Gráfico 5: Códigos primarios más comunes. 37
Gráfico 6: Entropía por sección de la clasificación NACE. 39
Gráfico 7: Tabla de frecuencias con el número de empresas y entropía por sección NACE. 39
Gráfico 8: Correlación entre Entropía y Número de Empresas por Sección NACE. 40
Gráfico 9: Distribución de empresas por sección asignada con LDA. 41
Gráfico 10: Nube de Palabras para la sección H. 41
Gráfico 11: Nube de Palabras para la sección K. 42
Gráfico 12: Nube de Palabras para la sección N. 42
Gráfico 13: Nube de Palabras para la sección R. 42
Gráfico 14: Entropía por sección de la clasificación LDA. 43
Gráfico 15: Tabla de frecuencias con el número de empresas y entropía por sección LDA. 44
Gráfico 16: Correlación entre Entropía y Número de Empresas por Sección LDA. 44
Gráfico 17: Distribución de empresas por sección asignada con BERTopic. 45
Gráfico 18: Nube de Palabras para la sección B. 46
Gráfico 19: Nube de Palabras para la sección D. 46
Gráfico 20: Nube de Palabras para la sección I. 47
Gráfico 21: Nube de Palabras para la sección U. 47
Gráfico 22: Entropía por sección de la clasificación BERTopic. 48
Gráfico 23: Tabla de frecuencias con el número de empresas y entropía por sección BERTopic. 48
Gráfico 24: Correlación entre Entropía y Número de Empresas por Sección BERTopic. 49
Gráfico 25: Distribución de empresas por sección asignada con NMF. 50
Gráfico 26: Nube de Palabras para la sección P. 50
Gráfico 27:Nube de Palabras para la sección B. 51
Gráfico 28: Nube de Palabras para la sección I. 51
Gráfico 29: Nube de Palabras para la sección R. 52
Gráfico 30: Entropía por sección de la clasificación NMF. 53
Gráfico 31: Tabla de frecuencias con el número de empresas y entropía por sección NMF. 53
Gráfico 32: Correlación entre Entropía y Número de Empresas por Sección NMF. 54
Gráfico 34: Nube de palabras para la sección B NACE 66
Gráfico 36: Nube de palabras para la sección D NACE 66
Gráfico 37: Nube de palabras para la sección E NACE 66
Gráfico 38: Nube de palabras para la sección F NACE 68
Gráfico 39: Nube de palabras para la sección G NACE 68
Gráfico 40: Nube de palabras para la sección H NACE 68
Gráfico 42: Nube de palabras para la sección J NACE 68
Gráfico 43: Nube de palabras para la sección K NACE 69

Gráfico 44: Nube de palabras para la sección L NACE 69
Gráfico 45: Nube de palabras para la sección M NACE 69
Gráfico 46: Nube de palabras para la sección N NACE 69
Gráfico 47: Nube de palabras para la sección O NACE 70
Gráfico 48: Nube de palabras para la sección P NACE 70
Gráfico 49: Nube de palabras para la sección Q NACE 70
Gráfico 50: Nube de palabras para la sección R NACE 70
Gráfico 51: Nube de palabras para la sección S NACE 71
Gráfico 52: Nube de palabras para la sección T NACE 71
Gráfico 53: Nube de palabras para la sección U 71
Gráfico 54: Nube de palabras para la sección A LDA 71
Gráfico 55: Nube de palabras para la sección B LDA 72
Gráfico 56: Nube de palabras para la sección C LDA 72
Gráfico 57: Nube de palabras para la sección D LDA 72
Gráfico 58: Nube de palabras para la sección E LDA 72
Gráfico 59: Nube de palabras para la sección F LDA 73
Gráfico 60: Nube de palabras para la sección G LDA 73
Gráfico 62: Nube de palabras para la sección I LDA 73
Gráfico 63: Nube de palabras para la sección J LDA 73
Gráfico 65: Nube de palabras para la sección L LDA 74
Gráfico 66: Nube de palabras para la sección M LDA 74
Gráfico 68: Nube de palabras para la sección O LDA 74
Gráfico 69: Nube de palabras para la sección P LDA 74
Gráfico 70: Nube de palabras para la sección Q LDA 75
Gráfico 72: Nube de palabras para la sección S LDA 75
Gráfico 73: Nube de palabras para la sección T LDA 75
Gráfico 74: Nube de palabras para la sección U LDA 75
Gráfico 75: Nube de palabras para la sección A BERTopic 76
Gráfico 77: Nube de palabras para la sección C BERTopic 76
Gráfico 79: Nube de palabras para la sección E BERTopic 76
Gráfico 80: Nube de palabras para la sección F BERTopic 76
Gráfico 81: Nube de palabras para la sección G BERTopic 77
Gráfico 82: Nube de palabras para la sección H BERTopic 77
Gráfico 84: Nube de palabras para la sección J BERTopic 77
Gráfico 85: Nube de palabras para la sección K BERTopic 77
Gráfico 86: Nube de palabras para la sección L BERTopic 78
Gráfico 87: Nube de palabras para la sección M BERTopic 78
Gráfico 88: : Nube de palabras para la sección N BERTopic 78
Gráfico 89: Nube de palabras para la sección O BERTopic 78
Gráfico 90: Nube de palabras para la sección P BERTopic 79
Gráfico 91: Nube de palabras para la sección Q BERTopic 79
Gráfico 92: Nube de palabras para la sección R BERTopic 79
Gráfico 93: Nube de palabras para la sección S BERTopic 79
Gráfico 94: Nube de palabras para la sección T BERTopic 80
Gráfico 96: Nube de palabras para la sección A NMF 80
Gráfico 98: Nube de palabras para la sección C NMF 80
Gráfico 99: Nube de palabras para la sección D NMF 80
Gráfico 100: Nube de palabras para la sección E NMF 81

Gráfico 101: Nube de palabras para la sección F NMF.....81
Gráfico 102: Nube de palabras para la sección G NMF.....81
Gráfico 103: Nube de palabras para la sección H NMF 81
Gráfico 104: Nube de palabras para la sección I NMF..... 82
Gráfico 105: Nube de palabras para la sección J NMF 82
Gráfico 106: Nube de palabras para la sección K NMF..... 82
Gráfico 108: Nube de palabras para la sección M NMF..... 82
Gráfico 109: Nube de palabras para la sección N NMF 83
Gráfico 110: Nube de palabras para la sección O NMF..... 83
Gráfico 112: Nube de palabras para la sección Q NMF 83
Gráfico 114: Nube de palabras para la sección S NMF..... 83
Gráfico 115: Nube de palabras para la sección T NMF..... 84
Gráfico 116: Nube de palabras para la sección U NMF 84

1. Introducción

1.1. Resumen

Este Trabajo Fin de Grado tiene como objetivo clasificar automáticamente empresas en sectores económicos a partir de descripciones textuales de su actividad. Hoy en día, son las propias empresas las que manualmente eligen cual es la sección de la NACE en la que mejor se clasifica la actividad que desempeñan. Muchas veces la elección de códigos NACE de las empresas no reflejan con precisión la actividad real de la empresa, lo que puede llevar a inexactitudes en análisis económicos y comparaciones sectoriales.

Con el objetivo de resolver o por lo menos reducir este problema, el proyecto explora el uso de técnicas de procesamiento de lenguaje natural (NLP) y aprendizaje automático, con la información obtenida de la base de datos SABI. Con estas tecnologías se pretende conseguir una asignación más precisa y coherente de códigos NACE, además de una mejora en eficiencia y precisión en la clasificación de sectores de actividad empresarial.

Durante el desarrollo de este proyecto, se detalla el proceso de selección y aplicación de algoritmos NLP y aprendizaje automático, diseñados para interpretar descripciones textuales y asignar automáticamente los códigos NACE correspondientes. Se espera que los resultados del estudio faciliten una mejor identificación y comparación entre empresas, mejorando tanto la calidad de los datos económicos disponibles como la toma de decisiones basada en estos.

1.2. Motivación

La correcta clasificación de empresas según su actividad económica es fundamental para la toma de decisiones en el ámbito público. Sin embargo, la tarea de clasificar manualmente cada empresa en el sistema NACE, ha demostrado tener errores, afectando a la precisión y fiabilidad de las estadísticas económicas. La motivación detrás de este Trabajo Fin de Grado surge de la necesidad de abordar estas limitaciones, proponiendo un sistema automatizado para mejorar la eficiencia y exactitud en la clasificación de empresas por sector de actividad.

Este proyecto está motivado por la creciente disponibilidad de grandes volúmenes de datos textuales y la evolución de las técnicas de procesamiento de lenguaje natural (NLP) y aprendizaje automático, que ahora permiten interpretar y clasificar de manera efectiva textos descriptivos más complejos. A través de la automatización, se pretende mejorar el método de clasificación manual, evitando la subjetividad en la interpretación de los códigos NACE, facilitando así una asignación que refleje de manera fiable la actividad real de las empresas.

1.3. Objetivos

El objetivo principal del TFG es desarrollar un sistema capaz de interpretar y clasificar automáticamente las descripciones textuales de actividades empresariales en categorías de la NACE. Para ello, podemos listar los siguientes objetivos:

- Explorar métodos de procesamiento de lenguaje natural (NLP), investigando y comprendiendo las técnicas y algoritmos de NLP que pueden ser aplicados para el análisis de descripciones textuales de actividades empresariales.
- Analizar la estructura y el funcionamiento de la Clasificación Europea Normalizada de Actividades Económicas Productivas (NACE) estudiando en profundidad la jerarquía de la NACE, sus categorías, códigos y principios metodológicos que rigen su funcionamiento.
- Evaluar la precisión y eficacia del sistema desarrollado realizando pruebas y evaluaciones del sistema de clasificación automática para determinar su precisión, coherencia y eficacia en la asignación de actividades empresariales a las categorías de la NACE.
- Proporcionar recomendaciones y mejoras: Basado en los resultados obtenidos, identificar áreas de mejora y proporcionar recomendaciones para optimizar el sistema de clasificación automática y su aplicación práctica en contextos empresariales reales.

1.4. Estructura del TFG

Este Trabajo Final de Grado se estructura en cinco capítulos los cuales se detallan a continuación, incluyendo su secuencia y contenido principal:

En primer lugar, se inicia con una introducción que expone los objetivos principales del proyecto, destacando la necesidad de abordar el problema existente con la clasificación manual de códigos NACE.

En segundo lugar, se describe el marco contextual, se analiza la actividad económica de una empresa, se explican las clasificaciones de actividades económicas y se revisan trabajos anteriores sobre clasificación de actividad económica a partir de texto.

A continuación, se detalla la metodología empleada, que incluye la descripción de la muestra de datos proveniente de la base de datos SABI, el preprocesamiento de los datos textuales de las actividades empresariales, así como el uso de técnicas de topic modeling y hierarchical topic modelling para la clasificación automática. También se introduce la entropía como medida de incertidumbre para comparar los resultados.

En cuarto lugar, los resultados se presentan en forma de análisis descriptivo de la distribución de los códigos NACE y las clasificaciones obtenidas por los distintos modelos aplicados, además de una comparación de dicha clasificación en secciones NACE.

Finalmente, en las conclusiones se evalúa el cumplimiento de los objetivos, se recapitulan los resultados, se discuten las limitaciones del estudio y se proponen posibles trabajos futuros derivados de esta investigación.

2. Marco contextual

En este capítulo se pone en contexto el trabajo realizado. En primer lugar, se examina la importancia de que las empresas declaren su objeto social, un requisito que queda reflejado en su registro en entidades como el Registro Mercantil y bases de datos como SABI. En segundo lugar, se ha indagado en la evolución de la NACE a lo largo de los años y se ha realizado un análisis profundo de su estructura y funcionamiento, en concreto de su codificación y normativa a la hora de clasificar actividades. Se destaca la NACE como estándar europeo y se exploran los procedimientos de actualización, así como los problemas ligados a este sistema de clasificación derivados en ocasiones por la aparición de actividades nuevas que no se encuentran contempladas en la clasificación existente. Por último, se revisan trabajos previos relacionados con la clasificación de actividad económica a partir de texto, proporcionando un contexto para este estudio.

2.1. La actividad económica de una empresa

La clasificación y análisis de las actividades económicas de las empresas se ha vuelto una tarea cada vez más compleja. Algunos sistemas como la Nomenclature of Economic Activities (NACE) Rev.2 de Eurostat han proporcionado un marco estándar para categorizar las actividades empresariales en la Unión Europea. Este sistema organiza las actividades económicas en un conjunto jerárquico de categorías basadas en el tipo de actividades realizadas por empresas, de esta forma se pueden comparar y analizar a nivel europeo (Eurostat, 2008).

Un aspecto importante en la clasificación de las empresas es la declaración de su objeto social, el cual debe ser especificado durante su constitución y se inscribe en el Registro Mercantil. Este Registro es el principal instrumento legal que dota de seguridad al tráfico mercantil y es esencial para el desarrollo económico, actuando como medio para reducir los costes de transacción. Las inscripciones en el Registro se realizan tras una calificación que controla la legalidad y validez de los actos y acuerdos sociales, así como la capacidad y legitimación de quienes los suscriben. El Registro Mercantil es administrado por registradores mercantiles, es decir, profesionales del derecho que ejercen una función pública asegurando la legalidad de todos los documentos que acceden a él. Existe también un Registro Mercantil Central encargado de atribuir las denominaciones de las sociedades y entidades mercantiles, subrayando la importancia de este sistema para la formalización de negocios en España (Colegio de Registradores, 2024).

La precisión en la definición del objeto social y su inscripción en el Registro Mercantil no solo cumple con una exigencia legal; representa también un acto de comunicación con el mercado y con la sociedad, ofreciendo una visión clara de las operaciones empresariales. Este proceso asegura que la actividad económica de la empresa esté claramente definida, evitando confusiones que puedan resultar en prácticas desleales o competencia desequilibrada. Asimismo, la inscripción en el registro mercantil facilita la clasificación de las empresas en sistemas como la NACE.

Además de la inscripción en el Registro Mercantil, herramientas como SABI (Sistema de Análisis de Balances Ibéricos) facilitan el acceso a información financiera detallada y actualizada de más de 2 millones de empresas, permitiendo construir ficheros permanentes de clientes y proveedores, realizar análisis de créditos personalizados, elaborar informes según necesidades específicas, comparar el posicionamiento de una empresa frente a sus competidores, identificar oportunidades para adquisiciones, fusiones, y mucho más. Esta herramienta, es muy fácil de usar y permite a los usuarios disponer de grandes cantidades de datos empresariales, contribuyendo a la toma de decisiones basada en análisis del entorno empresarial, además la mayoría de los datos disponibles en SABI proceden del registro mercantil (D&B, 2024).

2.2. Clasificaciones de actividades económicas

La NACE (Clasificación Nacional de Actividades Económicas) es un sistema utilizado para categorizar diversas actividades económicas en la Unión Europea desde 1970. Su principal función es proporcionar un marco para la recopilación y presentación de datos estadísticos relacionados con la actividad económica en áreas como la producción, el empleo, y las cuentas nacionales, entre otros aspectos. Esta clasificación permite que las estadísticas generadas sean comparables tanto a nivel europeo como global, siendo su uso obligatorio dentro del Sistema Estadístico Europeo. Esencialmente, la NACE organiza las actividades económicas de manera estandarizada, asignando códigos que permiten identificarlas y asociarlas con unidades estadísticas específicas. Sin embargo, la constante evolución del mercado y el surgimiento de nuevas industrias hacen que este sistema necesite ajustes y mejoras continuas en estos sistemas de clasificación. El proyecto ESSnet Trusted Smart Statistics – Web Intelligence Network ha abordado esta necesidad, implementando metodologías para actualizar y mejorar la precisión y aplicabilidad de los códigos NACE (Dabrowski, y otros, 2022).

2.2.1. Evolución de la NACE

El término NACE proviene del francés Nomenclature statistique des activités économiques dans la Communauté européenne, se trata de una clasificación de cuatro dígitos que proporciona el marco para recopilar y presentar datos estadísticos según la actividad económica en estadísticas europeas en el ámbito económico, social, ambiental y agrícola (Comisión Europea, 2020).

Entre 1961 y 1963, se desarrolló la "Nomenclatura de las industrias establecidas en las Comunidades Europeas" (NICE), abarcando industrias como las extractivas, productoras de energía, manufactura, y construcción. Posteriormente, en 1965, se creó la "Nomenclatura de comercio en la CEE" (NCE) para todas las actividades comerciales, seguidamente, en 1967 las clasificaciones de servicios y agricultura se agruparon en divisiones más amplias. En 1970, se introdujo la "Nomenclatura general de las actividades económicas en las Comunidades Europeas", que englobaba una amplia gama de actividades económicas. Sin embargo, esta primera versión de la NACE tenía dos desafíos principales: la recolección de datos se basaba en clasificaciones nacionales existentes, lo que dificultaba la comparación internacional, y carecía de reconocimiento dentro de un marco internacional. Para abordar estos problemas, se consideró armonizar la NACE con normas internacionales como la Clasificación Industrial Internacional Uniforme de todas las Actividades Económicas (CIIU Rev. 3), adoptada

por la Comisión de Estadística de las Naciones Unidas en 1989. Posteriormente, se desarrolló la NACE Rev. 1, basada en la estructura de la CIIU Rev. 3, seguida de una actualización en 2002 (NACE Rev. 1.1) para incorporar nuevas actividades y reflejar cambios tecnológicos y organizativos. Finalmente se adoptó el reglamento que establece la NACE Rev. 2 utilizada actualmente en diciembre de 2006, manteniendo sus características generales, pero equilibrando el nivel de detalle solicitado por los usuarios con la carga de trabajo de los institutos de estadística (Eurostat, 2008).

En 2023, se estableció la revisión 2 actualización 1 de NACE (NACE Rev. 2.1) que será implementada progresivamente en todos los dominios estadísticos relevantes a partir de 2025.

2.2.2. Estructura y codificación de la NACE

La NACE se organiza mediante una estructura jerárquica que se especifica en el Reglamento correspondiente. Esta estructura incluye:

1. Un nivel inicial que comprende secciones identificadas por códigos alfabéticos.
2. Un segundo nivel que consiste en divisiones identificadas por códigos numéricos de dos dígitos.
3. Un tercer nivel compuesto por grupos identificados por códigos numéricos de tres cifras.
4. Un cuarto nivel de clases identificadas por códigos numéricos de cuatro dígitos.

El sistema de codificación de la NACE asigna códigos alfanuméricos para identificar las diferentes divisiones, grupos y clases que describen actividades económicas específicas. En este sistema, el código de sección (alfabético) no está incluido en el código NACE que identifica la división, el grupo y la clase correspondientes. Por ejemplo, la actividad "Fabricación de colas" se identifica con el código 20.52, donde 20 representa la división, 20.5 el grupo y 20.52 la clase; la sección se puede obtener a través de los dos primeros números. Las divisiones se codifican de forma secuencial, pero se han reservado algunos números de código para futuras divisiones sin cambiar completamente la codificación. Además, cuando un nivel de la clasificación no se subdivide más, se utiliza "0" en la posición del código para el siguiente nivel más detallado. Por ejemplo, la clase "Actividades veterinarias" se codifica como 75,00 porque la división correspondiente no se divide más. Las categorías residuales se identifican con la cifra 9, como el grupo 08.9 "Explotación de minas y canteras no comprendidos en otras partes" y la clase 08.99 "Otras explotaciones extractivas n.c.o.p." (Eurostat, 2008).

2.2.3. Criterios adoptados para el desarrollo de la NACE y su armonización internacional

La Clasificación Nacional de Actividades Económicas (NACE) es un sistema esencial dentro de la Unión Europea para categorizar actividades económicas, su evolución va ligada a la adaptación a los cambios en la economía. Un aspecto importante de su desarrollo es la necesidad de armonización con clasificaciones internacionales, permitiendo así la comparabilidad y análisis económico tanto a nivel europeo como global. Esto se refleja en la revisión y actualización continua de la NACE (Eurostat, 2008).

Últimamente, la globalización de la economía y el avance tecnológico han impulsado revisiones de las nomenclaturas estadísticas, de esta forma se ha obtenido un sistema que armoniza las clasificaciones de actividades económicas y productos a escala mundial, europea y nacional. La versión nacional de la NACE, conocida como CNAE en España, es un ejemplo de cómo se adaptan estas clasificaciones a contextos nacionales manteniendo la coherencia con el estándar europeo (INE, 2006).

El proceso de revisión no solo está relacionado con cambios económicos sino también a la necesidad de mantener la clasificación alineada con estándares internacionales como la Clasificación Industrial Internacional Uniforme (CIIU) de las Naciones Unidas. La NACE y su armonización internacional destacan la importancia de tener clasificaciones comparables que faciliten análisis de competitividad, productividad, comercio exterior y empleo.

La adaptación de la NACE y el desarrollo de versiones como la CNAE subrayan el compromiso de los Estados miembros y la Comisión Europea por actualizar y uniformizar esta clasificación de manera simultánea en todos los Estados miembros. Este esfuerzo entre instituciones nacionales e internacionales, como el INE y Eurostat, junto con la coordinación por parte de divisiones estadísticas de las Naciones Unidas, refleja la dinámica de trabajo conjunto necesaria para enfrentar los retos de clasificar y analizar la economía moderna (Eurostat, 2008).

2.2.4. Reglas de clasificación de actividades y unidades

Cada empresa registrada en los datos estadísticos recibe un código NACE basado en su actividad económica principal, definida como aquella que más contribuye al valor agregado de la empresa. La asignación de este código se basa en las notas explicativas de la NACE, decisiones del comité de gestión de la NACE, tablas de correspondencia y referencias a otros sistemas de clasificación como la CIIU, la CPA, el SA y la CN. En el caso de empresas con una sola actividad económica, su actividad principal está determinada por la categoría de la NACE que describe esa actividad. Sin embargo, si la empresa realiza múltiples actividades económicas, la actividad principal se decide según el valor agregado asociado a cada actividad. El valor agregado, que es la diferencia entre la producción y los consumos intermedios refleja la contribución de contribución al producto interno bruto (PIB) (Eurostat, 2008).

Para la clasificación de las actividades se emplea el método descendente que sigue un principio jerárquico: la clasificación de una unidad en el nivel más bajo de la clasificación debe ser coherente con la clasificación de la unidad en los niveles superiores de la estructura. Para satisfacer esta condición, el proceso comienza con la identificación del nivel más alto pertinente y avanza hacia abajo a través de los niveles de la clasificación de la siguiente manera:

1. Se identifica la sección que tiene la mayor participación en el valor agregado.
2. Dentro de esta sección, se identifica la división que tiene la mayor participación en el valor agregado.
3. Dentro de esta división, se identifica el grupo que tiene la mayor participación en el valor agregado.
4. Dentro de este grupo, se identifica la clase que tiene la mayor participación en el valor agregado.

A continuación, se muestra un ejemplo donde una unidad lleva a cabo las siguientes actividades (participaciones en términos de valor añadido):

Section	Division	Group	Class	Description of the class	Share
C	25	25.9	25.91	Manufacture of steel drums and similar containers	10%
	28	28.1	28.11	Manufacture of engines and turbines, except aircraft, vehicle and cycle engines	6%
		28.2	28.24	Manufacture of power-driven hand tools	5%
		28.9	28.93	Manufacture of machinery for food, beverages and tobacco processing	23%
			28.95	Manufacture of machinery for paper and paperboard production	8%
G	46	46.1	46.14	Agents involved in the sale of machinery, industrial equipment, ships and aircraft	7%
		46.6	46.61	Wholesale of agricultural machinery, equipment and supplies	28%
M	71	71.1	71.12	Engineering activities and related technical consultancy	13%

Figura 1: Ejemplo de clasificación NACE de una empresa según las actividades que desarrolla.

Fuente: (Eurostat, 2008)

Por lo tanto, la clase correcta es la 28,93 Manufacture of machinery for food, beverages and tobacco processing, aunque la clase con mayor participación en el valor añadido es la clase: 46,61 Wholesale of agricultural machinery, equipment and supplies.

Seguidamente, se muestra la ruta de decisión seguida en el ejemplo anterior:

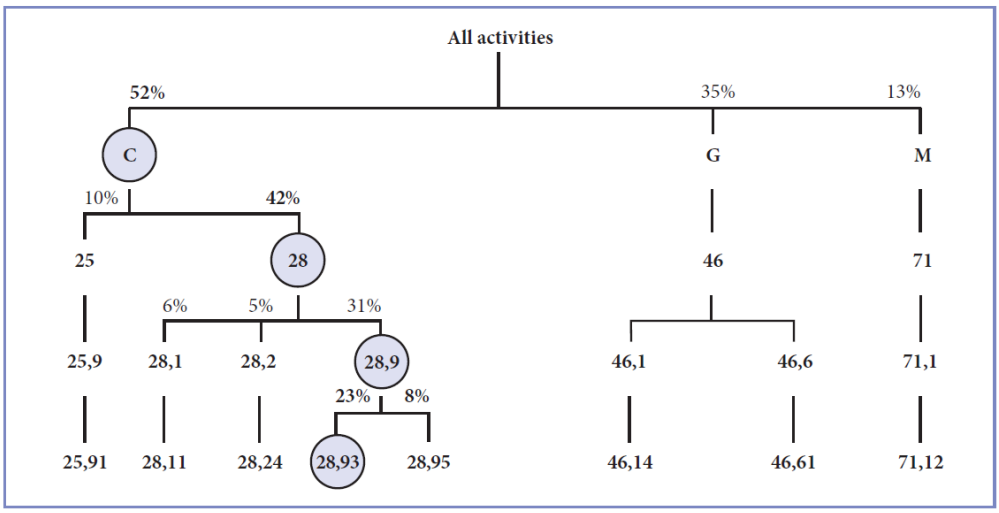

Figura 2: Ejemplo en formato de árbol de clasificación NACE de una empresa según las actividades que desarrolla.

Fuente: (Eurostat, 2008)

2.2.5. Problemáticas de la clasificación NACE

Cada vez hay más coincidencias de las limitaciones de la versión de la NACE Rev.2 actualmente en uso, entre ellas se encuentran:

- La incompatibilidad internacional de la NACE: surge como una problemática debido a la imposición de su uso en la Unión Europea, establecida por los Estados miembros y la Comisión. Aunque los reglamentos que rigen la NACE contemplan la posibilidad de que los Estados miembros desarrollen versiones nacionales derivadas de ella para uso interno, estas deben ajustarse al marco estructural y jerárquico de la NACE. Este enfoque ha llevado a la creación de diferentes versiones nacionales que, aunque siguen la misma lógica general,

pueden presentar discrepancias en cuanto a la clasificación de actividades económicas específicas.

- La desactualización de la NACE: se manifiesta como una problemática latente ante los constantes cambios en las estructuras económicas y los avances tecnológicos que generan la emergencia de nuevas actividades y productos. El informe del proyecto ESSnet resalta estas problemáticas, indicando cómo las actualizaciones recientes buscan mejorar la clasificación para reflejar de manera más precisa y eficiente la realidad económica actual, asegurando la relevancia continua del sistema NACE en el contexto regulatorio y estadístico europeo (Dabrowski, y otros, 2022).
- Dificultades para dar cabida a empresas que abarcan sectores: la clasificación de la NACE es exhaustiva y a la vez excluyente, lo que significa que todas las empresas se clasifican en un solo código. Esto tiene la ventaja de evitar la doble contabilización, pero podría crear dificultades para clasificar las empresas que realizan actividades capturadas en varios códigos NACE.
- Clasificación errónea: En conjunto, todas las razones anteriores llevan a la preocupación de que las propias empresas elijan incorrectamente su código de clasificación. A menudo, la persona responsable de esta tarea no es experta en clasificación de actividades económicas, lo que aumenta la probabilidad de errores. Además, es común que las empresas seleccionen un "código no clasificado en otra parte" incluso cuando hay un código adecuado disponible en la clasificación NACE. También es posible que la proporción de valor agregado entre las distintas actividades de una empresa cambie con el tiempo sin que esto se refleje en un ajuste de su clasificación.

2.3. Trabajos anteriores sobre clasificación de actividad económica a partir de texto

La clasificación automática de actividades económicas a partir de descripciones textuales ha sido un campo de investigación activo, que ha buscado superar las limitaciones de las taxonomías industriales tradicionales mediante el uso de técnicas avanzadas de procesamiento de lenguaje natural y aprendizaje automático. Este apartado revisa trabajos anteriores en esta área, integrando desarrollos en procesamiento de lenguaje natural (NLP) y aprendizaje automático para abordar las limitaciones de los sistemas de clasificación manual.

Uno de los primeros trabajos que se realizaron en este contexto es el realizado por Juan Mateos-Garcia y George Richardson (2022), se enfoca en evaluar y mejorar una taxonomía industrial emergente basada en descripciones de sitios web de negocios en el Reino Unido. La taxonomía propuesta busca superar las limitaciones del sistema SIC (Standard Industrial Classification) tradicional al incorporar una mayor cantidad de códigos SIC4, seleccionando parámetros de agrupamiento mediante evaluaciones consecuentes, y estableciendo medidas de confianza para la asignación de empresas a sectores basados en texto. El estudio utiliza datos web y métodos de aprendizaje automático para refinar la taxonomía y analizar su impacto en la composición sectorial de las economías locales, destacando la importancia de esta nueva taxonomía para informar políticas de desarrollo.

Por otro lado, se realizó también un segundo estudio por Alex Bishop, Juan Mateos-Garcia y George Richardson (2022), que explora cómo los datos de sitios web de empresas pueden usarse para superar las deficiencias del SIC, desarrollando una taxonomía industrial "desde abajo hacia arriba" basada en similitudes semánticas entre las descripciones de las compañías. El estudio propone métodos para descomponer códigos SIC no informativos en industrias más granulares y construir grupos industriales impulsados por el usuario, como la "economía verde". Además, plantea la construcción de índices de composición económica local que se correlacionan más fuertemente con el rendimiento económico local que los basados en la taxonomía SIC.

2.4. Recapitulación

En este capítulo, se ha tratado el objeto social de las empresas, resaltando cómo su declaración precisa y su registro en entidades como el Registro Mercantil y bases de datos especializadas como SABI son esenciales para la estructura y transparencia del de las empresas. La claridad en la definición del objeto social también sirve como un indicativo de las intenciones y capacidades de la empresa dentro del mercado.

Además, se ha discutido la importancia de las clasificaciones de actividades económicas, centrando la atención en la NACE como un estándar europeo que facilita la recopilación, presentación y análisis comparativo de datos económicos a través de distintos países y sectores. La NACE, con su estructura jerárquica y metodología de clasificación, proporciona un lenguaje común que es fundamental para el análisis económico y la formulación de políticas.

El capítulo también ha abordado los retos que conllevan la clasificación manual de empresas en cuanto a su actividad económica, destacando que este tipo de clasificación puede ser propensa a ser imprecisa y subjetiva. Este reto se agrava por la evolución constante del panorama empresarial, que introduce nuevas actividades que pueden no estar contempladas en clasificaciones existentes. Queda justificada la necesidad de métodos alternativos y automáticos para una clasificación más precisa. Finalmente, a través del análisis de trabajos previos, se ha creado un contexto para la investigación.

3. Metodología

En este capítulo se pone en contexto la metodología empleada, esta abarca desde la selección y preparación de la muestra de datos, extraída de la plataforma SABI, hasta la implementación y evaluación de modelos de NLP y algoritmos de aprendizaje automático diseñados para procesar y clasificar textos descriptivos complejos. Este capítulo desglosa el proceso en varias fases: la preparación y preprocesamiento de los datos, la selección de las técnicas y algoritmos para el modelado de temas y el agrupamiento jerárquico, y finalmente, los métodos empleados para evaluar la precisión y eficacia de los sistemas de clasificación desarrollados.

3.1. Muestra

El conjunto de datos principal para nuestro análisis se ha obtenido de la plataforma SABI (SABI, 2024). Esta plataforma está especializada en la recopilación y análisis de información financiera y comercial de empresas españolas y portuguesas (UOC, 2024).

3.1.1. Origen de datos

Se ha determinado que, con el objetivo de extraer una muestra representativa de las empresas que existen actualmente en España, se ha elegido un conjunto de 40000 empresas mediante un Muestreo Aleatorio Simple (MAS). De esta manera encontramos empresas de todos los tamaños y dedicadas a distintos sectores. Por otro lado, la muestra se encuentra en formato Excel.

3.1.2. Variables

En cuanto a las variables que componen la base de datos encontramos:

Nombre: Nombre de la empresa.

Número BvD: Identificador único de la empresa en la base de datos de Bureau van Dijk.

Código primario NACE Rev. 2: Código que clasifica la actividad principal de la empresa según la Clasificación Nacional de Actividades Económicas (NACE Rev. 2).

Código(s) secundario(s) NACE Rev. 2: Código(s) que clasifican las actividades secundarias de la empresa según la Clasificación Nacional de Actividades Económicas (NACE Rev. 2).

Descripción actividad: Breve descripción textual de la actividad principal de la empresa.

La base de datos descargada presenta el siguiente formato:

	Nombre	Número BvD	Código primario NACE Rev. 2	Código(s) segundario(s) NACE Rev. 2	Descripción actividad
1.	KL GRUT 78 SLU	ESB57648966	4332		a). Fabricacion, compraventa al detall y mayorista, y realizacion de toda clase de servicios de carpinteria metalica, lamas, cerrajeria, herreria, automatismo, electricidad, sistemas de seguridad y contra incendios, interfonia, cerramientos, mobiliario metalico y fontaneria....
2.	GENERAL DE ENCOFRADOS Y CIMENTACIONES SL	ESB10394724	4120		LA CREACION, ADQUISICION, REFORMA, EXPLOTACION, POR CUENTA PROPIA, AJENA O EN COMISION DE NEGOCIOS DE CONSULTORIA, TRAMITACION Y GESTION DE PROYECTOS. CREACION, ADQUISICION, REFORMA, EXPLOTACION DE ACADEMIAS Y CENTROS DE
3.	ELECTRO OSMA 2018 SOCIEDAD LIMITADA.	ESB02606168	4321		La realización de instalaciones eléctricas en general, montajes de alta, media y baja tensión, todo lo relacionado con la calefacción, climatización, aire acondicionado tanto particular como industrial, telecomunicaciones, energía solar, montajes electrónicos, comercio de materiales eléctricos y ele
4.	CHANRIL TRADE SL	ESB82716226	6820		VENTA AL POR MAYOR DE TODO TIPO DE ARTICULOS DE MENAJE, HOGAR Y REGALOS.

Figura 3: Formato inicial de la base de datos descargada de SABI.

Fuente: Elaboración propia.

3.2. Preprocesamiento de datos

Para el preprocesamiento de los datos, en primer lugar, se han cambiado los nombres de determinadas columnas de la base de datos para que su manipulación resultase más sencilla, la columna 'Nombre' pasó a llamarse 'Nombre empresa', la columna 'Código primario NACE Rev. 2' se llama 'Código primario' y finalmente, 'Código(s) segundario(s) NACE Rev. 2' se llama 'Códigos secundarios'.

Por otro lado, existían ciertas empresas que contaban con varios códigos secundarios NACE Rev. 2 y que en la base de datos aparecían representados en filas diferentes, generando de esta forma filas con valores faltantes en todas las columnas excepto en la de código secundario. Para solucionar este problema se han fusionado estas filas concatenando los códigos secundarios mediante una coma. También se han eliminado aquellas filas cuya columna 'Descripción de actividad' aparecía vacía ya que no era de utilidad para el estudio.

A continuación, se ha procedido a asignar una letra que corresponde a cada empresa en función de la Sección NACE a la que pertenece. Esto se basa en las dos primeras cifras del código NACE primario, las cuales son extraídas y utilizadas para clasificar las empresas en sectores específicos, tal como se detalla en la clasificación NACE que se muestra en la imagen. Para cada rango de códigos, que representa un sector específico desde la 'A' hasta la 'U', se han definido condiciones utilizando estos dos dígitos. Por ejemplo, el sector 'A', que comprende 'Agricultura, silvicultura y pesca', incluye a todas las empresas cuyos códigos comiencen con '01', '02' o '03'.

Section	Title	Divisions
A	Agriculture, forestry and fishing	01 – 03
B	Mining and quarrying	05 – 09
C	Manufacturing	10 – 33
D	Electricity, gas, steam and air conditioning supply	35
E	Water supply; sewerage, waste management and remediation activities	36 – 39
F	Construction	41 – 43
G	Wholesale and retail trade; repair of motor vehicles and motorcycles	45 – 47
H	Transportation and storage	49 – 53
I	Accommodation and food service activities	55 – 56
J	Information and communication	58 – 63
K	Financial and insurance activities	64 – 66
L	Real estate activities	68
M	Professional, scientific and technical activities	69 – 75
N	Administrative and support service activities	77 – 82
O	Public administration and defence; compulsory social security	84
P	Education	85
Q	Human health and social work activities	86 – 88
R	Arts, entertainment and recreation	90 – 93
S	Other service activities	94 – 96
T	Activities of households as employers; u0ndifferentiated goods- and services-producing activities of households for own use	97 – 98
U	Activities of extraterritorial organisations and bodies	99

Figura 4: Estructura general de clasificación de empresas en sectores según su código primario. Fuente: (Eurostat, 2008)

Para implementar esta clasificación, se emplea la función np.select de NumPy. Esta función selecciona un valor de una lista de posibles valores basándose en una lista de condiciones asociadas. Si los dos primeros dígitos del código de una empresa coinciden con alguna de las condiciones establecidas, se asigna la letra del sector correspondiente a la columna 'sector' de la empresa.

A continuación, se ha realizado el procesamiento de los datos textuales de la columna ‘Descripción de actividad’, el objetivo de este paso es transformar el texto en una forma que sea más fácil de procesar por los algoritmos de aprendizaje automático, eliminando la variabilidad y complejidad innecesarias.

Para realizar el procesamiento de los datos textuales se ha utilizado la librería de Python Spacy (Spacy Developers, 2024). Esta librería, según Vasiliev (2020), ofrece una plataforma robusta para diversas tareas de NLP, como la tokenización, la lematización, y la identificación de entidades nombradas, entre otras. Se ha decidido utilizar spaCy debido a su alto rendimiento en comparación con otras librerías de Python y a su capacidad para manejar grandes volúmenes de texto de manera rápida y precisa. En el libro también se detalla cómo spaCy facilita el análisis y la manipulación de texto mediante una interfaz intuitiva y modelos preentrenados que soportan múltiples idiomas.

Para el procesamiento de la columna ‘Descripción de actividad’ con Spacy se han realizado los siguientes pasos:

- **Carga del modelo de lenguaje y configuración inicial**

  Se carga el modelo de lenguaje 'es_core_news_sm' de spaCy, que es específico para el español. Se desactivan el analizador sintáctico ('parser') y el reconocedor de

entidades nombradas ('ner') para optimizar el rendimiento, ya que las tareas requeridas son principalmente la tokenización y la lematización, y no se necesitan análisis sintáctico ni reconocimiento de entidades.

- **Limpieza de Texto**

La limpieza de texto ha implicado la eliminación de elementos irrelevantes para el análisis, como pueden ser:

- **Normalización de texto**: Se ha convertido el texto a minúsculas para unificar variantes de palabras.
- **Eliminación de signos de puntuación**: Se han eliminado los signos de puntuación ya que no aportan al significado deseado para la clasificación.
- **Eliminación de números**: Se han eliminado los números que aparecen en la columna de descripción de actividad ya que los únicos números que nos interesan para el posterior análisis son aquellos que aparecen en las columnas 'Código primerio' y 'Código Secundario'.
- **Eliminación de espacios extra**: Se han eliminado los espacios, tabulaciones o líneas nuevas extras para dejar un solo espacio entre palabras.
- **Eliminación de acentos**: Se han convertido los caracteres acentuados en su equivalente sin acento. Esto se ha logrado mediante la normalización Unicode NFKD, que descompone los caracteres en sus componentes básicos y elimina los acentos.

- **Tokenización y lematización**

Se han implementado procesos de tokenización y lematización para preparar los datos de texto para su posterior análisis. La tokenización se ha realizado mediante la herramienta spaCy, esta técnica permite descomponer el texto en unidades básicas o tokens, similar a los métodos discutidos en el documento 'Sistema para el pre-procesamiento de textos para el procesamiento del lenguaje natural' (Vila Rodríguez, y otros, 2009). Este proceso es fundamental para separar las palabras y signos de puntuación del texto, permitiendo un tratamiento más eficaz de los datos. Posteriormente, se han aplicado técnicas de lematización, también con spaCy, para reducir las palabras a su forma base o lema, eliminando variaciones morfológicas y centrándose en el significado esencial de los términos. Este paso ha permitido minimizar la redundancia de datos y mejorar la precisión de los algoritmos de PLN.

- **Eliminación de Stop Words**

Por otro lado, se han identificado y eliminado las palabras irrelevantes, conocidas como stopwords, este paso es muy importante para mejorar el procesamiento del lenguaje natural aplicado a textos técnicos. En el artículo de Sarica y Luo (2021) explican cómo desarrollar una lista de stopwords adaptada a cada tipo de texto que se vaya a procesar. En este artículo, los autores destacan la importancia de identificar stopwords que son frecuentes en textos técnicos pero que son poco informativas para tareas específicas de PLN. Utilizan métodos estadísticos para identificar estas palabras a partir de grandes bases de datos de patentes, lo que fue interesante para adaptar técnicas similares en este proyecto.

En el código implementado, se han adaptado las ideas expuestas en el artículo para crear una lista específica de stopwords adecuada al dataset, asegurando que los términos eliminados no aportaran valor semántico relevante, de manera similar a cómo ellos identificaron términos poco informativos en textos de ingeniería.

- **Implementación**

La implementación de estos pasos de preprocesamiento se ha realizado mediante librerías de Python especializadas en procesamiento de lenguaje natural, como NLTK o spaCy. Se ha desarrollado un flujo de preprocesamiento y se han aplicado secuencialmente estas técnicas a cada descripción de actividad empresarial de la base de datos. Una vez llevado a cabo este preprocesamiento, se ha sustituido la columna 'Descripción de actividad' por la nueva columna con las descripciones procesadas y ha pasado a llamarse 'Descripción actividad preprocesada'.

3.3. Topic Modeling

Hoy en día, casi todo el contenido textual está digitalizado y se distribuye a través de numerosas fuentes, creando un gran volumen de información. Estos datos textuales están accesibles universalmente y las actividades de leer, comprenderlos y analizarlos se han convertido en tareas esenciales de la vida cotidiana. El modelado de temas es una técnica estadística utilizada para descubrir y revelar la estructura semántica subyacente en grandes colecciones de documentos. Al analizar el texto, esta técnica identifica temas o tópicos recurrentes en un conjunto de datos, agrupando palabras y frases que comparten una relación temática común. Esta técnica se emplea en áreas como la minería de texto, el análisis de redes sociales, la bioinformática y la ingeniería de software, permitiendo extraer información relevante y patrones ocultos en textos masivos (Kherwa & Bansal, 2019).

Actualmente, los modelos más significativos para el modelado de temas son LDA, NMF, Top2Vec y BERTopic. En el trabajo realizado por Egger y Yu (2022) se comparan estos modelos aplicados a tweets, que se caracterizan por su naturaleza breve, densa y no estructurada. Estos modelos proporcionan perspectivas totalmente nuevas para interpretar fenómenos sociales a través de análisis computacionales, permitiendo a los investigadores en ciencias sociales evaluar y comparar la eficacia de diversas técnicas en un contexto real. Además, este trabajo también explica que modelos como BERTopic y Top2Vec utilizan enfoques basados en embeddings para manejar eficientemente grandes volúmenes de datos y descubrir patrones semánticos intrincados que pueden ser cruciales para la comprensión avanzada de los datos sociales digitales.

Los temas representan descripciones subyacentes dentro de un cuerpo extenso de texto. Por lo general, se espera que documentos centrados en un tema particular incluyan algunas palabras con más frecuencia que otras. Por lo tanto, un modelo de temas revisa estos documentos y forma agrupaciones de palabras que sean similares. En esencia, los modelos temáticos identifican palabras y las agrupan en temas para formar conglomerados.

A continuación, en la figura 5 se ilustra cómo funciona el Topic Modelling.

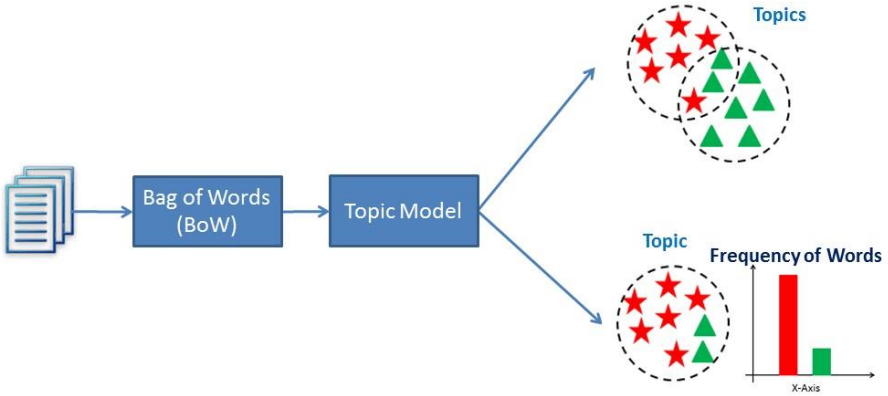

Figura 5: Diagrama de cómo funciona el Topic Modelling.
Fuente: (DataCamp, 2023)

A continuación, se explican detalladamente los distintos modelos que podrían ser útiles para realizar el estudio y se exponen algunos contextos y trabajos anteriores en los que han sido aplicados.

3.3.1. Latent Dirichlet Allocation (LDA)

El modelo de Latent Dirichlet Allocation (LDA) es un método de aprendizaje automático no supervisado, generativo y no parametrizado, presentado como modelo gráfico por Blei, Ng y Jordan (2003). LDA se emplea comúnmente para identificar una cantidad de temas definida por el usuario, que son comunes a los documentos dentro de un conjunto de textos. En este contexto, cada documento representa una observación, las características corresponden a la presencia o frecuencia de cada palabra, y las categorías son los temas. Como LDA es un método no supervisado, los temas no están predefinidos y no hay certeza de que los temas aprendidos sigan las categorizaciones naturales previas de los documentos. Los temas se modelan como distribuciones de probabilidad basadas en las palabras presentes en los documentos, y cada documento se caracteriza por una mezcla de estos temas.

Aunque dos documentos puedan compartir combinaciones similares de temas y, por lo tanto, usar un subconjunto común de palabras con mayor frecuencia que documentos de combinaciones temáticas distintas, el contenido exacto de estos documentos no será idéntico. Este aspecto permite a LDA identificar y agrupar estas palabras comunes para formar los temas.

Un uso práctico de LDA en ingeniería implica la clasificación automática de documentos y la evaluación de su relevancia en relación con diversos temas. En este modelo, se considera que cada documento consta de una combinación de varios temas. Esta característica es similar al análisis semántico latente probabilístico (pLSA), con la diferencia principal de que LDA incluye una distribución previa de Dirichlet para los temas. La utilización de Dirichlet sugiere que cada documento está asociado principalmente a un limitado número de temas y que estos temas, a su vez, utilizan un conjunto reducido de palabras frecuentes. Este enfoque permite una mejor clarificación de las palabras y asigna los documentos a los temas de manera más precisa. LDA, por tanto, es una extensión del modelo pLSA, siendo esencialmente equivalente cuando se emplea una distribución previa uniforme de Dirichlet.

Un ejemplo sencillo de cómo funciona LDA, a partir de un conjunto de documentos en los que las únicas palabras que aparecen en ellos son: comer, dormir, jugar, maullar y ladrar, LDA produciría temas como estos:

Tema	*comer*	*dormir*	*jugar*	*maullar*	*ladrar*
Tema 1	0.1	0.3	0.2	0.4	0.0
Tema 2	0.2	0.1	0.4	0.0	0.3

Figura 6: Ejemplo de cómo funciona LDA.

Fuente: (Amazon Web Services, 2024)

Se puede deducir que los documentos que tienen una mayor probabilidad de formar parte del Tema 1 tienen que ver con los gatos (que es más probable que maúllen y duerman) y es más probable que los documentos que formen parte del tema 2 tengan que ver con los perros (que prefieren jugar y ladrar). Se pueden encontrar estos temas a pesar de que las palabras "perro" y "gato" no aparezcan nunca en ninguno de los textos (Amazon Web Services, 2024).

A continuación, se muestra un diagrama del modelo y se explica su funcionamiento:

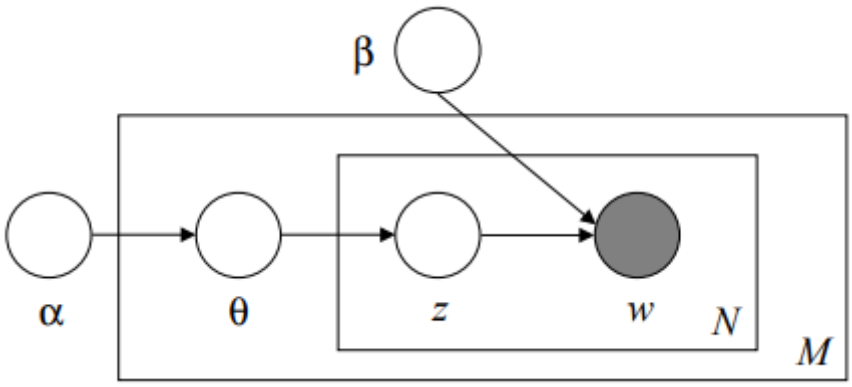

Figura 7: Diagrama del modelo LDA.

Fuente: (Blei, Ng, & Jordan, 2003)

$\propto$: Es un vector de hiperparámetros de la distribución de Dirichlet. Representa los parámetros de la distribución de Dirichlet que genera la distribución de temas θ para cada documento.

θ: Es un vector de probabilidades de los temas en un documento específico. Cada documento tiene su propia distribución de temas, θ_d, que se genera a partir de una distribución de Dirichlet con el parámetro α.

β: Es una matriz de palabras-tema (también llamada matriz de tópicos o distribuciones de palabras), donde cada fila de β es una distribución de probabilidad sobre las palabras para un tema particular.

Z: Es la variable latente que indica el tema asignado para una palabra específica en un documento. Para cada palabra en un documento, se selecciona un tema Z_n de acuerdo con la distribución de temas θ.

W: Es la palabra observada en el documento. Para cada palabra w_n, se selecciona una palabra de acuerdo con la distribución de palabras asociada al tema Z_n , que se describe por β.

M: Representa el número total de documentos en el corpus. El rectángulo grande encapsula todos los documentos, lo que indica que el proceso descrito se repite para cada uno de los M documentos.

N: Representa el número de palabras en un documento específico. El rectángulo más pequeño encapsula w y z, indicando que el proceso se repite para cada una de las N palabras en un documento.

Círculos vacíos: Representan variables latentes o parámetros no observados que el modelo estima durante el entrenamiento (por ejemplo, θ, z).

Círculos llenos: Representan variables observadas en los datos (en este caso, w, que son las palabras en los documentos).

El diagrama representa el proceso generativo de LDA, que es una manera de describir cómo se podrían haber generado los datos (los documentos y las palabras que contienen). En este proceso:

- Para cada documento, se elige una distribución de temas (θ_d) de una distribución de Dirichlet parametrizada por α.
- Para cada palabra en el documento:
  - Se selecciona un tema (z_n) de la distribución de temas del documento (θ_d).
  - Se escoge una palabra (w_n) de la distribución de palabras del tema seleccionado (β_k), que a su vez está parametrizada por η. (Blei, Ng, & Jordan, 2003)

Algunos ejemplos donde se ha aplicado el método Latent Dirichlet Allocation es en el trabajo de Bastani, Namavari y Shaffer (2019) donde explora la aplicación del modelo LDA para analizar y clasificar las quejas de consumidores recibidas por la Consumer Financial Protection Bureau (CFPB). La investigación demuestra cómo LDA puede descubrir temas latentes dentro de grandes volúmenes de narrativas de quejas, facilitando la identificación de tendencias y problemas comunes reportados por los consumidores en relación con productos y servicios financieros. Utilizando el modelo LDA, los autores lograron extraer temas significativos de las narrativas de las quejas, lo que permitió una comprensión más profunda de las inquietudes de los consumidores y la identificación de áreas específicas que podrían necesitar atención regulatoria o mejoras por parte de las instituciones financieras.

Otro ejemplo podría ser el trabajo de Johan Risch (2016) donde el modelo LDA se emplea para analizar y clasificar temas en Twitter. La metodología evaluada se centra en la capacidad del modelo LDA para capturar la distribución temática de los tweets y

su efectividad para asignar temas a mensajes nuevos que no se habían visto anteriormente. Para abordar la naturaleza continua y masiva de los datos de Twitter, se adapta LDA al aprendizaje en línea, permitiendo que el modelo se actualice de manera incremental a medida que se reciben nuevos tweets. Este enfoque permite que LDA maneje efectivamente el flujo constante de datos, manteniendo la relevancia del modelo en la clasificación temática y demostrando ser adecuado para aplicaciones de detección de temas en flujos de datos de tamaño manejable, con sugerencias adicionales para escalar a volúmenes de datos mayores.

Finalmente, en el trabajo de fin de grado de Gonzalo Hernández (2020) se utiliza el modelo LDA para analizar y predecir movimientos del mercado financiero basándose en las noticias. En primer lugar, se prepara un corpus de noticias mediante técnicas de procesamiento del lenguaje natural para luego extraer las palabras más comunes que se usarán para definir los vectores de tópicos. Luego se aplica LDA para transformar la bolsa de palabras en una matriz de distribución de tópicos, donde cada noticia se representa como un vector de distribución de tópicos. Estos vectores se utilizan para entrenar diversos modelos de aprendizaje automático y predecir si el mercado cerrará al alza o a la baja en un día determinado. Los vectores de temas se correlacionan con los movimientos del mercado financiero, intentando prever los cambios en el índice IBEX 35.

3.3.2. Non-negative Matrix Factorization (NMF)

El modelo de Factorización de Matrices No Negativas (NMF, por sus siglas en inglés) es un método de reducción de dimensionalidad y análisis de grupos utilizado en aprendizaje automático. Esta técnica es útil para descomponer una matriz grande en el producto de dos matrices más pequeñas, asegurando que ninguna de estas matrices contenga valores negativos (Núñez Martínez, 2005). Esto es especialmente útil en contextos donde los valores negativos no tienen sentido interpretativo, como en el análisis de datos donde todas las cantidades son positivas por.

A continuación, se ilustra el concepto de Factorización de Matrices No Negativas. Aquí se muestra cómo una matriz grande *V* puede ser aproximadamente factorizada en el producto de dos matrices más pequeñas, *W* y *H*.

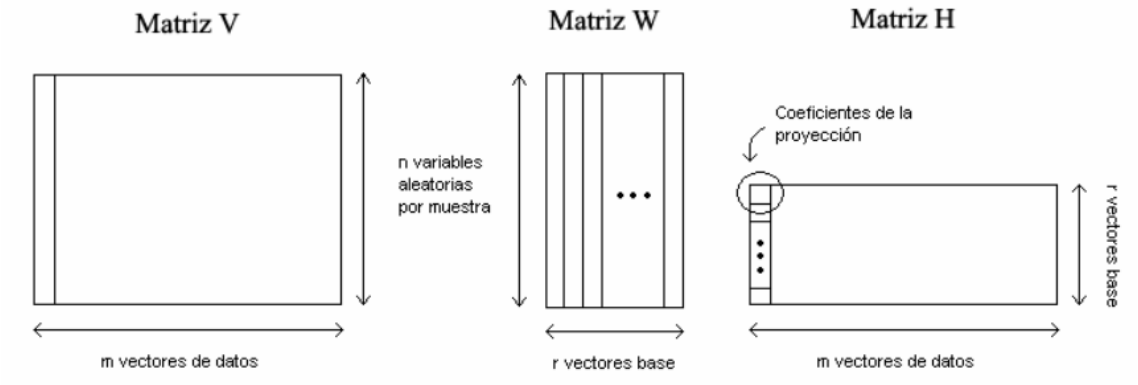

Figura 8: Diagrama del modelo NMF.

Fuente: (Núñez Martínez, 2005)

- **Matriz *W*:** Esta es una matriz de dimensiones más pequeñas cuyas columnas pueden ser interpretadas como componentes base o características latentes del conjunto de datos representado por *V*. En el contexto de análisis de texto, las columnas de *W* representan los temas.
- **Matriz *H*:** Esta es otra matriz más pequeña cuyas filas representan los coeficientes o pesos asociados a cada una de las características latentes en *W* para reconstruir la matriz original *V*. En otras palabras, *H* indica la presencia y la intensidad de las características (columnas de *W*) en cada una de las observaciones originales (columnas de *V*).
- **Matriz *V*:** Esta es la matriz original que se desea descomponer. Puede representar diferentes tipos de datos, en el caso de este proyecto términos de documento en análisis de texto.

La relación $\boldsymbol{V \approx W \times H}$ indica que el producto de $\boldsymbol{W}$ y $\boldsymbol{H}$ es una aproximación de $\boldsymbol{V}$. El objetivo del NMF es encontrar $\boldsymbol{W}$ y $\boldsymbol{H}$ tales que minimicen la diferencia entre $\boldsymbol{V}$ y $\boldsymbol{W \times H}$, típicamente usando una función de pérdida como la distancia Euclidiana o la divergencia de Kullback-Leibler, dependiendo de la naturaleza de los datos y el contexto específico del problema. (Núñez Martínez, 2005)

La Factorización de Matrices No Negativas (NMF), se puede utilizar para el modelado de temas de la siguiente manera: Se empieza con una matriz de entrada V de dimensiones m×n, que representa la matriz término-documento en análisis de texto. Este método descompone V en dos matrices, W y H, donde W tiene dimensiones m×k y H tiene dimensiones k×n.

En este contexto:

- La matriz V representa la matriz término-documento, donde cada fila corresponde a un documento y cada columna a un término específico.
- La matriz H, cada fila se puede considerar como un vector de características o embedding de una palabra.
- La matriz W, cada columna indica la importancia o peso que cada palabra tiene en cada documento, es decir, la relación semántica de las palabras con cada documento.

La condición fundamental de NMF en este escenario es que todos los elementos de las matrices W y H son positivos, dado que todos los valores en V también lo son, lo que es habitual en los datos de conteo de palabras.

Algunos ejemplos donde se ha aplicado el modelo de Factorización de Matrices No Negativas son en el trabajo de Lee y Seung (1999) donde el algoritmo se utiliza para identificar partes constituyentes de objetos en conjuntos de imágenes. Este enfoque se basa en la capacidad de NMF para descomponer una gran matriz de datos de imagen (donde cada columna representa una imagen aplanada) en componentes que representan características comunes o partes de los objetos. Cada componente o columna de la matriz W aprendida por NMF puede interpretarse como una "parte" del objeto que contribuye significativamente a la reconstrucción de las imágenes originales. El objetivo es que estas partes sean lo más independientes posible entre sí, facilitando así una representación más interpretable y esencial de las imágenes en la matriz V.

Por otro lado, en el estudio de los autores Shi, Kang, Choo y Reddy (2018) se utiliza una versión mejorada de NMF para modelar temas en textos cortos, donde la escasez de palabras hace que los métodos tradicionales de modelado de temas sean menos efectivos. En este enfoque, NMF se enriquece con correlaciones de contexto de palabras locales para mejorar la calidad de los temas extraídos. Esto se logra integrando una matriz de co-ocurrencia de palabras en el proceso de factorización, lo que ayuda a capturar mejor las relaciones semánticas entre palabras que aparecen juntas frecuentemente.

3.3.3. BERTopic

BERTopic (Bidirectional Encoder Representations from Transformers) es un modelo de agrupación de temas que utiliza embeddings de documentos generadas por modelos de lenguaje basados en transformers, aplicando después un procedimiento basado en TF-IDF para identificar y representar los tópicos.

TF-IDF (Frecuencia de Término – Frecuencia Inversa de Documento) es una técnica que mide la relevancia de una palabra en un documento en relación con una colección de textos. Se calcula combinando la frecuencia del término en el documento (TF) con la frecuencia inversa del término en toda la colección (IDF), ajustada como log(N/n), donde N es el total de documentos y n es el número de documentos que contienen el término. (Zhai & Massung, 2016)

Esta técnica ofrece varias ventajas, como la capacidad de manejar documentos que evolucionan en el tiempo, permitiendo ajustar las representaciones de tópicos para reflejar cambios temporales sin la necesidad de incrustar y agrupar documentos continuamente, lo que resulta en un proceso más rápido y escalable. Este modelo es capaz de adaptar las representaciones de los tópicos para incorporar diferentes tipos de metadatos, como el autor o la revista en la que se publicó un documento. Esto es útil para analizar cómo los tópicos pueden cambiar en función del contexto en el que se producen los documentos. BERTopic ha sido validado utilizando varios conjuntos de datos, incluyendo 20 NewsGroups, BBC News y tweets de Trump, lo que demuestra su utilidad en una variedad de contextos y tipos de datos (Grootendorst, 2022).

A continuación, se expone un ejemplo simple de cómo funciona BERTopic utilizando como conjunto de datos un grupo de 20 noticias:

En primer lugar, se realiza el preprocesamiento de texto donde los artículos son limpiados y preprocesados. Luego se realiza la transformación con BERT donde cada artículo es convertido en un vector numérico utilizando BERT. Estos vectores capturan la esencia semántica del texto de manera que textos con significados similares están cerca en el espacio vectorial.

A continuación, se realiza la reducción de dimensionalidad ya que los vectores de BERT pueden ser muy grandes y complejos. Los vectores reducidos son agrupados utilizando un algoritmo como HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise). Este paso agrupa los artículos en clusters basados en sus similitudes semánticas. Para cada cluster, BERTopic identifica las palabras y frases más representativas que caracterizan los documentos en ese grupo. Estas palabras clave forman la "etiqueta" del tópico.

Finalmente, los resultados pueden visualizarse y analizarse para entender qué temas son predominantes, cómo se relacionan los documentos entre sí, y cómo los tópicos cambian con el tiempo si se dispone de datos temporales.

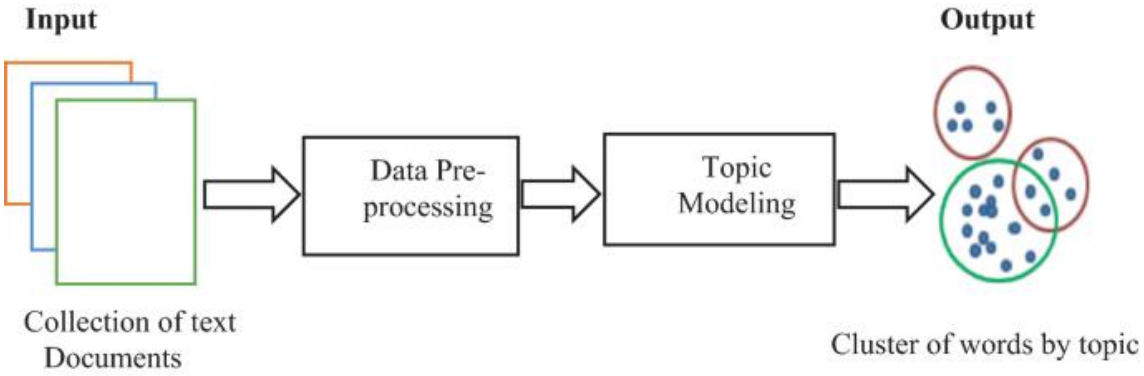

Figura 9: Diagrama del modelo BERTopic.
Fuente: (Motahhir & Bossoufi, 2023)

Algunos ejemplos donde se ha aplicado BERTopic es en el estudio publicado en el Jurnal Media Informatika Budidarma (2023) donde el modelo se aplicó para agrupar 13,027 resúmenes de artículos de Scopus relacionados con el procesamiento del lenguaje natural. Los resúmenes fueron primero preprocesados y luego transformados en vectores utilizando el modelo de incrustación MiniLM-L6-v2. BERTopic agrupó estos vectores en temas coherentes, revelando estructuras y trayectorias temáticas dentro de la literatura académica. El estudio destacó la capacidad de BERTopic para identificar conexiones temáticas y proporcionar una exploración más focalizada de la literatura relevante, facilitando así la identificación de trabajos relacionados y tendencias de investigación emergentes.

Por otro lado, en el estudio de Hutama y Suhartono (2022) BERTopic se combinó con modelos transformadores multilingües (XLM-R y mBERT) para mejorar la clasificación de noticias falsas en Indonesia, un idioma considerado de bajos recursos. BERTopic ayudó a mejorar la representación contextual de las noticias mediante la distribución de temas, que luego se utilizó junto con las características extraídas por los modelos transformadores para clasificar las noticias. Este enfoque no solo mejoró la precisión de la clasificación de noticias falsas, sino que también permitió explorar cómo las distribuciones temáticas pueden enriquecer los modelos de detección de noticias falsas en idiomas con recursos limitados.

Finalmente, en el estudio realizado para el NPJ Digital Medicine (2023), el algoritmo se aplicó para comparar la experiencia de usuario de aplicaciones de salud móvil reguladas y no reguladas en Alemania, analizando tanto las calificaciones promedio de las tiendas de aplicaciones como las reseñas escritas. Esta investigación destacó cómo BERTopic puede revelar diferencias significativas en la percepción del usuario, mostrando que las aplicaciones mHealth reguladas, conocidas como DiGAs, recibieron valoraciones más altas que sus contrapartes no reguladas en términos de servicio al cliente, personalización y facilidad de uso. Además, se identificaron desafíos específicos como errores de software y procesos de registro engorrosos en las DiGAs, mientras que las principales preocupaciones para las aplicaciones no reguladas giraron en torno a los precios excesivos.

3.4. Hierarchical Topic Modelling

El modelado de temas jerárquico es una técnica avanzada en el análisis de textos que supera las limitaciones del modelado de temas tradicional al permitir la modelación de las correlaciones entre temas, lo que no es posible con un enfoque de distribución única en cada documento. Este enfoque utiliza estructuras jerárquicas, como árboles o grafos acíclicos dirigidos (DAG), para organizar los temas de manera que reflejen las relaciones subyacentes entre ellos.

El modelado de temas jerárquico, al igual que el método de clasificación de la NACE, se basa en la idea de organizar información en estructuras jerárquicas para facilitar su análisis y comprensión. En el caso del modelado de temas jerárquico, esta técnica permite capturar y representar las relaciones entre diferentes temas dentro de un conjunto de documentos, estructurándolos en niveles jerárquicos que van desde temas más generales hasta subtemas más específicos. De manera similar, la NACE utiliza una estructura jerárquica para clasificar actividades económicas, comenzando con secciones amplias y descendiendo a divisiones, grupos y clases más detalladas.

Ambos enfoques comparten el objetivo de descomponer ya sea un conjunto de documentos o el conjunto de actividades económicas en componentes organizados de manera jerárquica, donde cada nivel proporciona una mayor especificidad y precisión. En el modelado de temas, esta organización permite entender cómo se relacionan los diferentes temas dentro de un corpus, mientras que en la NACE, facilita la categorización coherente de las actividades económicas.

Además, ambos métodos utilizan la jerarquía no solo para organizar, sino también para mantener la coherencia y evitar la redundancia. En el modelado de temas, los subtemas se construyen a partir de temas más generales, de modo que cada nivel de la jerarquía es consistente con los niveles superiores. De manera análoga, la NACE asegura que la clasificación de una actividad económica en un nivel más detallado sea coherente con su clasificación en niveles más generales, asegurando una estructura lógica y precisa en la clasificación.

Dentro del modelado de temas jerárquico, el hLDA (Hierarchical Latent Dirichlet Allocation) y el PAM (Pachinko Allocation Model) son dos de los modelos más representativos. El hLDA utiliza un proceso conocido como el proceso del restaurante chino anidado (nCRP) para aprender jerarquías de temas de manera no supervisada. En este modelo, los documentos generan temas a lo largo de un camino específico en una estructura de árbol, haciendo que cada documento esté asociado con una ruta de temas desde la raíz hasta las hojas. (Liu, Tang, He, Zhou, & Shaowen, 2016)

El algoritmo ejecuta LDA sobre el corpus original, lo que resulta en un modelo de temas y asignaciones de palabras a temas. Estas asignaciones de palabras a temas se utilizan para crear documentos sintéticos, uno para cada par documento/tema. Los documentos sintéticos se agrupan en corpus sintéticos por tema, y se ejecuta LDA para cada uno de los corpus sintéticos. Este proceso continúa recursivamente hasta que el corpus sintético y los documentos son demasiado pequeños para modelar. El resultado es una jerarquía de distribuciones de temas. (Smith, Hawes, & Myers, 2014)

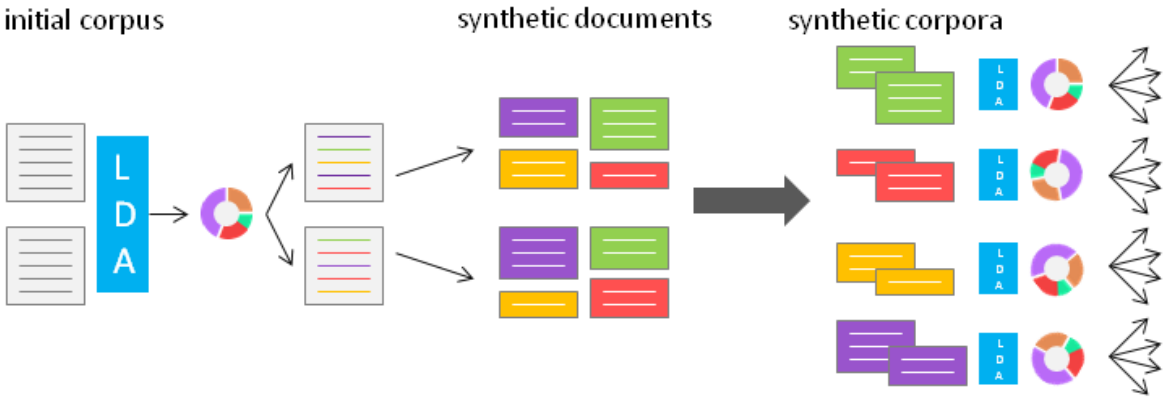

Figura 10: Descripción general del algoritmo HLDA.

Fuente: (Smith, Hawes, & Myers, Hierarchie: Interactive Visualization for Hierarchical Topic Models, 2014)

Por otro lado, el PAM extiende la idea del modelo hLDA al permitir que las relaciones entre los temas se modelen como un Grafo Acíclico Dirigido, donde los nodos interiores representan temas que son distribuciones sobre sus temas hijos, y no solo sobre palabras. Esto permite una flexibilidad mucho mayor en la representación de las relaciones entre temas, aunque a diferencia del hLDA, no todos los nodos en el PAM están asociados con distribuciones sobre palabras (Liu, Tang, He, Zhou, & Shaowen, 2016).

El gráfico que se muestra a continuación titulado "Ejemplo correlación de temas en PAM" ilustra la relación entre supertemas y subtemas. Cada círculo representa un supertema, mientras que los recuadros simbolizan los subtemas. Un supertema puede vincularse a múltiples subtemas, mostrando así su correlación. Los números en los bordes indican los valores de α para cada par (supertema, subtema). (Li & McCallum, 2006)

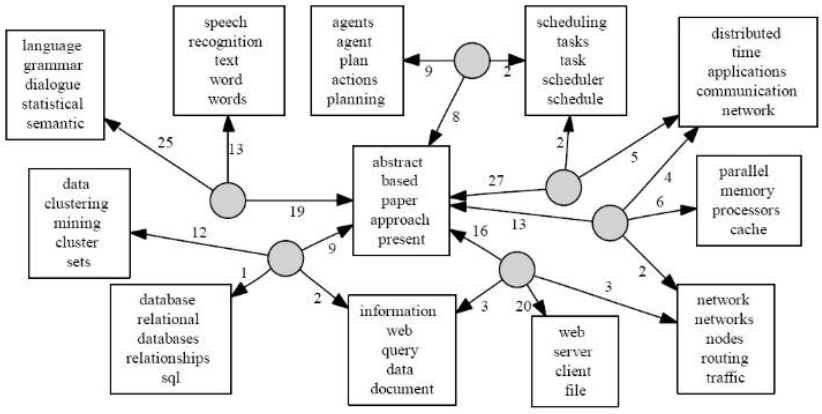

Figura 11: Ejemplo correlación de temas en PAM.

Fuente: (Li & McCallum, 2006)

3.5. Evaluación de la clasificación de actividades

A la hora de clasificar las empresas, la NACE sigue una serie de esquemas de clasificación establecidos mientras que en este trabajo se utilizan los modelos

explicados anteriormente como LDA, BERTopic o NMF. Una vez realizadas las nuevas clasificaciones se compararán entre sí mediante el cálculo de sus entropías.

3.5.1. Entropía

La entropía, en el contexto de la teoría de la información, es un concepto desarrollado por Claude Shannon (1948). La entropía, en esta teoría, se utiliza para medir la incertidumbre o el contenido informativo promedio de un mensaje generado por una fuente de información.

Definición matemática de Entropía:

La entropía H de una fuente de información se define matemáticamente como:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_b P(x_i)$$

donde $P(xi)$ es la probabilidad de que ocurra el evento xi de un conjunto de posibles eventos X, y n es el número de eventos posibles. La base b del logaritmo determina la unidad de medida de la entropía. Si la base es 2, la entropía se mide en bits. (Shannon, 1948)

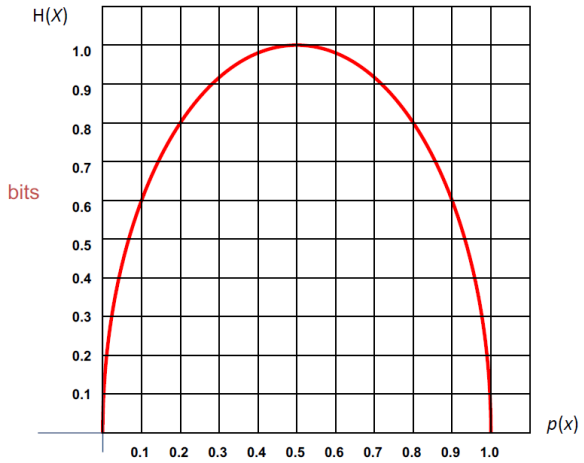

Figura 12: Entropía de la información en un ensayo de Bernoulli X.
Fuente: (Fuentes de Información Entropía, 2021)

La imagen anterior representa la entropía de la información en un ensayo de Bernoulli X (experimento aleatorio en que X puede tomar los valores 0 o 1). La entropía depende de la probabilidad P(X=1) de que X tome el valor 1. Cuando P(X=1)=0.5, todos los resultados posibles son igualmente probables, por lo que el resultado es poco predecible y la entropía es máxima. (Fuentes de Información Entropía, 2021)

Uso de la Entropía como medida de incertidumbre

La entropía mide cuánta información se espera revelar en promedio al observar una variable aleatoria desde una fuente. Una entropía alta significa que hay mucha incertidumbre en la información producida por la fuente, por lo tanto, cada mensaje que se recibe de tal fuente tiene un alto contenido informativo. Por el contrario, una

entropía baja indica que los mensajes son predecibles y contienen menos información (Cover & Thomas, 2006).

En este proyecto se ha utilizado la entropía para evaluar la clasificación de las empresas dentro de los sectores tanto por parte de la NACE como con la utilización de modelos para generar una nueva clasificación. Una vez clasificadas las empresas, se calcula la entropía dentro de cada grupo y finalmente la entropía media, de esta forma se mide como de "bien" o "adecuadamente" están organizados los datos en los grupos designados, basándose en su contenido informativo y diversidad.

Los resultados se interpretan de la siguiente manera:

- **Entropía Alta:** Si la entropía es alta, significa que hay una gran variedad de actividades dentro del grupo, lo que podría indicar una clasificación menos precisa u homogénea. Esto puede sugerir que el grupo es demasiado amplio o que algunas empresas podrían estar mejor clasificadas en otros grupos.
- **Entropía Baja:** Una entropía baja en un grupo sugiere que las actividades de las empresas dentro del grupo son muy similares, indicando una buena clasificación.

De esta forma, se calcula la entropía una vez clasificadas las empresas con distintos modelos y se comparan para determinar qué modelo realiza la mejor clasificación.

3.5.2. Coeficiente de Gini

Para la comparación de las distintas Entropías obtenidas con los métodos de clasificación, en este proyecto se ha utilizado el coeficiente de Gini.

El coeficiente de Gini es un indicador utilizado en economía para medir la desigualdad en la distribución de los ingresos dentro de una población. Se deriva de la curva de Lorenz, que representa gráficamente la distribución acumulada del ingreso en función de la población acumulada ordenada de manera ascendente según sus ingresos.

- **Curva de Lorenz**: Muestra la relación entre la proporción acumulada de la población y la proporción acumulada de ingresos que esa población recibe. Cuanto más se aleja la curva de la línea de igualdad perfecta (una línea diagonal que representa la igualdad total), mayor es la desigualdad.
- **Cálculo del coeficiente de Gini**: Este coeficiente se calcula como la razón entre el área que se encuentra entre la curva de Lorenz y la línea de igualdad perfecta, y el área total bajo la línea de igualdad perfecta.

$$G = \frac{A}{A + B}$$

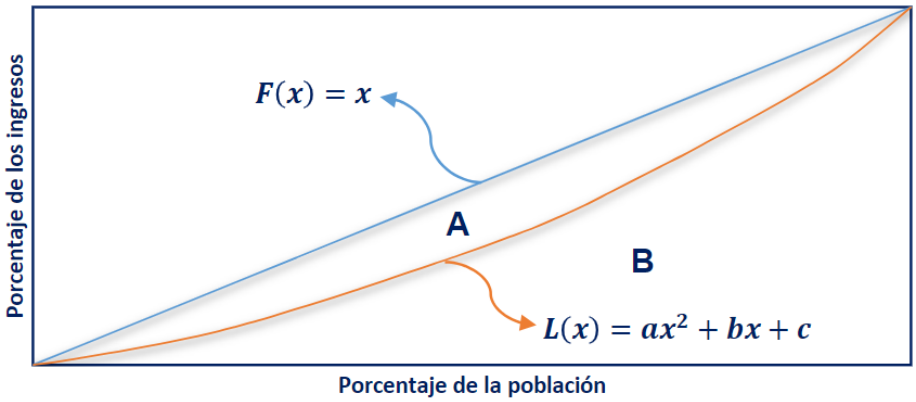

Figura 13: Recta de igualdad perfecta, curva de Lorenz y área de Gini.

Fuente: (Brenes González, 2020)

El valor que toma el coeficiente de Gini varía entre 0 y 1. Un valor de 0 indica perfecta igualdad (donde todos los individuos tienen los mismos ingresos), mientras que un valor de 1 indica máxima desigualdad (donde una sola persona tiene todo el ingreso y los demás no tienen nada). (Brenes González, 2020)

En el contexto de este proyecto en lugar de distribución de ingresos entre la población se utiliza la distribución de entropías entre las distintas secciones en las que clasifica el modelo o la NACE. De esta manera, un coeficiente de Gini alto para uno de nuestros modelos de clasificación implica mayor variabilidad en los grupos.

4. Resultados

En este apartado se exponen los resultados obtenidos tras la aplicación de las distintas técnicas de modelado de datos introducidas en el capítulo anterior. En primer lugar, se ha realizado un análisis descriptivo en el que a través de gráficas y visualizaciones se puede observar la distribución de las empresas en las distintas secciones y la entropía de cada una de ellas clasificadas mediante el método empleado por la NACE. Por otro lado, se han clasificado las empresas según su descripción de actividad mediante las técnicas de modelado de tópicos LDA, BERTopic y NMF y se ha calculado la entropía de cada una de estas nuevas clasificaciones con el fin de comparar estos resultados con los originales. Finalmente, se han comparado los resultados para valorar las mejoras obtenidas respecto a la clasificación original.

4.1. Análisis descriptivo

Se ha realizado un análisis descriptivo de los datos a través de varios métodos estadísticos y de visualización, utilizando herramientas de análisis de datos en Python.

En primer lugar, se ha creado una tabla de frecuencias y un gráfico de barras en los que se puede observar cuantas empresas de la muestra de datos pertenecen a cada sector asignadas siguiendo el método de la NACE. Podemos observar en la gráfica 1 que los sectores con mayor número de empresas son aquellos relacionados con el comercio al por mayor y menor, construcción y manufactura, destacándose especialmente el sector G con 10075 empresas asignadas, este sector tiene por título 'Comercio al por mayor y al por menor; reparación de vehículos de motor y motocicletas'. Le sigue el sector F 'Construcción' con 6693 empresas pertenecientes. Por otro lado, los sectores con menos empresas son el T 'Actividades de los hogares como empleadores de personal doméstico' y la U 'Actividades de organizaciones y organismos extraterritoriales'.

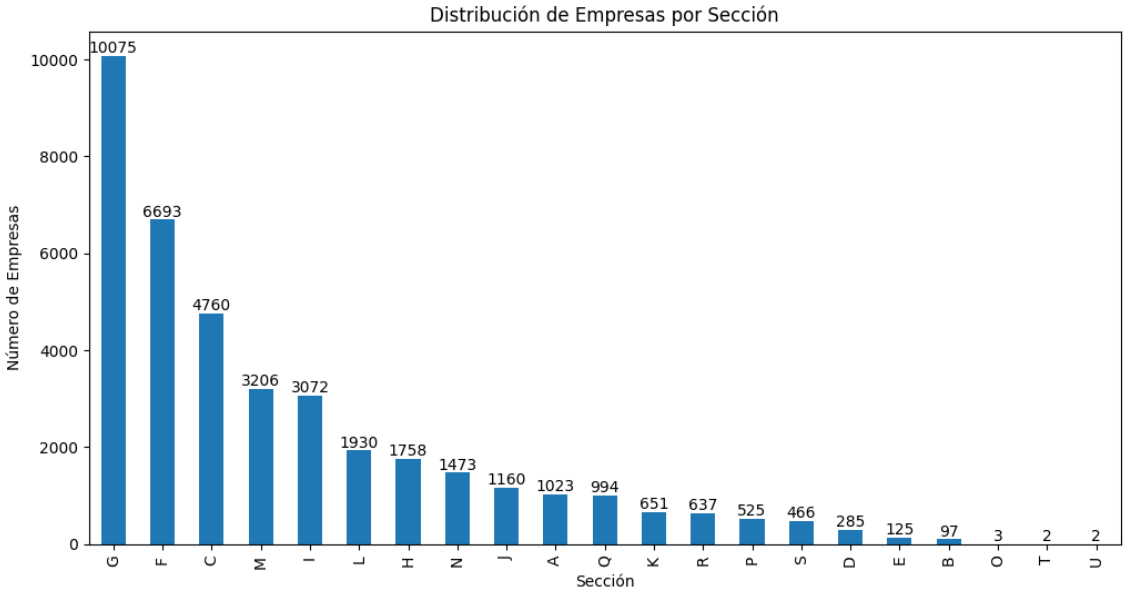

Gráfico 1: Distribución de Empresas por sección asignado con NACE.

Fuente: Elaboración propia.

A continuación, se ha realizado la técnica de nube de palabras para identificar y visualizar las palabras más frecuentes dentro de las descripciones de actividades de las empresas por sector. Con este análisis se pueden deducir las actividades principales de las empresas asignadas a cada sector y comprobar si coincide con el título asignado por la NACE. Además, este tipo de gráfico ayuda a entender los términos más comunes y su relevancia dentro de cada sector. Seguidamente, se muestran las nubes de algunos sectores:

Gráfico 2: Nube de palabras para la sección A con NACE.

Fuente: Elaboración propia.

Nube de palabras sector A: Como se puede observar en la nube de palabras, los términos que más se repiten en las descripciones que componen las empresas de este sector, son: explotación, agrícola, cultivo, ganadero, producto agrícola entre otros. El sector A tiene por nombre 'Agricultura, ganadería, silvicultura y pesca', por lo que las palabras clave de este sector coinciden con su título.

Gráfico 3: Nube de palabras para la sección C con NACE.

Fuente: Elaboración propia.

Nube de palabras sector C: Como se puede observar en la nube de palabras, las palabras más frecuentes en las descripciones son: fabricación, compra, venta, menor, mayor, compraventa o exportación entre otros. El sector C tiene por nombre 'Manufactura', por lo que existe una fuerte correlación entre el título y las palabras clave de las descripciones.

Gráfico 4: Nube de palabras para la sección I con NACE.

Fuente: Elaboración propia.

Nube de palabras sector I: Como se puede observar en la nube de palabras, los términos que más se repiten en las descripciones de este sector, destacan: bar, cafetería, restauración, restaurante, hostelería, comida entre otros. El sector I tiene por nombre 'Actividades de alojamiento y restauración, por lo que las palabras clave de este sector están fuertemente relacionadas con el título asignado.

Similar al análisis por sector, se ha realizado una visualización de los códigos primarios más comunes, gráfico 5. Los códigos primarios clasifican las actividades empresariales en categorías específicas, con este análisis se pueden identificar las áreas de actividad predominantes.

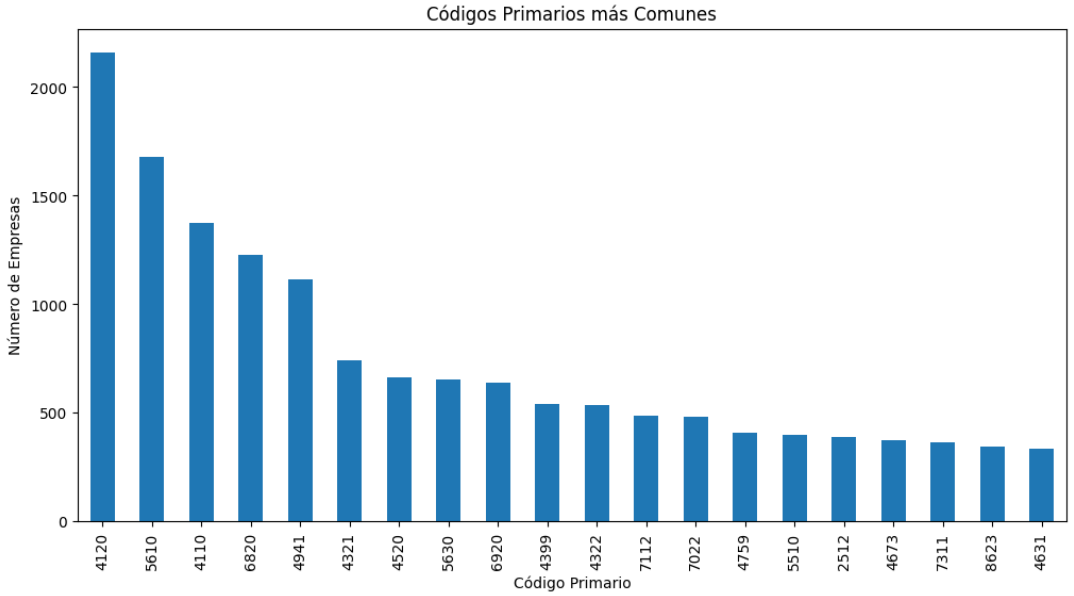

Gráfico 5: Códigos primarios más comunes.

Fuente: Elaboración propia.

Se puede observar cómo los códigos más frecuentes son: el código primario NACE 4120 que pertenece al tipo "Construcción de edificios residenciales y no residenciales", el código NACE 5610 que corresponde a "Restaurantes y puestos de comidas" y finalmente el código NACE 4110 corresponde a la "Promoción inmobiliaria".

Finalmente, para acabar con este análisis descriptivo, se ha calculado el mínimo, máximo y promedio de la longitud de las descripciones de las actividades de las empresas sobre la base de datos previamente procesada. Este análisis cuantitativo proporciona información sobre el nivel de detalle que las empresas utilizan para describir sus actividades, con una longitud media de aproximadamente 12 palabras, una mínima de 1 y una máxima de 83.

4.2. Entropía por sección NACE

En este apartado, se ha calculado la Entropía para cada sección asignada mediante la metodología NACE y se ha comparado con la Entropía obtenida clasificando las empresas utilizando los Topic Modeligs descritos en el capítulo anterior.

4.2.1. Cálculo de la Entropía para la clasificación realizada por la NACE

En primer lugar, para calcular la entropía dentro de cada sección NACE, se han obtenido para cada sección los correspondientes vectores TF-IDF (Frecuencia de Término - Frecuencia Inversa de Documento) de las descripciones de actividades. Para esto, se usa la clase TfidfVectorizer para transformar la columna 'Descripción actividad preprocesada' en una matriz TF-IDF. Esta matriz mide la importancia de un término en un documento en relación con una colección de documentos o corpus. El peso TF-IDF aumenta proporcionalmente con el número de veces que un término aparece en el documento, pero se compensa por la frecuencia del término en el corpus.

Una vez calculada la matriz, se ha definido la función para calcular la entropía de un vector TF-IDF. Esta función calcula la entropía como la suma del producto de cada valor TF-IDF por el logaritmo de ese valor.

Después, se aplica esta función de entropía a cada sección. Para cada grupo, se filtran las filas correspondientes, se extraen las puntuaciones TF-IDF, y después se calcula la entropía. Este proceso se realiza iterando sobre cada sección única.

En cuanto a los resultados obtenidos, podemos observar en la gráfica 6 como el sector G que se dedica al comercio al por mayor y al por menor; reparación de vehículos de motor y motocicletas, es el que presenta una mayor entropía, 34960,84. Podemos observar también en la tabla de frecuencias 7, que esta misma sección es la que agrupa un mayor número de empresas, 10075. Seguidamente, el sector F (destinado a la construcción) es la segunda sección con mayor entropía, 26153,12, además de ser la segunda sección en recopilar un mayor número de empresas, 6693.

Por el contrario, las secciones T destinada a las actividades en el hogar y U destinada a las actividades de organizaciones son las secciones con menor entropía, ambas 7, y también las que menos secciones agrupa, 7,32 y 7,17 empresas respectivamente.

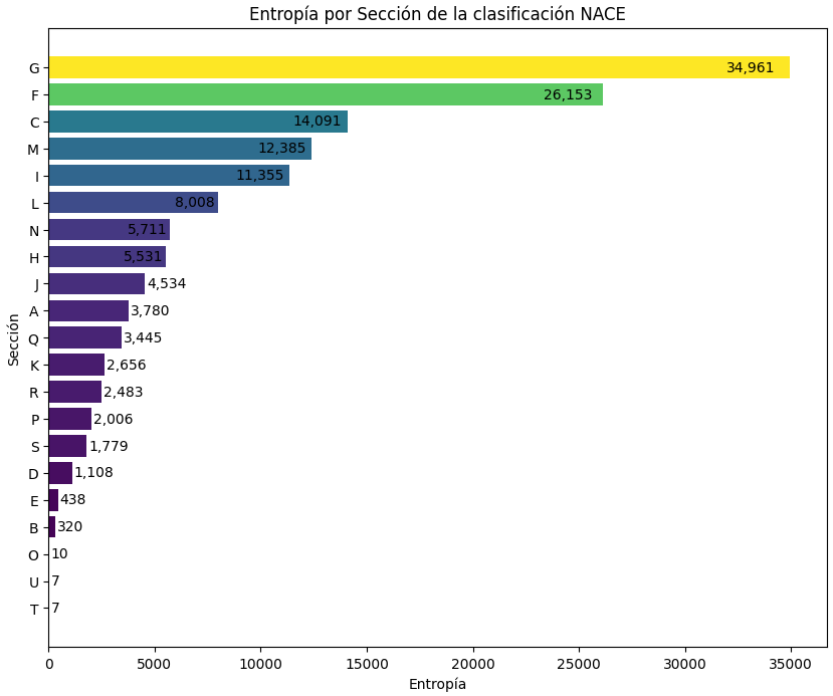

Gráfico 6: Entropía por sección de la clasificación NACE.

Fuente: Elaboración propia.

Sección	Número de Empresas	Entropía
U	2	7,323720
T	2	7,173614
O	3	10,263470
B	97	319,586811
E	125	438,345477
D	285	1108,341864
S	466	1778,785915
P	525	2005,505350
R	637	2482,659043
K	651	2655,782922
Q	994	3445,231601
A	1023	3780,066.851
J	1160	4534,015265
N	1473	5710,674484
H	1758	5530,719799
L	1930	8008,413600
I	3072	11354,987337
M	3206	12385,260737
C	4760	14091,468503
F	6693	26153,123905
G	10075	34960,847215

Gráfico 7: Tabla de frecuencias con el número de empresas y entropía por sección NACE.

Fuente: Elaboración propia.

En el gráfico 8 de dispersión se puede observar cómo existe una correlación entre la entropía y el número de empresas por sección. Por un lado, en el eje X se muestra la entropía que va de 0 hasta 35000 y en el eje Y están el número de empresas que va de 0 hasta 10000. Los puntos representan las 21 secciones y en este caso están dispersos, pero con una tendencia notable en la que los sectores con menor entropía tienden a tener un menor número de empresas. Los puntos parecen formar una ligera curva ascendente, sugiriendo que a medida que aumenta la entropía, también puede aumentar el número de empresas.

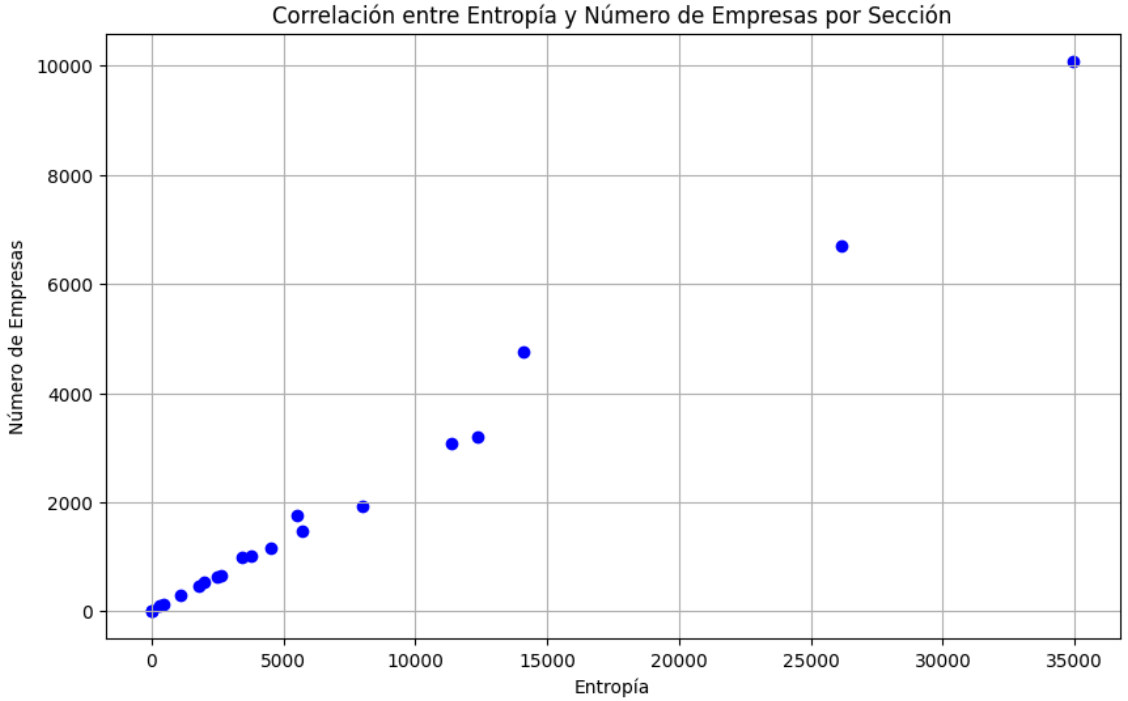

Gráfico 8: Correlación entre Entropía y Número de Empresas por Sección NACE.

Fuente: Elaboración propia.

Finalmente, se ha calculado la entropía media para la clasificación NACE que es de 6703,26, la entropía mediana que es 3445,23, la entropía mínima que es 7.17, la entropía máxima que es 34960,84 y el coeficiente de Gini que es 0.62 para posteriormente comparar estos datos con los que se obtengan con la clasificación realizada mediante Topic Modelling.

4.2.2. Cálculo de la Entropía para la clasificación realizada con el modelo LDA

En esta sección, se han clasificado las empresas en secciones sin tener en cuenta su código primario, es decir, utilizando la técnica de LDA (Latent Dirichlet Allocation).

El primer paso que se ha realizado ha sido la vectorización de las descripciones de actividad, se ha utilizado la función CountVectorizer para convertir la columna 'Descripción actividad' en una matriz de frecuencias de palabras. A continuación, se ha ajustado el modelo con 21 tópicos equivalente a las 21 secciones en las que la NACE clasifica las actividades.

Una vez ejecutado el modelo, se ha asignado a cada descripción el tópico más probable y en el gráfico 9 se puede observar cómo han quedado clasificadas las empresas.

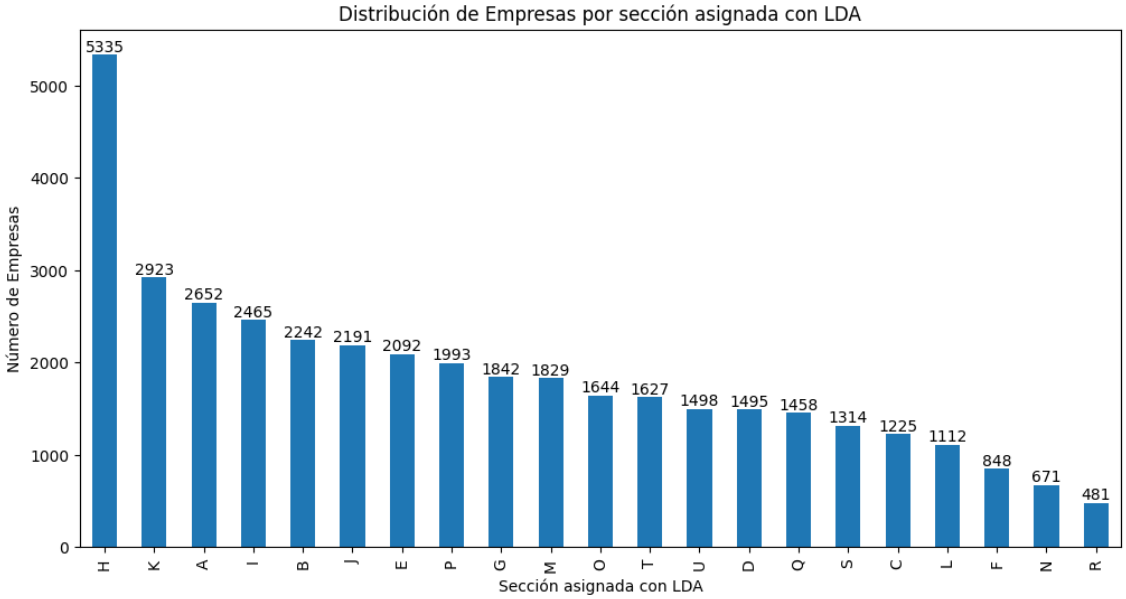

Gráfico 9: Distribución de empresas por sección asignada con LDA.

Fuente: Elaboración propia.

Para cada sección se ha generado una nube de palabras para poder tener una idea de que tipo de empresas se han agrupado en cada una de ellas. La sección que se compone de un mayor número de empresas es la H, con 5335, que si nos fijamos en su nube de palabras 10 engloba aquellas empresas relacionadas con el mercado inmobiliario ya que algunas de sus palabras clave son: compra, venta, alquiler, finca, urbano o promoción entre otras. Le sigue la sección K, con 2923 empresas, que según su nube de palabras esta sección está destinada a la industria hostelera ya que abarca términos como restaurante, bar, cafetería, hostelería, y servicio entre otros.

Gráfico 10: Nube de Palabras para la sección H.

Fuente: Elaboración propia.

Gráfico 11: Nube de Palabras para la sección K.

Fuente: Elaboración propia.

Por otro lado, las secciones con un menor número de empresas son las N y la R. La sección N agrupa 671 empresas que engloba aquellas empresas dedicadas procesos de fabricación y construcción ya que en su nube de palabras 12 aparecen términos como construcción y fabricación, materiales como la madera y el metal, y procesos como el corte, montaje y comercialización. La sección R es la que menos empresas abarca, 481, en esta sección se agrupan empresas relacionadas con servicios y gestión en el sector turístico y comercial.

Gráfico 12: Nube de Palabras para la sección N.

Fuente: Elaboración propia.

Gráfico 13: Nube de Palabras para la sección R.

Fuente: Elaboración propia.

Una vez clasificadas las empresas en las secciones que ha estimado el modelo LDA se ha seguido el mismo procedimiento que en el apartado anterior cuando las secciones eran asignadas por la NACE. Se han obtenido para cada sección los correspondientes vectores TF-IDF de las descripciones de actividad, se ha definido la función para calcular la entropía de un vector TF-IDF y se ha aplicado para cada sección asignada con LDA.

Con respecto a los datos obtenidos, la gráfica 14 muestra que el sector H, que abarca empresas relacionadas con el mercado inmobiliario como hemos visto en su nube de palabras, tiene la mayor entropía con un valor de 2391,99. Igualmente, la tabla de frecuencias 15 indica que este sector contiene el mayor número de empresas, sumando un total de 5335. En segundo lugar, el sector K, dedicado a la industria hostelera, presenta la segunda mayor entropía, con un valor de 1002,39, y también ocupa el segundo lugar en cuanto al número de empresas, con 2923. Por otro lado, las secciones R y N, muestran la menor entropía, ambas con un valor de 1033,44 Y 1509,66 respectivamente, y también tienen el menor número de empresas, con 481 y 671 empresas cada una.

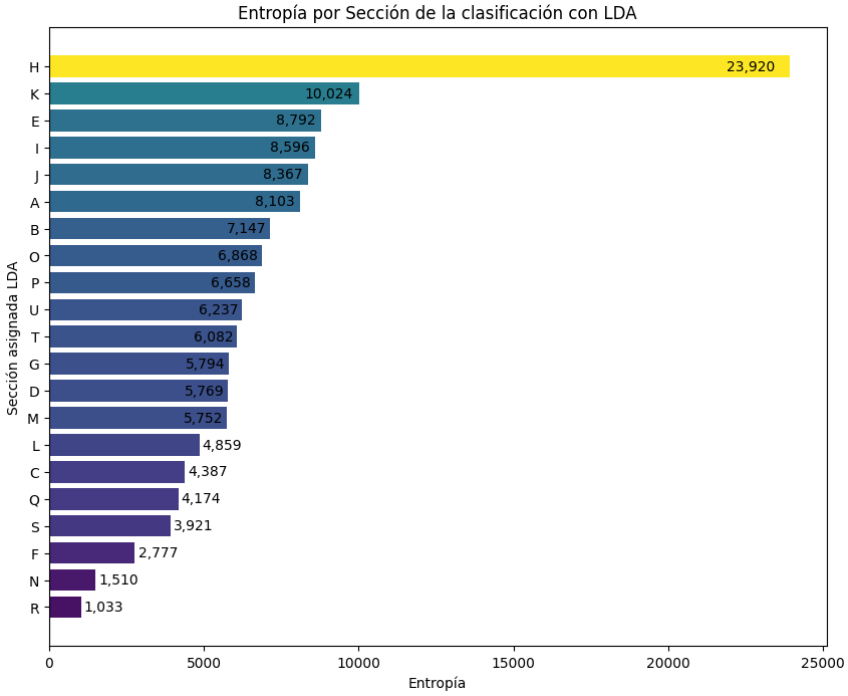

Gráfico 14: Entropía por sección de la clasificación LDA.

Fuente: Elaboración propia.

Sección asignada LDA	Número de Empresas	Entropía
R	481	1033,445318
N	671	1509,657339
F	848	2776,827467
L	1112	4858,935627
C	1225	4387,287768
S	1314	3920,799052
Q	1458	4174,201269

D	1495	5768,602287
U	1498	6236,959315
T	1627	6081,80478
O	1644	6867,923719
M	1829	5751,741394
G	1842	5793,739235
P	1993	6658,211558
E	2092	8791,517806
J	2191	8366,982311
B	2242	7147,124519
I	2465	8595,882139
A	2652	8103,048965
K	2923	1002,391137
H	5335	2391,997424

Gráfico 15: Tabla de frecuencias con el número de empresas y entropía por sección LDA.

Fuente: Elaboración propia.

También se ha obtenido el gráfico de dispersión 16 que ilustra la correlación entre la entropía y el número de empresas por sección. En el eje horizontal, la entropía se mide desde 0 hasta aproximadamente 25,000. En el eje vertical, el número de empresas varía de 0 a cerca de 5,000. Los puntos en el gráfico representan 21 secciones diferentes, distribuidos de manera que reflejan una tendencia ascendente. La mayoría de los puntos están agrupados en la parte inferior izquierda del gráfico, indicando que las secciones con entropías más bajas tienden a tener menos empresas. A medida que la entropía aumenta, el número de empresas también parece crecer, como se observa en el punto más distante a la derecha, que tiene la mayor entropía y el mayor número de empresas. El patrón de dispersión sugiere que podría existir una relación positiva entre la entropía y el número de empresas, aunque esta relación no parece ser estrictamente lineal, dada la dispersión y agrupación variada de los puntos en todo el gráfico.

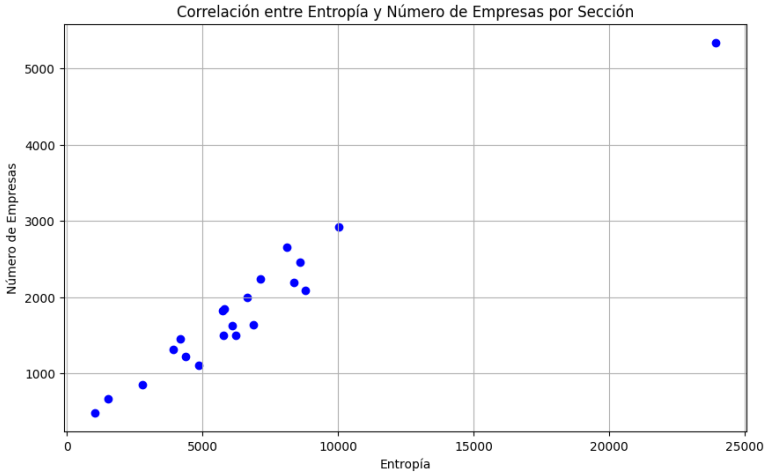

Gráfico 16: Correlación entre Entropía y Número de Empresas por Sección LDA.

Fuente: Elaboración propia.

Finalmente, se ha calculado la entropía media para la clasificación con LDA que es de 6703,26, la entropía mediana que es 6081,80, la entropía mínima que es 1033,44, la entropía máxima que es 23919,97 y el coeficiente de Gini que es 0.30.

4.2.1. Cálculo de la Entropía para la clasificación realizada con el modelo BERTopic

En este apartado, se han clasificado las empresas en secciones utilizando la técnica de modelado de temas de BERTopic (Bidirectional Encoder Representations from Transformers).

El primer paso que se ha realizado ha sido instalar la librería de Python correspondiente para poder utilizar el modelo de BERTopic, a continuación, se han extraído las descripciones de actividad de las empresas y se han almacenado en una lista. Luego, se ha creado una instancia del modelo BERTopic, configurado para identificar 21 tópicos diferentes (las 21 secciones que existen en la NACE).

Después de ejecutar el modelo, se ha asignado a cada descripción el tópico más probable. En el gráfico 17, se muestra la clasificación resultante en secciones de las empresas.

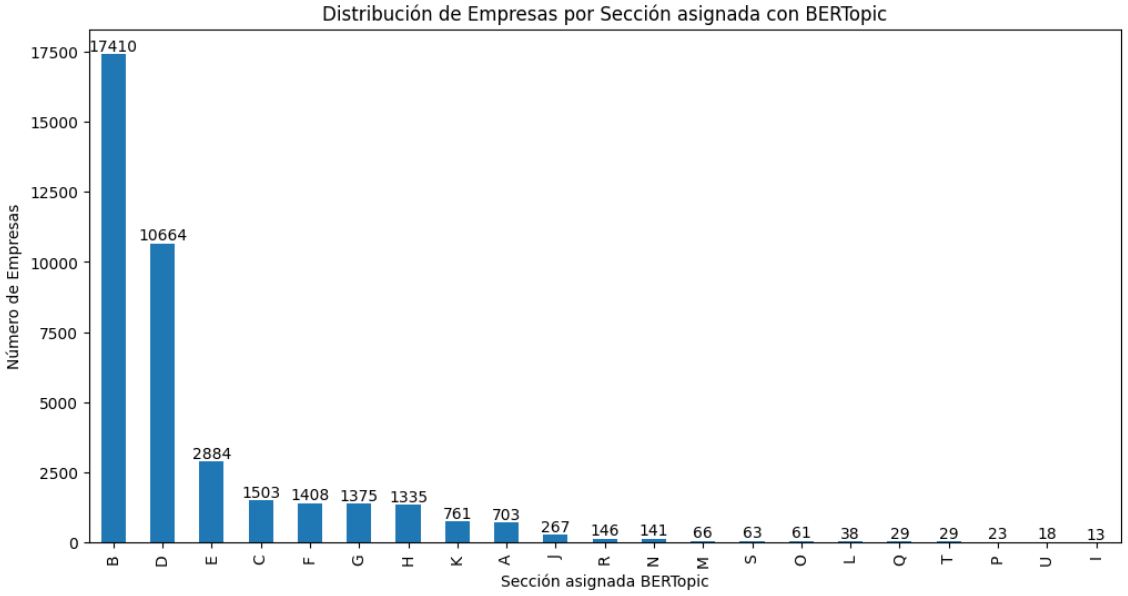

Gráfico 17: Distribución de empresas por sección asignada con BERTopic.

Fuente: Elaboración propia.

Para cada sección, se creó una nube de palabras para identificar el tipo de empresas agrupadas en cada una. La sección B, que incluye el mayor número de empresas, con un total de 17410, muestra en su nube de palabras 18 términos relacionados con el comercio y la construcción como compra, venta, importación, exportación, y construcción, mostrando la amplitud de actividades comerciales y de construcción presentes en el sector. Por otro lado, la sección D, que agrupa a 1664 empresas, está vinculada también al comercio y la construcción, como se refleja en su nube de palabras 19 que incluye términos como compra, venta, importación, exportación, servicio, y construcción. El hecho de que haya varios sectores cuya nube de palabras agrupe

palabras similares o iguales indica que son palabras genéricas en el ámbito empresarial como puede ser ‘compra’, ‘venta’ o ‘servicio’.

Gráfico 18: Nube de Palabras para la sección B.

Fuente: Elaboración propia.

Gráfico 19: Nube de Palabras para la sección D.

Fuente: Elaboración propia.

Por otro lado, las secciones con el menor número de empresas incluyen las secciones I y U. La sección I cuenta con 18 empresas, que se concentran en los procesos de fabricación y montaje con tecnología láser. Esto se refleja en su nube de palabras 20, que incluye términos como corte, pieza, montaje, láser y metálico. En cuanto a la sección U, con solo 12 empresas, agrupa a compañías enfocadas en la enseñanza ya que sus palabras clave son enseñanza, formación, colegio, estudio, escuela o idioma.

Gráfico 20: Nube de Palabras para la sección I.

Fuente: Elaboración propia.

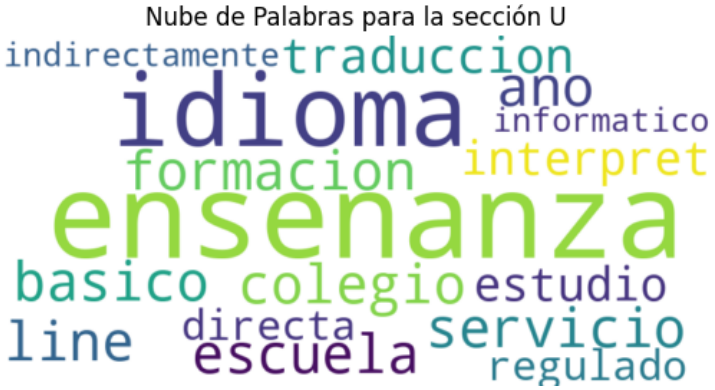

Gráfico 21: Nube de Palabras para la sección U.

Fuente: Elaboración propia.

Después de clasificar las empresas en las secciones determinadas por el modelo BERTopic, se ha aplicado el mismo método usado anteriormente cuando las secciones se asignaban según la clasificación NACE. Para cada sección, se han generado los vectores TF-IDF correspondientes a partir de las descripciones de las actividades. Luego, se ha establecido una función para calcular la entropía de estos vectores TF-IDF, y se ha aplicado a cada sección determinada por BERTopic.

En cuanto a los datos obtenidos, la gráfica 22 muestra que el sector B, enfocado en comercio y la construcción como se indica en su correspondiente nube de palabras, registra la entropía más alta con un valor de 63648,46. Además, la tabla de frecuencias 23 indica que este sector incluye la mayor cantidad de empresas, alcanzando un total de 17410. Por otro lado, el sector D, También vinculado a la industria y la construcción, ocupa el segundo lugar tanto en entropía, con un valor de 41209,31, como en número de empresas, sumando 10664. Por otro lado, las secciones U y T muestran la menor entropía, con valores de 10,54 y 11,03 respectivamente, y también son dos de las secciones con menos empresas asignadas, 18 y 29 respectivamente.

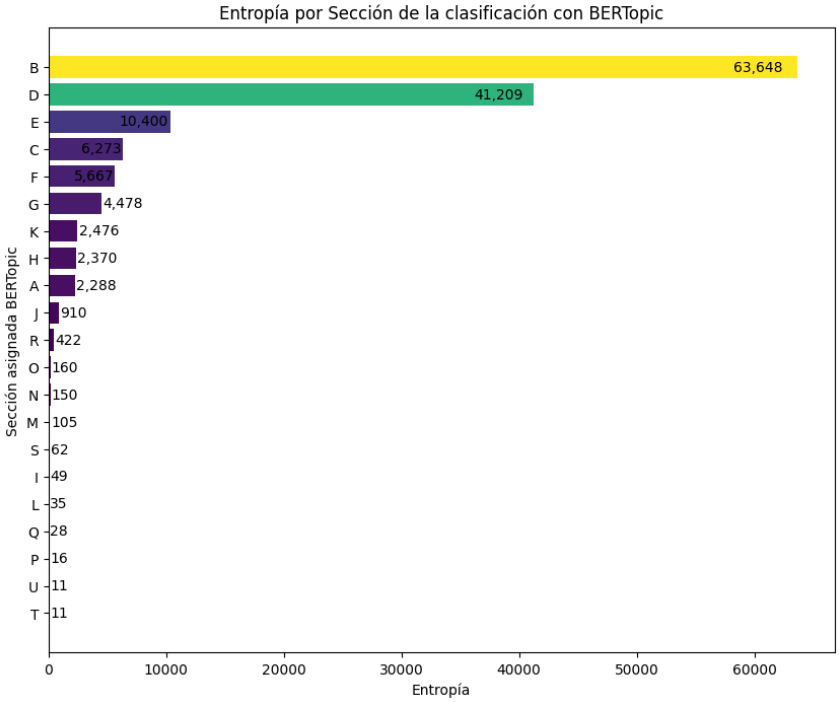

Gráfico 22: Entropía por sección de la clasificación BERTopic.

Fuente: Elaboración propia.

Sección asignada BERTopic	Número de Empresas	Entropía
I	13	48,746987
U	18	11,029099
P	23	15,842948
T	29	10,540268
Q	29	27,912310
L	38	35,434419
O	61	160,493942
S	63	61,616018
M	66	104,561412
N	141	150,181991
R	146	422,383145
J	267	910,230279
A	703	2288,350382
K	761	2475,725657
H	1335	2370,436234
G	1375	4477,690391
F	1408	5666,516482
C	1503	6273,468443
E	2884	10399,638472
D	10664	41209,317866
B	17410	63648,460739

Gráfico 23: Tabla de frecuencias con el número de empresas y entropía por sección BERTopic.

Fuente: Elaboración propia.

Se ha elaborado un gráfico de dispersión 24 que muestra la relación entre la entropía y la cantidad de empresas por sección asignadas por el modelo BERTopic. La escala de entropía, representada en el eje horizontal, abarca desde 0 hasta aproximadamente 60,000, y el número de empresas, representado en el eje vertical, varía de 0 a 17,500. En el gráfico se pueden observar varios puntos distribuidos principalmente a lo largo de dos zonas: una agrupación de puntos cerca del origen, donde tanto la entropía como el número de empresas son bajos, y unos pocos puntos dispersos hacia la derecha, indicando un aumento en ambos valores. El punto más alejado hacia la derecha, situado en el extremo superior, representa la sección con la entropía más alta y el mayor número de empresas. La dispersión de los puntos sugiere que a medida que aumenta la entropía, tiende a aumentar también el número de empresas, aunque no de manera uniforme, lo que podría indicar una relación positiva pero no necesariamente lineal entre estas dos variables.

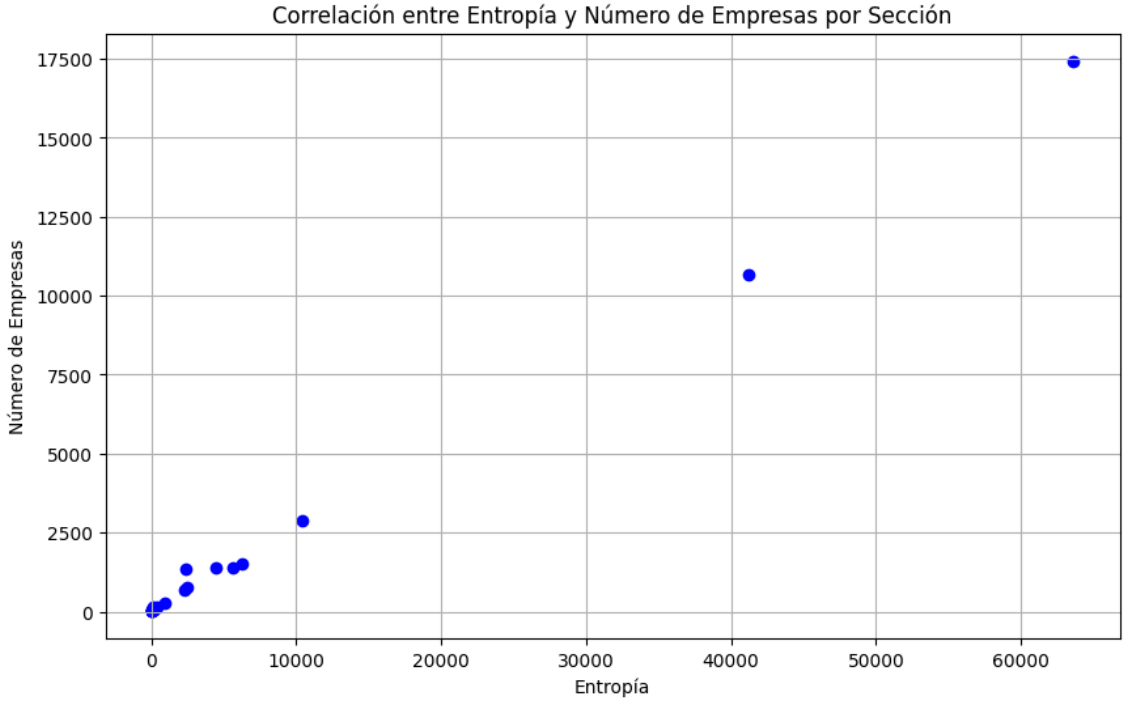

Gráfico 24: Correlación entre Entropía y Número de Empresas por Sección BERTopic.

Fuente: Elaboración propia.

Finalmente, se ha calculado la entropía media para la clasificación con BERTopic que es de 6703,26, la entropía mediana que es 422,38, la entropía mínima que es 10,54, la entropía máxima que es 63648,46 y el coeficiente de Gini que es 0.79.

4.2.2. Cálculo de la Entropía para la clasificación realizada con el modelo NMF

En este apartado, se han clasificado las empresas en secciones utilizando la factorización de matrices no negativas (NMF).

Primero, se ha transformado el texto en vectores utilizando CountVectorizer para que el modelo pueda procesarlo. Después, se ha configurado el modelo NMF con 21 grupos diferentes y se ha ajustado a los datos preparados. Este modelo identifica qué grupo o tema es el más relevante para cada descripción. Una vez que el modelo procesa los

datos, asigna a cada descripción un tema principal. Se puede observar cómo han quedado clasificadas las empresas en el gráfico 25.

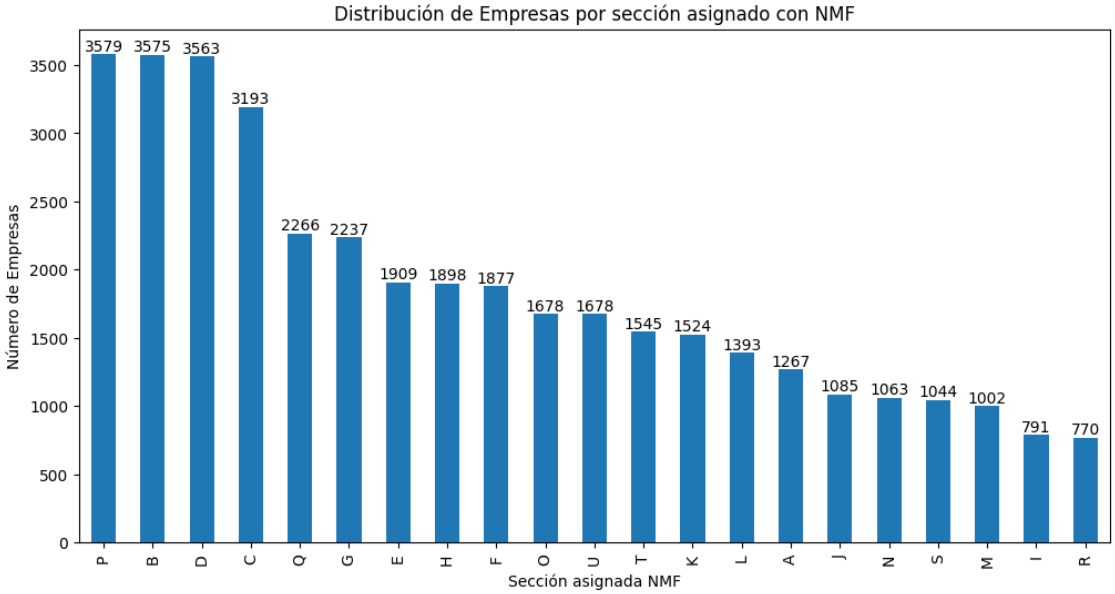

Gráfico 25: Distribución de empresas por sección asignada con NMF.

Fuente: Elaboración propia.

A continuación, se han creado nubes de palabras para cada sección con el objetivo de identificar los tipos de empresas agrupadas en cada una. La sección P, que incluye el mayor número de empresas con un total de 3579, muestra en su nube de palabras términos asociados a la construcción y el inmobiliario, como construcción, promoción, reparación, edificio, vivienda, urbanización, entre otros. En segundo lugar, la sección B cuenta con 3575 empresas y su nube de palabras muestra términos muy similares a los obtenidos en la sección P, incluyendo palabras como promoción, construcción, edificio, edificación o material entre otras.

Gráfico 26: Nube de Palabras para la sección P.

Fuente: Elaboración propia.

Gráfico 27:Nube de Palabras para la sección B.

Fuente: Elaboración propia.

Por otro lado, las secciones que incluyen un número menor de empresas son la I y la R. La sección I, con 791 empresas, abarca aquellas enfocadas en el comercio y la industria. Esto se observa en su nube de palabras, que incluye términos como comercio, venta y compra, producto, explotación o industrial. En cuanto a la sección R, que tiene el menor número de empresas con 770, esta se dedica a agrupar empresas relacionadas con el sector inmobiliario como se puede observar en su nube de palabras ya que encontramos términos como propiedad, inmobiliario, alquiler compra o venta.

Gráfico 28: Nube de Palabras para la sección I.

Fuente: Elaboración propia.

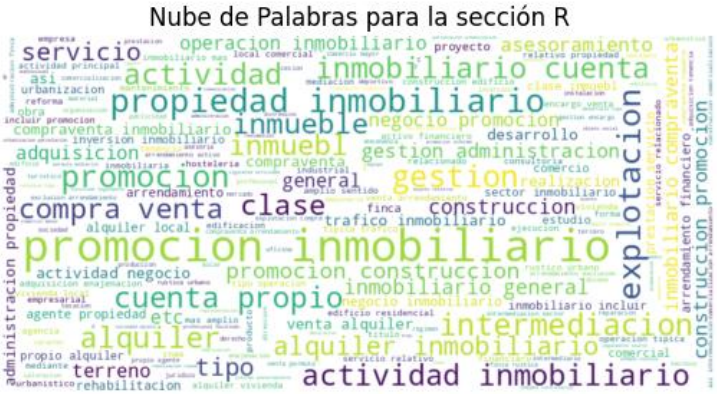

Gráfico 29: Nube de Palabras para la sección R.

Fuente: Elaboración propia.

Después de clasificar las empresas en las secciones determinadas por el modelo NMF, se ha aplicado el mismo método que se usó previamente cuando las secciones se asignaban según la clasificación NACE o según las otras técnicas de modelado de tópicos. Para cada sección, se generaron vectores TF-IDF a partir de las descripciones de las actividades de las empresas. Además, se estableció una función para calcular la entropía de estos vectores TF-IDF, y esta función se utilizó para analizar cada una de las secciones asignadas por el modelo NMF.

En base a los resultados, la gráfica muestra que el sector B, que incluye empresas del sector de la construcción y el inmobiliario como se observa en su nube de palabras, registra la mayor entropía, alcanzando un valor de 13707,04. Asimismo, la tabla de frecuencias muestra que esta sección también se encuentra entre las dos primeras con mayor cantidad de empresas, con un total de 3575. El sector D se sitúa en segundo lugar en cuanto a entropía, con un valor de 13482,62 y en tercer lugar en cuanto a número de empresas, sumando 3563. En contraste, las secciones A y R tienen la entropía más baja, con valores de 2562,03 y 2835,02 respectivamente, y forman parte de las secciones con menor número de empresas agrupadas.

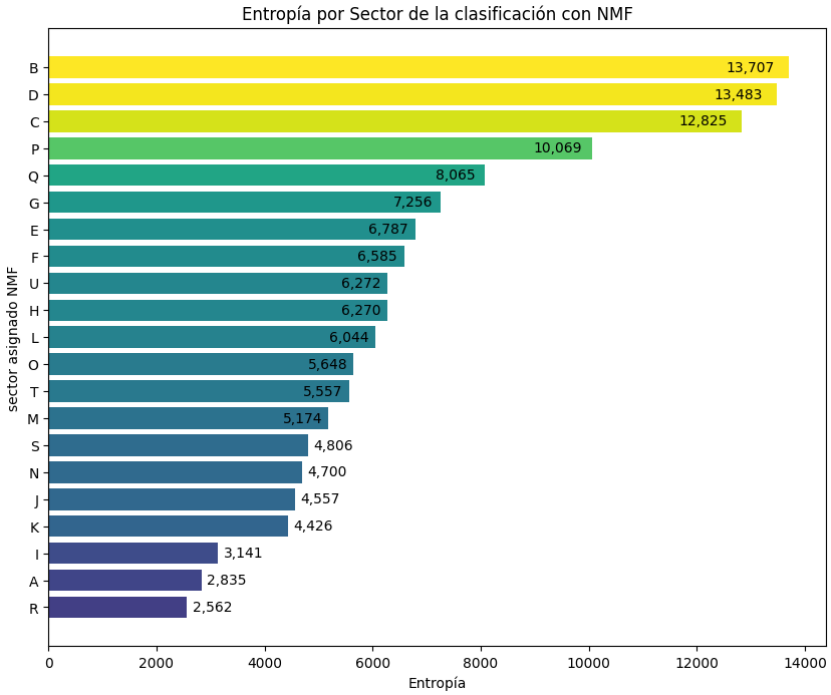

Gráfico 30: Entropía por sección de la clasificación NMF.

Fuente: Elaboración propia.

Sección asignada NMF	Número de Empresas	Entropía
R	770	2562,0313
I	791	3141,0535
M	1002	5173,6904
S	1044	4806,4428
N	1063	4700,3834
J	1085	4557,2790
A	1267	2835,0231
L	1393	6043,8374
K	1524	4426,4041
T	1545	5556,8372
U	1678	6272,1027
O	1678	5647,7430
F	1877	6584,7951
H	1898	6269,6379
E	1909	6787,2633
G	2237	7255,7532
Q	2266	8065,0126
C	3193	12824,8512
D	3563	13482,6254
B	3575	13707,0426
P	3579	10068,7681

Gráfico 31: Tabla de frecuencias con el número de empresas y entropía por sección NMF.

Fuente: Elaboración propia.

También se ha obtenido el gráfico de dispersión 32 que muestra la correlación entre la entropía y el número de empresas por sección. En el eje horizontal, la entropía se mide desde 2500 hasta aproximadamente 14000. En el eje vertical, el número de empresas varía de 700 a cerca de 4,000. Los puntos en el gráfico parecen formar dos grupos distintos: uno en el rango inferior de entropía con un número relativamente más bajo de empresas, y otro grupo en el rango superior de entropía con un número más alto de empresas. Este comportamiento sugiere que, a mayor entropía, que implica una mayor variedad de actividad de las empresas de una sección, corresponde también un mayor número de empresas agrupadas en dicha sección. La disposición de los puntos muestra una relación positiva, indicando que el incremento en la entropía está asociado a un aumento en la cantidad de empresas que forman la sección.

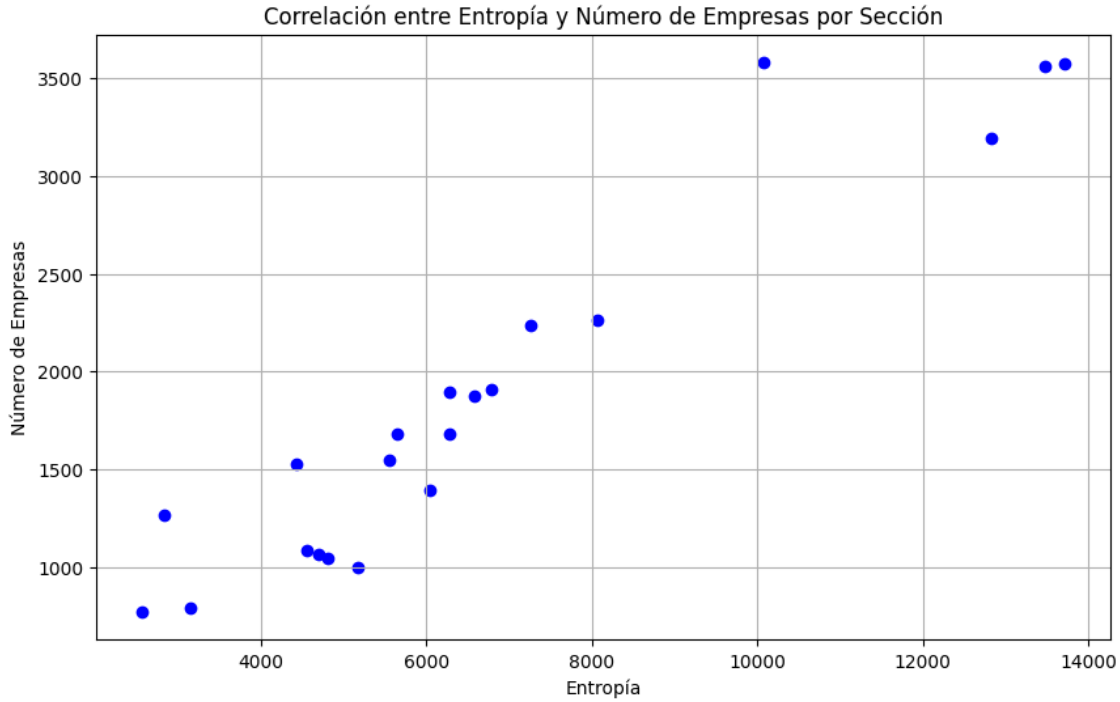

Gráfico 32: Correlación entre Entropía y Número de Empresas por Sección NMF.

Fuente: Elaboración propia.

Finalmente, se ha calculado la entropía media para la clasificación con NMF que es de 6703,26, la entropía mediana que es 6043,83, la entropía mínima que es 2562,03, la entropía máxima que es 13707,04 y el coeficiente de Gini para la clasificación con NMF es: 0.25.

4.2.3. Comparación de resultados

En este apartado se comparan los resultados de las nuevas clasificaciones con la clasificación original de la NACE.

En primer lugar, la clasificación original por NACE mostró una correlación significativa entre el número de empresas y la entropía calculada, sugiriendo que las secciones con más empresas agrupadas tienden a ser más variadas. Los sectores de comercio al por mayor y menor, así como la construcción, destacaron por su alta entropía, indicando una amplia gama de subactividades dentro de estas categorías.

Al aplicar LDA, se observó que la distribución de empresas por sección era menos uniforme en comparación con NACE. Utilizando BERTopic, la entropía alcanzó los valores más altos en sectores relacionados con el comercio y la construcción. Esto sugiere que BERTopic puede ser eficaz para distinguir pequeñas variaciones en descripciones de empresas que operan en sectores económicos complejos y diversificados. Finalmente, la clasificación mediante NMF exhibió un comportamiento similar a LDA en términos de entropía, pero con una distribución ligeramente más equilibrada de empresas por sección.

En todos los métodos, existe una correlación general entre el número de empresas en una sección y la entropía de esta, indicando que, a mayor número de empresas en una sección, generalmente existe una mayor diversidad de actividades descritas. Esta correlación fue más pronunciada en la clasificación BERTopic, posiblemente debido a su capacidad para capturar mejor la semántica de las descripciones de actividades.

Por otro lado, si comparamos la entropía media obtenida de cada una de las clasificaciones observamos que es idéntica para todas las técnicas, situada en 6703,26. Esto sugiere que, en promedio, cada técnica clasifica las descripciones de la actividad empresarial de una manera que resulta en un nivel similar de diversidad o variabilidad de información entre los grupos.

En cuanto a las entropías medianas y máximas, la clasificación NACE tiene una entropía mediana relativamente alta y la entropía máxima es la más alta entre la obtenida de todos los modelos, lo que indica que, aunque algunos sectores clasificados están altamente diversificados, muchos están cerca de la mediana, reflejando una distribución más homogénea. La clasificación obtenida mediante LDA presenta una entropía máxima menor que NACE y una entropía mediana menor, indicando una menor variabilidad en la diversidad entre los sectores analizados. Esto podría reflejar una concentración más homogénea de empresas en sectores específicos.

Por otro lado, la clasificación con BERTopic obtiene la entropía máxima más alta, mucho mayor que la de los otros modelos, mientras que su mediana es significativamente más baja. Esto sugiere que algunos sectores son extremadamente diversificados, mientras que muchos otros podrían estar bastante especializados o concentrados. Finalmente, la clasificación realizada con NMF tiene la menor entropía máxima, y una entropía mediana cercana a la de LDA, lo que indica una menor dispersión general en la clasificación de empresas en términos de diversidad.

Además, el coeficiente de Gini calculado para las entropías de cada modelo refuerza estos hallazgos. NACE presenta un coeficiente de Gini de 0,61, indicando una distribución relativamente equitativa de la diversidad dentro de las secciones. LDA y NMF muestran coeficientes de Gini de 0,30 y 0,25 respectivamente, lo que sugiere una mayor homogeneidad en la clasificación de empresas en comparación con NACE. BERTopic, sin embargo, exhibe el coeficiente de Gini más alto, con un valor de 0,80, lo que confirma que este modelo tiende a crear secciones con una gran disparidad en la diversidad interna: algunas muy especializadas y otras extremadamente diversas.

Los resultados obtenidos sugieren que BERTopic podría ser útil para identificar una variedad de subtemas o especializaciones dentro de un sector más grande mientras que NMF y LDA pueden ser preferibles para estudios que requieran una mayor

uniformidad en la clasificación y análisis, particularmente cuando se buscan patrones más generales o tendencias en sectores sin la necesidad de profundizar en subclases. Aunque, la clasificación NACE a pesar de ser un sistema estandarizado, sigue siendo útil para la clasificación de empresas en sectores.

5. Conclusiones

Para finalizar este proyecto, en este capítulo se presentan las conclusiones obtenidas. La problemática que motivó este estudio fue el mal funcionamiento en ciertas ocasiones por parte de la NACE a la hora de clasificar en sectores las diferentes empresas. Para tratar de solventar o disminuir el error de esta clasificación, se han explorado diversos métodos de procesamiento de lenguaje natural en los que no es la propia empresa la que elige el código en el que se clasifica, sino que éste se asigna de forma automática en función de su similitud con la descripción de la actividad de otras empresas. Se han implementado diversas técnicas de Topic Modelling para clasificar empresas en secciones según su descripción de actividad, como Latent Dirichlet Allocation, Non-negative Matrix Factorization o BERTopic.

Una vez implementados estos modelos, se han comparado los resultados utilizando la entropía como medida de incertidumbre. En cuanto a los resultados obtenidos, se puede llegar a la conclusión de que para el estudio realizado los modelos de LDA y NMF son los que mejor han clasificado las empresas en sectores, aunque, si en un futuro se quisiera seguir clasificando las empresas en subcategorías (divisiones, grupos y clases) lo más adecuado sería utilizar BERTopic. Este modelo ha demostrado ser el más adecuado para identificar una amplia variedad de subtemas o especializaciones dentro de un sector más grande, mientras que LDA y NMF han sido los que han distribuido de una forma más equitativa las empresas en los 21 sectores requeridos.

En cuanto a los objetivos propuestos al inicio del proyecto, se ha cumplido con el objetivo principal de desarrollar un sistema capaz de interpretar y clasificar automáticamente las descripciones de actividades empresariales en categorías. Para lograr cumplir con este objetivo, se ha realizado un análisis profundo de la estructura y el funcionamiento de la NACE, desde su evolución lo largo de los años, pasando por su codificación y normativa a la hora de clasificar actividades hasta los problemas ligados a este sistema de clasificación. Una vez establecido dicho contexto se han empleado diversas técnicas de limpieza y preprocesamiento de datos textuales y finalmente se han desarrollado los modelos que se han considerado más adecuados.

Una vez implementados los modelos y comparados los resultados se ha valorado que sería útil el empleo de las técnicas de topic modelling a la hora de clasificar las empresas en sectores. Es decir, además de ser la propia empresa la que selecciona su categoría, valorar su elección a través de su descripción de actividad empleando los modelos desarrollados en este estudio, de esta forma, se podría reducir considerablemente la posibilidad de error a la hora de clasificar las empresas. Además, al disponer de clasificaciones más fiables, se habilitaría la monitorización sectorial de la actividad económica de forma más precisa, detectando mejor caídas o incrementos sectoriales de actividad económica.

En cuanto a limitaciones de este estudio, la muestra empleada para la implementación y prueba de los modelos ha sido muy reducida en comparación con la cantidad de datos que puede proporcionar SABI. Con un ordenador más potente se podría haber seleccionado una muestra más grande y de esta forma la clasificación en sectores podría haber quedado más equiparada, es decir, la muestra representaría mejor a todos los sectores en los que clasifica la NACE. Por otro lado, habría sido interesante haber

empleado alguna técnica de hierarchical topic modelling para comparar los resultados con los obtenidos mediante las técnicas de topic modelling. Además, el estudio podría haber seguido clasificando las empresas en divisiones, grupos y clases empleando las mismas técnicas de clasificación y llegando resultados más concretos y precisos.

En cuanto a los conocimientos adquiridos a lo largo del grado que han sido empleados en este trabajo, principalmente surgen de la asignatura "Lenguaje Natural y recuperación de la información". En esta asignatura se aprende a extraer información de textos para posteriormente procesarlos y analizarlos con técnicas de procesamiento de lenguaje natural como puede ser la tokenización, lematización o vectorización, todas ellas técnicas aplicadas en este proyecto. Por otro lado, a la hora de implementar los modelos escogidos se emplearon los conocimientos adquiridos en asignaturas como "Modelos descriptivos y predictivos" o "Evaluación, despliegue y monitorización de modelos". Además, para generar el contexto del estudio y adquirir la base de datos, se han empleado conocimientos adquiridos en "Comportamiento económico y social". También han sido muy importantes en el desarrollo de este proyecto los conocimientos adquiridos en "Programación" para la realización del código implementado en Python o los conocimientos sobre la Entropía como medida de incertidumbre adquiridos en la asignatura "Teoría de la Información". Además, los conocimientos adquiridos en la asignatura de "Visualización" fueron también fundamentales para realizar los gráficos adecuados para la posterior interpretación.

Finalmente, para la realización de este proyecto, además del conocimiento obtenido en las asignaturas del grado, se han aplicado diversas competencias adquiridas durante estos años, como, por ejemplo, la "Planificación y gestión del tiempo", el "Aprendizaje permanente", la "Aplicación y pensamiento crítico" y el "Análisis y resolución de problemas".

En cuanto a futuros proyectos, como se ha comentado anteriormente sería interesante continuar desarrollando la clasificación de empresas a nivel de división, grupo y clase y de esta manera determinar que modelo es el que queda más acertado en la clasificación. Por otro lado, también sería conveniente tratar de combinar la selección manual de grupo con la clasificación realizada a través de los modelos con tal de perfeccionar el sistema ya existente. Además, se podrían probar técnicas de hierarchical topic modelling para comprobar si la clasificación con este tipo de modelos es más efectiva.

En cuanto al legado principal de este proyecto queda la propuesta de clasificación de empresas según su similitud semántica, que puede ser de utilidad para evaluar la actividad económica con clasificaciones sectoriales más homogéneas. Además, queda a disposición de todos los interesados en continuar con el estudio toda la información y datos necesarios a continuación: https://github.com/irenecanovaas/TFG_IRENE_CANOVAS.git. En el repositorio GitHub adjunto se encuentra la base de datos empleada en el proyecto y todos los scripts relacionados con la limpieza y preprocesamiento de los datos textuales además de los modelos implementados.

Bibliografía

Amazon Web Services. (2024). *LDA Algorithm*. Obtenido de https://docs.aws.amazon.com/es_es/sagemaker/latest/dg/lda.html

Analytics Vidhya. (junio de 2021). *Step-by-step guide to master NLP: Topic modelling using NMF*. Obtenido de https://www.analyticsvidhya.com/blog/2021/06/part-15-step-by-step-guide-to-master-nlp-topic-modelling-using-nmf/

Bastani, K., Namavari, H., & Shaffer, J. (2019). Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints. *Expert Systems with Applications*. Obtenido de https://doi.org/10.1016/j.eswa.2019.03.001

Bishop, A., Mateos-Garcia, J., & Richardson, G. (2022). Using text data to improve industrial statistics in the UK. *Economic Statistics Centre of Excellence*.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet Allocation. . *Journal of Machine Learning Research 3*.

Brenes González, H. A. (2020). La curva de Lorenz y el coeficiente de Gini como medidas de la desigualdad de los ingresos. *Revista Electrónica de Investigación en Ciencias Económicas*.

Colegio de Registradores. (2024). *Colegio de Registradores de España*. Obtenido de https://www.registradores.org/el-colegio/registro-mercantil

Comisión Europea. (2020). *Eurostat: Statistical classification of economic activities in the European Community*. Obtenido de https://ec.europa.eu/eurostat/statistics-explained/SEPDF/cache/1320.pdf

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory*. Wiley-Interscience.

D&B, I. (2024). *eInforma*. Obtenido de https://www.einforma.com/ayuda/soluciones-y-herramientas/que-es-sabi

Dabrowski, D., Lasslop, G., Munter, P., Cierpial-Wolan, M., van Delden, A., & Phelps, S. (2022). NACE codes from text: WP3 1st Interim technical report.

DataCamp. (2023). *What is topic modeling?* Obtenido de https://www.datacamp.com/tutorial/what-is-topic-modeling

Devopedia. (2020). *Latent Dirichlet Allocation*. Obtenido de https://devopedia.org/latent-dirichlet-allocation

Egger, R., & Yu, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*.

eInforma. (2024). *SABI: Sistema de Análisis de Balances Ibéricos*. Obtenido de ¿Qué es SABI?: https://www.einforma.com/ayuda/soluciones-y-herramientas/que-es-sabi

Eurostat. (2008). *NACE Rev. 2: Statistical classification of economic activities in the European Community*. Luxembourg: Office for Official Publications of the European Communities.

Fuentes de Información Entropía. (2021). *Departamento de sistemas Informáticos y comunicación*.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv*.

HackerNoon. (10 de abril de 2023). *NLP Tutorial: Topic Modeling in Python with BERTopic*. . Obtenido de https://hackernoon.com/es/nlp-tutorial-tema-modelado-en-python-con-bertopic-372w35l9

Hernández Fernández, G. (2020). Integración de Data Mining sobre noticias para predicción en mercados financieros. *Trabajo de fin de grado, Universidad Politécnica de Madrid*.

Hutama , L., & Suhartono, D. (2022). Indonesian hoax news classification with multilingual transformer model and BERTopic. *Informatica*.

INE. (2006). *NATIONAL CLASSIFICATION OF ECONOMIC ACTIVITIES (CNAE)*. Madrid.

Kherwa, P., & Bansal, P. (2019). Topic modeling: A comprehensive review. *EAI Endorsed Transactions on Scalable Information Systems*. Obtenido de https://levity.ai/blog/what-is-topic-modeling

Lee, D., & Seung, S. (1999). Learning the parts of objects by non-negative matrix factorization. *Letters to nature*.

Li, W., & McCallum, A. (2006). Pachinko Allocation: DAG-Structured Mixture Models of Topic Correlations. *Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA*.

Liu, L., Tang, L., He, L., Zhou, W., & Shaowen, Y. (2016). An Overview of Hierarchical Topic Modeling. *8th International Conference on Intelligent Human-Machine Systems and Cybernetics*.

Mateos-García, J., & Richardson, G. (2022). A bottom up industrial taxonomy for the UK: Refinements and an application. *Economic Statistics Centre of Excellence*.

Motahhir, S., & Bossoufi, B. (2023). *Digital Technologies an Aplications*. Fez: Springer Nature.

Núñez Martínez, J. (2005). Estudio de nuevos algoritmos de descomposición lineal de observaciones en componenetes. *Universidad de Sevilla*.

Risch, J. (2016). Detecting Twitter topics using latent Dirichlet allocation. *Uppsala Universitet, Department of Information Technology*.

Rodríguez Vega, A. I. (s.f.). Factorizacón No Negativa de Matrices. *Trabajo de Fin de Grado de la Universidad de Valladolid*.

SABI. (2024). *SABI Infroma*. Obtenido de https://sabi.informa.es/version-20230626-8-0/home.serv?product=SabiInforma&

Samsir, Surbakti Saragih, R., Subagio, S., Aditya, R., & Watrianthos, R. (2023). BERTopic Modeling of Natural Language Processing Abstracts: Thematic Structure and Trajectory. *Jurnal Media Informatika Budidarma*.

Sarica, S., & Luo, J. (2021). Stopwords in technical language processing. *PLOS ONE*. Obtenido de https://doi.org/10.1371/journal.pone.0254937

Shannon, C. (1948). *A Mathematical Theory of Communication*.

Shi, T., Kang, K., Choo, J., & Reddy, C. (2018). Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. *World wide web conference*.

Smith, A., Hawes, T., & Myers, M. (2014). Hierarchie: Interactive Visualization for Hierarchical Topic Models. *DECISIVE ANALYTICS Corporation*.

Smith, A., Hawes, T., & Myers, M. (2014). Hiérarchie: Interactive Visualization for Hierarchical Topic Models. *Decisive Analytics Corporation*.

Spacy Developers. (2024). *spaCy 101: Everything you need to know*. Obtenido de https://spacy.io/usage/spacy-101

Uncovska, M., Freitag, B., Meister, S., & Fehring, L. (2023). Rating analysis and BERTopic modeling of consumer versus regulated mHealth app reviews in Germany. *NPJ Digital Medicine*.

UOC, B. d. (2024). *SABI: Sistema de Análisis de Balances Ibéricos*. Obtenido de https://biblioteca.uoc.edu/es/Coleccion-digital-por-areas-de-estudio/coleccion/SABI-Sistema-de-Analisis-de-Balances-Ibericos/

Vasiliev, Y. (2020). *Natural language processing with Python and spaCy: A practical introduction*.

Vila Rodríguez, K., Fernández Orquín, A., Collazo Amable, A., Pérez Martín, R., Cobarrubia Echarte, Á. L., & Cano Morera, D. (2009). Sistema para el pre-procesamiento de textos para el Procesamiento del Lenguaje Natural.

Zhai, C., & Massung, S. (2016). *Text data management and analysis : a practical introduction to information*.

6. Anexo I. Objetivos de Desarrollo Sostenible

Objetivos de Desarrollo Sostenibles	Alto	Medio	Bajo	No Procede
ODS 1. **Fin de la pobreza**				x
ODS 2. **Hambre cero**				x
ODS 3. **Salud y bienestar**				x
ODS 4. **Educación de calidad**				x
ODS 5. **Igualdad de género**				x
ODS 6. **Agua limpia y saneamiento**				x
ODS 7. **Energía asequible y no contaminante**				x
ODS 8. **Trabajo decente y crecimiento económico**	x			
ODS 9. **Industria, innovación e infraestructuras**	x			
ODS 10. **Reducción de las desigualdades**				x
ODS 11. **Ciudades y comunidades sostenibles**				x
ODS 12. **Producción y consumo responsables**		x		
ODS 13. **Acción por el clima**				x
ODS 14. **Vida submarina**				x
ODS 15. **Vida de ecosistemas terrestres**				x
ODS 16. **Paz, justicia e instituciones sólidas**				x
ODS 17. **Alianzas para lograr objetivos**		x		

Figura 14: Grado de relación del trabajo con los Objetivos de Desarrollo Sostenible

Fuente: Elaboración propia.

El objetivo principal de este proyecto ha sido la mejora en la clasificación de empresas según su actividad económica mediante el uso de técnicas de procesamiento de lenguaje natural y aprendizaje automático. Este objetivo puede relacionarse con los siguientes Objetivos de Desarrollo Sostenible (ODS):

- **ODS 8 (Trabajo decente y crecimiento económico):** Este objetivo promueve el crecimiento económico sostenido, inclusivo y sostenible además del empleo pleno, productivo y decente. La correcta clasificación de actividades económicas contribuye a una mejor comprensión y análisis de los sectores económicos, lo que puede influir positivamente en la elaboración de estrategias

de desarrollo económico. Una clasificación más exacta permite identificar con claridad las áreas de crecimiento y los sectores que requieren intervención.

- **ODS 9 (Industria, innovación e infraestructura):** Este objetivo busca construir infraestructuras resilientes, promover la industrialización inclusiva y sostenible, y fomentar la innovación. Este TFG contribuye a este objetivo al innovar en el método de clasificación de empresas. Al automatizar el proceso de clasificación, se facilita la gestión y el análisis de grandes conjuntos de datos empresariales, apoyando la toma de decisiones basada en información más precisa y actualizada.

Además de los ODS 8 y 9, el proyecto también está relacionado menos directamente con los objetivos 12 y 17:

- **ODS 12 (Producción y consumo responsables):** Una clasificación empresarial más exacta puede ayudar a identificar prácticas de producción y patrones de consumo, contribuyendo a estrategias para un consumo más sostenible.
- **ODS 17 (Alianzas para lograr los objetivos):** La mejora en la clasificación y el análisis de datos empresariales puede fomentar la colaboración entre entidades económicas y académicas, promoviendo la compartición de datos y estrategias para el desarrollo sostenible.

7. Anexo II. Nubes de palabras

En este anexo se muestran las nubes de palabras de cada una de las secciones obtenidas tanto utilizando el método NACE como los métodos de Topic Modelling implementados en el desarrollo de este proyecto.

7.1. Nubes de palabras para las secciones NACE

Gráfico 33: Nube de palabras para la sección B NACE

Fuente: Elaboración propia

Gráfico 34: Nube de palabras para la sección D NACE

Fuente: Elaboración propia

Gráfico 35: Nube de palabras para la sección E NACE

Fuente: Elaboración propia

Gráfico 36: Nube de palabras para la sección F NACE

Fuente: Elaboración propia

Gráfico 37: Nube de palabras para la sección G NACE

Fuente: Elaboración propia

Gráfico 38: Nube de palabras para la sección H NACE

Fuente: Elaboración propia

Gráfico 39: Nube de palabras para la sección J NACE

Fuente: Elaboración propia

Gráfico 40: Nube de palabras para la sección K NACE

Fuente: Elaboración propia

Gráfico 41: Nube de palabras para la sección L NACE

Fuente: Elaboración propia

Gráfico 42: Nube de palabras para la sección M NACE

Fuente: Elaboración propia

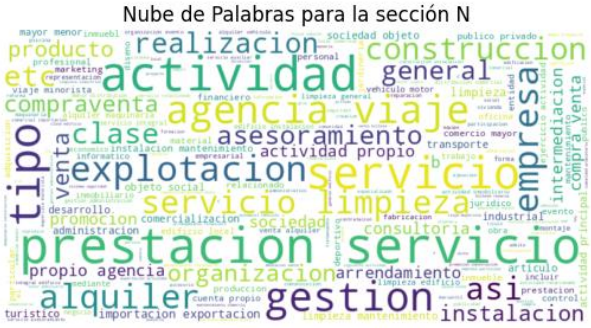

Gráfico 43: Nube de palabras para la sección N NACE

Fuente: Elaboración propia

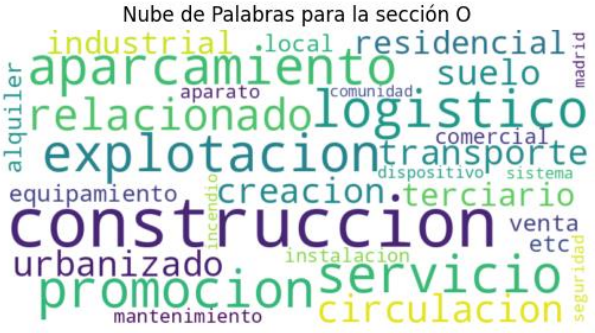

Gráfico 44: Nube de palabras para la sección O NACE

Fuente: Elaboración propia

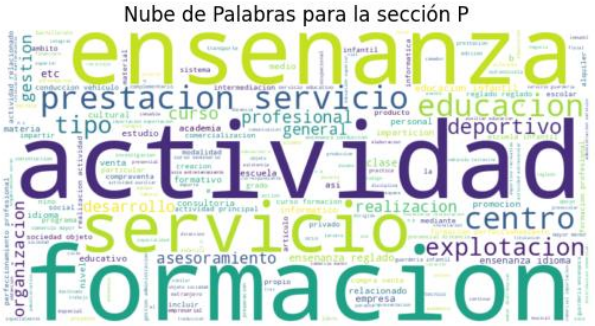

Gráfico 45: Nube de palabras para la sección P NACE

Fuente: Elaboración propia

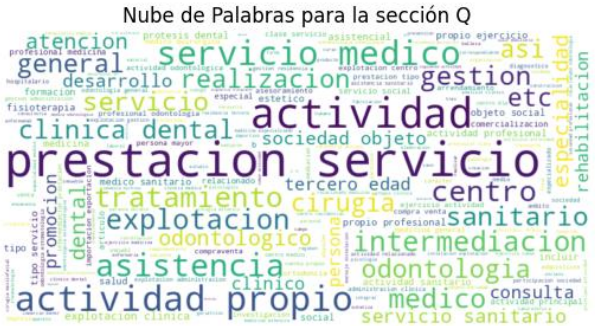

Gráfico 46: Nube de palabras para la sección Q NACE

Fuente: Elaboración propia

Gráfico 47: Nube de palabras para la sección R NACE

Fuente: Elaboración propia

Gráfico 48: Nube de palabras para la sección S NACE

Fuente: Elaboración propia

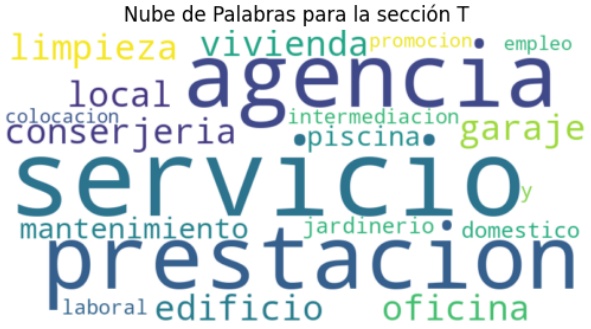

Gráfico 49: Nube de palabras para la sección T NACE

Fuente: Elaboración propia

Gráfico 50: Nube de palabras para la sección U

Fuente: Elaboración propia

7.2. Nubes de palabras para las secciones LDA

Gráfico 51: Nube de palabras para la sección A LDA

Fuente: Elaboración propia

Gráfico 52: Nube de palabras para la sección B LDA

Fuente: Elaboración propia

Gráfico 53: Nube de palabras para la sección C LDA

Fuente: Elaboración propia

Gráfico 54: Nube de palabras para la sección D LDA

Fuente: Elaboración propia

Gráfico 55: Nube de palabras para la sección E LDA

Fuente: Elaboración propia

Gráfico 56: Nube de palabras para la sección F LDA

Fuente: Elaboración propia

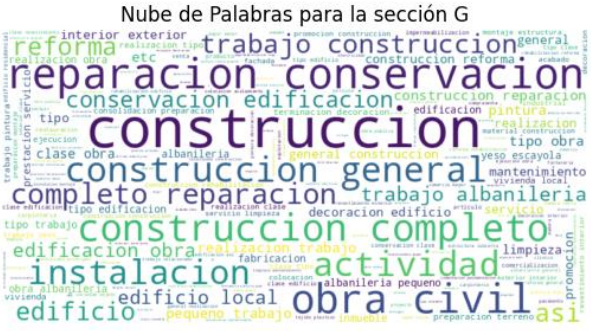

Gráfico 57: Nube de palabras para la sección G LDA

Fuente: Elaboración propia

Gráfico 58: Nube de palabras para la sección I LDA

Fuente: Elaboración propia

Gráfico 59: Nube de palabras para la sección J LDA

Fuente: Elaboración propia

Gráfico 60: Nube de palabras para la sección L LDA

Fuente: Elaboración propia

Gráfico 61: Nube de palabras para la sección M LDA

Fuente: Elaboración propia

Gráfico 62: Nube de palabras para la sección O LDA

Fuente: Elaboración propia

Gráfico 63: Nube de palabras para la sección P LDA

Fuente: Elaboración propia

Gráfico 64: Nube de palabras para la sección Q LDA

Fuente: Elaboración propia

Gráfico 65: Nube de palabras para la sección S LDA

Fuente: Elaboración propia

Gráfico 66: Nube de palabras para la sección T LDA

Fuente: Elaboración propia

Gráfico 67: Nube de palabras para la sección U LDA

Fuente: Elaboración propia

7.3. Nubes de palabras para las secciones BERTopic

Gráfico 68: Nube de palabras para la sección A BERTopic

Fuente: Elaboración propia

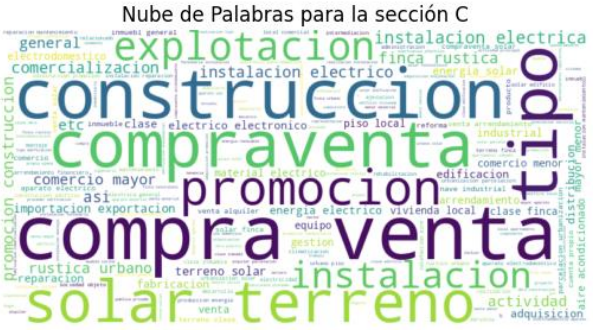

Gráfico 69: Nube de palabras para la sección C BERTopic

Fuente: Elaboración propia

Gráfico 70: Nube de palabras para la sección E BERTopic

Fuente: Elaboración propia

Gráfico 71: Nube de palabras para la sección F BERTopic

Fuente: Elaboración propia

Gráfico 72: Nube de palabras para la sección G BERTopic

Fuente: Elaboración propia

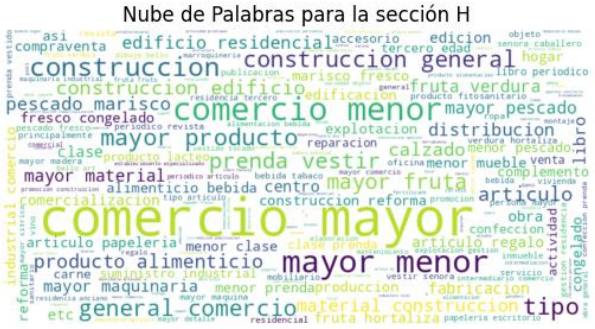

Gráfico 73: Nube de palabras para la sección H BERTopic

Fuente: Elaboración propia

Gráfico 74: Nube de palabras para la sección J BERTopic

Fuente: Elaboración propia

Gráfico 75: Nube de palabras para la sección K BERTopic

Fuente: Elaboración propia

Gráfico 76: Nube de palabras para la sección L BERTopic

Fuente: Elaboración propia

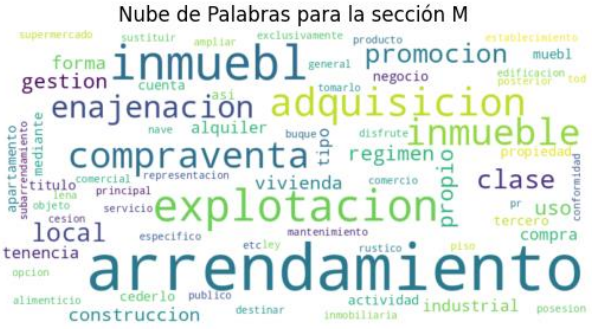

Gráfico 77: Nube de palabras para la sección M BERTopic

Fuente: Elaboración propia

Gráfico 78: : Nube de palabras para la sección N BERTopic

Fuente: Elaboración propia

Gráfico 79: Nube de palabras para la sección O BERTopic

Fuente: Elaboración propia

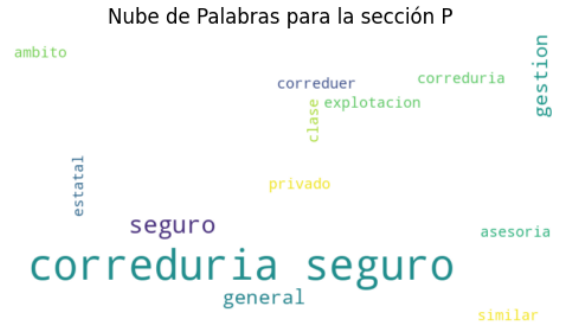

Gráfico 80: Nube de palabras para la sección P BERTopic

Fuente: Elaboración propia

Gráfico 81: Nube de palabras para la sección Q BERTopic

Fuente: Elaboración propia

Gráfico 82: Nube de palabras para la sección R BERTopic

Fuente: Elaboración propia

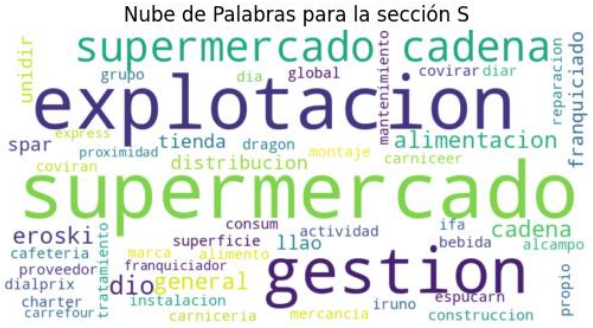

Gráfico 83: Nube de palabras para la sección S BERTopic

Fuente: Elaboración propia

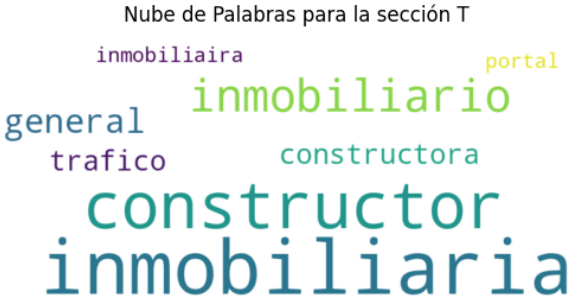

Gráfico 84: Nube de palabras para la sección T BERTopic

Fuente: Elaboración propia

7.4. Nubes de palabras para las secciones NMF

Gráfico 85: Nube de palabras para la sección A NMF

Fuente: Elaboración propia

Gráfico 86: Nube de palabras para la sección C NMF

Fuente: Elaboración propia

Gráfico 87: Nube de palabras para la sección D NMF

Fuente: Elaboración propia

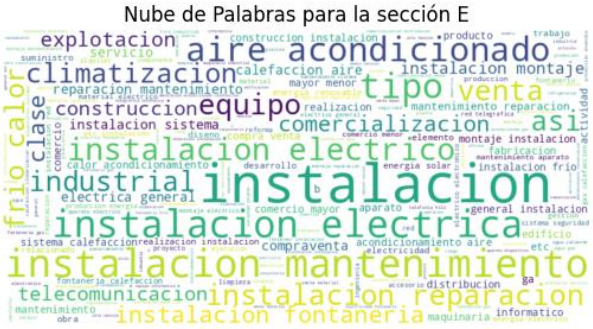

Gráfico 88: Nube de palabras para la sección E NMF

Fuente: Elaboración propia

Gráfico 89: Nube de palabras para la sección F NMF

Fuente: Elaboración propia

Gráfico 90: Nube de palabras para la sección G NMF

Fuente: Elaboración propia

Gráfico 91: Nube de palabras para la sección H NMF

Fuente: Elaboración propia

Gráfico 92: Nube de palabras para la sección I NMF

Fuente: Elaboración propia

Gráfico 93: Nube de palabras para la sección J NMF

Fuente: Elaboración propia

Gráfico 94: Nube de palabras para la sección K NMF

Fuente: Elaboración propia

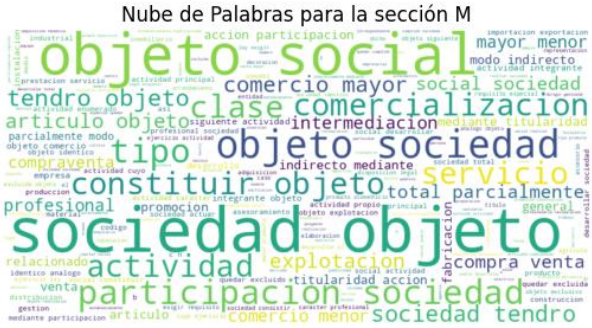

Gráfico 95: Nube de palabras para la sección M NMF

Fuente: Elaboración propia

Gráfico 96: Nube de palabras para la sección N NMF

Fuente: Elaboración propia

Gráfico 97: Nube de palabras para la sección O NMF

Fuente: Elaboración propia

Gráfico 98: Nube de palabras para la sección Q NMF

Fuente: Elaboración propia

Gráfico 99: Nube de palabras para la sección S NMF

Fuente: Elaboración propia

Gráfico 100: Nube de palabras para la sección T NMF

Fuente: Elaboración propia

Gráfico 101: Nube de palabras para la sección U NMF

Fuente: Elaboración propia