COMILLAS
UNIVERSIDAD PONTIFICIA

Facultad de Ciencias Económicas y Empresariales

APROXIMACIÓN A LA INFORMACIÓN NO FINANCIERA DEL IBEX 35 MEDIANTE *TEXT MINING* Y *TOPIC MODELING*

Autor: Paula de la Torre Solera
Tutor: Lourdes Fernández Rodríguez

MADRID | Junio 2023

Resumen

Este trabajo de fin de grado busca examinar la información no financiera del IBEX 35 correspondiente al ejercicio 2021 utilizando técnicas de *text mining*, al ser la más actualizada al realizar el estudio. El objetivo es identificar los temas prioritarios en la agenda ESG del IBEX 35 mediante *topic modeling*. Para llevar a cabo este estudio, se realizará un análisis exhaustivo de los términos y bigramas encontrados en los EINFs del IBEX 35, además de utilizar un modelo LDA que ayudará a determinar las temáticas subyacentes en el *corpus* generado. Los resultados obtenidos resaltan la necesidad de establecer un formato homogéneo para la publicación de información en sostenibilidad, así como la consolidación de la sostenibilidad social como prioritaria. Además, se observa la integración de la sostenibilidad en el modelo de negocio, otorgándole la misma importancia que a la operatividad de la empresa.

Palabras Clave

Sostenibilidad, ESG, Información No Financiera, IBEX 35, *Text Mining*, *Topic Modeling*

Abstract

This final degree project aims to examine the non-financial information of the IBEX 35 for the year 2021 using text mining techniques, being the most updated at the time of the study. The objective is to identify the priority topics within the ESG agenda of the IBEX 35 using topic modeling. To carry out this study, an exhaustive analysis of the terms and bigrams found in non-financial reports of the IBEX 35 will be carried out, in addition to using an LDA model that will help determine the underlying topics in the generated corpus. The results obtained highlight the need to establish a homogeneous format for the publication of information on sustainability, as well as the consolidation of social sustainability as a priority topic. In addition, the integration of sustainability in the business model is observed, giving it the same importance as the company's operations.

Key Words

Sustainability, ESG, Non-Financial Information, IBEX 35, Text Mining, Topic Modeling

Índice

Índice de Figuras

Índice de Tablas

1. Introducción

1.1. Justificación del tema

Debido a los retos ambientales y sociales, la limitación de recursos y la creciente sensibilización sobre los desafíos que afronta la sociedad, la sostenibilidad y su integración en las operaciones empresariales se han convertido en un tema central en la agenda de cualquier industria. En España, se ha presenciado una tendencia empresarial emergente en la que la sostenibilidad ha ganado cada vez más relevancia llevando a las empresas a ampliar su compromiso y acciones en materia ambientales, sociales y de buen gobierno (ESG) (Transcendent, 2022).

Todo esto se materializa en las memorias de sostenibilidad publicadas anualmente por las empresas, equiparable al informe de gestión en materia operativa. Actualmente, la divulgación de información no financiera (INF) viene marcada por la Directiva 2014/95/UE y su trasposición al Derecho Nacional en la Ley 11/2018. Se establece un marco de divulgación en el cual las empresas afectadas están obligadas a publicar informes sobre aspectos como la gestión ambiental, la gestión social y la diversidad, o las cuestiones de buen gobierno (Deloitte, 2017). Sirve para comunicar más allá de los resultados financieros y centrarse en el impacto social, ambiental y ético de las empresas en la sociedad y el planeta.

Aun con un marco regulatorio definido, no existen estándares de obligatorio cumplimiento que homogenicen la INF y faciliten la comparación y análisis de madurez de la integración de la sostenibilidad a escala sectorial o por ámbito geográfico. También, el volumen de información publicada hace inabarcable un análisis manual exhaustivo de este.

Este trabajo de fin de grado se apalanca en el *text mining* para afrontar esta problemática y obtener temas transversales a la INF del IBEX 35. El uso de *text mining* reduce exponencialmente el tiempo dedicado a analizar cualquier tipo de información no estructurada y facilita la obtención de resultados coherentes.

La elección de este tema viene justificada por la gran importancia que tiene la sostenibilidad en el panorama empresarial actual, la constante actualización regulatoria en materia de INF y el interés personal del autor de este trabajo sobre cómo las empresas responden ante estos retos sociales y regulatorios.

1.2. Objetivos

El objetivo de este trabajo de fin de grado es estudiar y determinar cuáles son los temas prioritarios en la agenda ESG (en inglés, *Enviromental*, *Social* y *Governance*) de las empresas del IBEX 35, como representantes del tejido empresarial de nuestro país. Además, se busca establecer si el enfoque hacia la sostenibilidad sigue centrado en la Responsabilidad Social Corporativa (RSC) o ha evolucionado hacia una integración en las operaciones, cadena de valor y órganos de gobierno de las empresas.

Para llevar a cabo este estudio, se analizarán los Estados de Información No Financiera (EINFs) de cada empresa, o informes equivalentes, correspondientes al año 2021, dado que son los últimos publicados disponibles. Estos informes se consideran la fuente de información principal en materia de sostenibilidad que cumple con los requisitos legales establecidos por la legislación vigente.

Toda este estudio se respaldará con los resultados obtenidos mediante un proceso de *text mining* (minería de texto) y *topic modeling* (modelado de temas). Para ello, se utilizará el conjunto de EINFs como *corpus*, es decir, como base de datos objeto de estudio; y se evaluará si el formato de la información no financiera es adecuado o si es necesario evolucionar hacia una presentación más uniforme y comparable, que facilite este proceso.

1.3. Metodología

La metodología de este trabajo se sustenta en dos pilares. El primero es el estudio de la sostenibilidad, la regulación de INF y el *text mining*. El segundo es el análisis de la INF de las empresas del IBEX 35 mediante *text mining*, específicamente utilizando *el topic modeling*.

Se ha estudiado y definido un marco teórico con el objetivo de contextualizar los aspectos más relevantes relacionados con la sostenibilidad, la regulación y el *text mining*. Esta información ha sido obtenida de repositorios académicos como Google Scholar o Research Gate, así como de las páginas oficiales de las instituciones relevantes en el tema abordado.

Por otro lado, se han revisado las técnicas de análisis de datos no estructurados aprendidas durante el grado de E2+Analytics para seleccionar el algoritmo adecuado para resolver el problema planteado, eligiendo el *topic modeling* como técnica a emplear. Esto se fundamenta, por una parte, en la búsqueda de patrones comunes o interesantes en la información seleccionada, y por otra, en la naturaleza y dimensionalidad de estos informes.

En paralelo, se ha creado una base de datos que contiene la INF del IBEX 35 correspondiente al año fiscal 2020-2021, dado que, en el momento de realizar este trabajo, los informes del 2021-2022 no habían sido publicados. Sobre esta, se ha realizado el pre-procesamiento, buscando homogeneizar y limpiar la información recogida, así como la implementación del modelo LDA (Asignación Latente de Dirichlet), obteniendo los resultados sobre los cuales se basan las conclusiones de este trabajo de fin de grado.

1.4. Estructura del trabajo

Para lograr la metodología descrita, este trabajo de fin de grado sigue una estructura deductiva yendo de un entendimiento general a la realización de un análisis específico sobre la base de datos utilizada.

Para ello, se comenzará el trabajo realizando una explicación sobre el origen y evolución de la sostenibilidad, así como su integración actual en las prioridades de las empresas, haciendo hincapié en el concepto de información no financiera. A continuación, se describirá el marco regulatorio europeo y nacional que afecta a las compañías del IBEX 35 en materia de INF, introduciendo conceptos como los estándares internacionales de referencia más utilizados. Posteriormente, se introducirá el *text mining*, junto con las posibles aplicaciones en el ámbito empresarial, así cómo se profundizará en el *topic modeling*, como técnica principal dentro del análisis de este trabajo. Acto seguido, se

analizará la base de datos, implementando técnicas exploratorias de pre-procesamiento del texto, así como se implementará el modelo LDA, para obtener los temas principales en sostenibilidad del IBEX 35 y analizar los resultados. Finalmente, se expondrán las conclusiones obtenidas tras la realización del trabajo, teniendo en cuenta las distintas ideas destacadas a lo largo del mismo, buscando responder a los objetivos establecidos.

2. La sostenibilidad: evolución, dimensiones e integración en las empresas

2.1. Origen y evolución del concepto sostenibilidad

Al hablar de sostenibilidad se hace referencia al desarrollo de los procesos que permiten cubrir las necesidades actuales sin poner en peligro la capacidad de las futuras generaciones de satisfacer las suyas propias (Comisión Mundial sobre el Medio Ambiente y el Desarrollo, 1987). En otras palabras, la sostenibilidad implica el uso responsable y consciente de los recursos naturales, la protección del medio ambiente y la promoción del bienestar social y económico.

El concepto de sostenibilidad moderna nace en la década de 1970 como respuesta a la preocupación por el creciente impacto humano en el medio ambiente y la necesidad de encontrar mejores formas de administrar los recursos naturales. En 1972, el Club de Roma publica, en colaboración con el Grupo de Dinámica de Sistemáticas del Instituto de Tecnología de Massachussets (MIT), su primer informe titulado "Los límites del crecimiento". Este se centra en analizar cuál es el impacto sobre el planeta del aumento de población, industrialización de procesos, consumo alimentario, contaminación y explotación de los recursos naturales así como sus implicaciones en las tendencias de consumo (Meadows et al., 1972). Las conclusiones del informe ponen por primera vez en valor la necesidad de establecer un modelo productivo basado en la sostenibilidad ecológica y social, ya que de no cambiar los patrones de consumo los modelos productivos basados en el crecimiento llevarían a la escasez de recursos.

De forma paralela, en junio de 1972 en Estocolmo, se celebra la Conferencia Mundial de las Naciones Unidas sobre el Medio Humano, siendo esta la primera conferencia con alcance internacional que ponía al medio ambiente como sujeto y tema principal de su celebración. Esta conferencia coloca las cuestiones ambientales en el orden del día internacional mediante el establecimiento de 26 principios alineados con la conservación y desarrollo del medio ambiente en la "Declaración de Estocolmo", así como la creación del "Plan de Acción para el Medio Humano" como marco de actuación con recomendaciones para la acción medioambiental internacional (Organización de las Naciones Unidas, 1973). La conferencia sirvió como inspiración para el posterior

desarrollo de políticas y legislación a nivel institucional en temáticas relacionadas con medio ambiente.

Durante la década de 1980, la Comisión Mundial sobre el Medio Ambiente y el Desarrollo (CMMAD) impulsa el concepto de desarrollo sostenible como el medio para integrar la sostenibilidad como objetivo clave de las instituciones y en el ámbito político. Esta comisión se forma con el objetivo de revisar los problemas medioambientales y de desarrollo, sugerir nuevos modelos de colaboración internacional y elevar la compresión entre instituciones y ciudadanos en materia de sostenibilidad (Comisión Mundial sobre el Medio Ambiente y el Desarrollo, 1987). En 1987, se publica el informe "Nuestro Futuro Común", o Informe de Brundtland, en el cual se introduce el concepto de desarrollo sostenible como "el desarrollo que satisface las necesidades del presente sin comprometer la capacidad de las generaciones futuras para satisfacer sus propias necesidades" (Comisión Mundial sobre el Medio Ambiente y el Desarrollo, 1987). El informe introduce también la necesidad de que el crecimiento económico tiene que estar alineado con el desarrollo sostenible ambiental y social de forma equitativa. Esto lleva a que la sostenibilidad y el desarrollo sostenible tomen un mayor peso en el debate institucional y se empiezan a considerar de forma reiterada en las conversaciones institucionales y políticas.

A partir de la década de los 90, se consolida la institucionalización de la sostenibilidad y el desarrollo sostenible mediante la celebración de varias cumbres internacionales para abordar temas ambientales y sociales críticos, como el cambio climático, la pérdida de biodiversidad y la pobreza, así como el establecimiento de objetivos y políticas a escala global. También, la sostenibilidad empieza integrarse en las grandes corporaciones, contribuyendo al reconocimiento por parte de la sociedad de la necesidad de realizar un cambio en los patrones de consumo y producción.

En 1992 tiene lugar la Cumbre de las Naciones Unidas sobre Medio Ambiente y Desarrollo en Río de Janeiro con el objetivo de establecer mecanismos que faciliten la promoción del desarrollo sostenible, y la reducción y mitigación de los desarrollos medioambientales (Organización de las Naciones Unidas, 1992). Por una parte, se populariza que el desarrollo sostenible es alcanzable por instituciones y por individuos.

Por otra, se establece la integración de la cuestión económica, ambiental y social bajo un mismo y nuevo enfoque de los modelos productivos (Organización de las Naciones Unidas, 1992).

De esta Cumbre surgen cinco documentos relacionados con la promoción del desarrollo sostenible alineado con la sostenibilidad ambiental y social, destacando el "Programa 21". Se definen estrategias a llevar a cabo por parte de los Gobiernos e instituciones con el objetivo de conseguir un desarrollo sostenible holístico (Organización de las Naciones Unidas, 1992). Este documento, sumado con la atención que le da el Informe de Brundtland, facilita la incorporación y consolidación del desarrollo sostenible y la sostenibilidad en el centro del discurso político a escala internacional, pero no llega a fijar objetivos concretos y medibles en materia de desarrollo sostenible.

Cabe mencionar diferentes hitos dentro de la consolidación de la sostenibilidad como son, por una parte, el Protocolo de Kioto (1997) y el Acuerdo de París (2015); así como la definición de los Objetivos de Desarrollo Sostenible (ODS) (2015), sirviendo como marco de medición y cumplimiento de las empresas e instituciones con el cambio climático y el desarrollo sostenible (Convención Marco de las Naciones Unidas sobre el Cambio Climático, 1997; Organización de las Naciones Unidas, 2015).

Tanto el Protocolo de Kioto como el Acuerdo de París establecen compromisos en el ámbito de cooperación internacional en materia de sostenibilidad y cambio climático. Ambos reflejan la creciente conciencia y preocupación mundial por la sostenibilidad ya que establecen mecanismos a favor de un modelo productivo menos dependiente de los gases de efecto invernadero, así como métodos para abordar el cambio climático desde políticas institucionales (Convención Marco de las Naciones Unidas sobre el Cambio Climático 1997; 2015). Aunque el Protocolo de Kioto fue criticado por no tener en cuenta a los países en desarrollo, el Acuerdo de París establece un enfoque más inclusivo y anima a todos los países a participar en la reducción de emisiones.

Cabe mencionar la importancia de la definición de los ODS como parte de la Agenda 2030, ya que suponen el consenso a escala global de un marco de medición y contribución al desarrollo sostenible por parte de cualquiera que se adscriba a ellos (Organización de las Naciones Unidas, 2015).

En 2015, Naciones Unidas (ONU) introduce los 17 objetivos con sus respectivas 169 metas como parte de la Agenda 2030 para facilitar el compromiso de los Estados y organizaciones a la hora de movilizar los recursos necesarios para su implementación y cumplimiento (Ministerios de Derechos Sociales y Agenda 2030, s.f.).

Los ODS abarcan, desde diferentes enfoques, aspectos en materia medioambiental, social y económica, que suelen estar interrelacionados entre sí. Estos recogen "los desafíos globales a los que nos enfrentamos día a día, como la pobreza, la desigualdad, el clima, la degradación ambiental, la prosperidad, la paz y la justicia" (Organización de las Naciones Unidad, 2015).

Por otra parte, marcan el camino y los hitos a realizar para cumplir con la Agenda 2030 y facilitar la integración del desarrollo sostenible en todas las operaciones de los Estados y organizaciones. Es por ello que las empresas no solo los incorporan a sus estrategias, sino que establecen cuales son más prioritarios dentro de su actividad empresarial. De ese modo, se busca maximizar el impacto positivo y mitigar los riesgos en materia ESG por parte de las empresas en sus grupos de interés.

Los ODS están ampliamente aceptados por el mundo empresarial como marco de aproximación a la sostenibilidad de sus operaciones. Un 100% de las empresas del IBEX 35 están comprometidas públicamente con el cumplimiento de los ODS y los vinculan a sus estrategias (Pacto Mundial de la ONU, 2022). No solo supone un compromiso con la sostenibilidad, sino que también es una palanca de generación de valor económico positivo.

Es evidente que, en la última década, la sostenibilidad se ha vuelto cada vez más importante en la agenda empresarial y de Gobierno. Las empresas y los Gobiernos están tomando medidas para integrar la sostenibilidad en sus estrategias y operaciones alineando la legislación corporativa y demás políticas con la "Agenda 2030 para el Desarrollo Sostenible", así como con los ODS, como marco común en sostenibilidad corporativa.

2.2. Dimensiones de la sostenibilidad

Dentro del concepto de sostenibilidad se diferencian tres dimensiones en las cuales se centran los esfuerzos del desarrollo sostenible: medio ambiental, económica y social. Estas interactúan entre sí de diversas maneras, y actualmente, no hay un único modelo sobre cómo estas se priorizan o se interrelacionan, ya que las tres son igual de importantes y contribuyen de la misma manera al desarrollo sostenible (Purvis et al., 2017). A continuación se explican brevemente estas dimensiones, para poder delimitar el alcance de cada una y poner en valor como acaban interactuando entre ellas.

- La dimensión ambiental de la sostenibilidad se refiere a la conservación y el uso responsable de los recursos naturales, así como a la protección y restauración del medio ambiente. Se centra en todas las acciones que llevan a una mejor gestión de los recursos naturales y la reducción del impacto negativo en el ecosistema por parte del modelo productivo y de consumo actual (Organización de las Naciones Unidas, s.f.).
- La dimensión económica de la sostenibilidad se centra en garantizar que las actividades económicas sean viables a largo plazo y contribuyan al bienestar de la sociedad en general. Esto implica fomentar el comercio justo y el crecimiento económico sostenible en todas las comunidades, así como apoyar la innovación y el progreso tecnológico para disminuir los efectos medioambientales y aumentar la eficiencia (Organización de las Naciones Unidas, s.f.).
- La dimensión social de la sostenibilidad se refiere a garantizar la justicia social, la igualdad y la equidad para todas las personas, en particular las que se encuentran en situación de vulnerabilidad. Esta se centra en acciones que garanticen la mejora de la calidad de las relaciones de la sociedad con el individuo, garantizando la igualdad de oportunidades y su bienestar (United Nations Global Compact, s.f.).

Es evidente que las tres dimensiones presentan matices similares entre ellas. Uno de los modelos más comunes para representar la interacción de las dimensiones es un modelo con tres círculos concéntricos, con la sostenibilidad como intersección de los tres (Purvis et al., 2017). También, se defiende que el pilar medioambiental predomina ante el económico y el social, pero eso es resultado de que históricamente la sostenibilidad ha

sido enfocada mayoritariamente desde la dimensión medioambiental. A partir de la década de los 2000 se empieza a introducir el concepto de sostenibilidad económica y sostenibilidad social como palancas a activar junto con la medioambiental para realmente poder aspirar al desarrollo sostenible (Purvis et al., 2017).

2.3. Camino hacia la Sostenibilidad Corporativa

Como se ha mencionado previamente, tras la realización de la Cumbre de Río de Janeiro en 1992, la sostenibilidad empieza a ser considerada como un elemento crucial para el desarrollo del mundo corporativo. Se defiende que las empresas y Gobiernos no deberían ser participantes, sino ejemplos y referentes a seguir en materia de desarrollo sostenible (Purvis et al., 2017).

La incorporación de la sostenibilidad en las corporaciones se da, en un primer momento, de la mano de la Responsabilidad Social Corporativa (RSC). Esta abarca "las expectativas económicas, legales, éticas y filantrópicas que pone la sociedad sobre la actividad de las empresas" (Carroll, 1991). La RSC es el compromiso que tienen las empresas de crear valor a sus grupos de interés mediante distintas acciones impulsadas por ellas, pero sin llegar a estar incluidas como parte de su propia estrategia.

Una de las grandes críticas hacia el concepto de RSC es su percepción y aceptación como un compromiso público social y no como una palanca de creación de valor para las corporaciones (Andreu & Fernández, 2011). Esto limita la integración de la sostenibilidad de manera holística en las cadenas de valor de las corporaciones y la posiciona como un elemento adicional de comunicación o marketing en las empresas.

El concepto de sostenibilidad corporativa nace de la necesidad de tener un enfoque integral hacia la sostenibilidad por parte de las empresas. Presenta una evolución sobre el concepto de RSC y añade la temporalidad como factor diferencial. Por ello, la sostenibilidad corporativa se puede definir como la capacidad de las empresas para crear valor a largo plazo en sus grupos de interés apalancándose en oportunidades y riesgos derivados del desarrollo económico, social y medioambiental (Andreu & Fernández, 2011; Martín García, 2019).

Dentro de la sostenibilidad corporativa encontramos los criterios ESG como "factores que convierten a una compañía sostenible a través de su compromiso social, ambiental y de buen gobierno, sin descuidar nunca los aspectos financieros" (Deloitte, 2021).

- El criterio Ambiental (*Enviromental*) evalúa el impacto que tienen las actividades de las empresas en el medio ambiente. Este incluye actividades que no solo contribuyan a la mitigación de afectos adversos en el entorno, sino a la generación de un impacto positivo como la descarbonización, uso de energía renovables o protección de la diversidad (Deloitte, 2021; Iberdrola, s.f.).
- El criterio Social (*Social*) analiza como contribuye la empresa tanto a las relaciones principales con sus grupos de interés directos (empleados o clientes) en materia de salud, bienestar o diversidad, así como con sus comunidades locales. (Deloitte, 2021; Iberdrola, s.f.).
- El criterio de Buen Gobierno (*Governance*) hace referencia a aquellos mecanismos dentro de las empresas relacionados con la transparencia, cultura interna, cumplimiento legal y ético (Deloitte, 2021; Iberdrola, s.f.).

Estos ayudan a delimitar las áreas en las que se tienen que focalizar las acciones de las empresas para maximizar el valor generado por ellas, y también crean el marco que mide el desempeño en sostenibilidad de las empresas.

Según Andreu y Fernández (2019), la consolidación de la sostenibilidad corporativa como palanca de creación de valor monetario en las empresas pasa por la consecución de cuatro premisas:

- Alejar la sostenibilidad de la filantropía
- Construir un *business case* sobre la sostenibilidad
- Gestionar la sostenibilidad de forma eficiente
- Eliminar la asociación exclusiva de sostenibilidad al medioambiente

En el siguiente apartado, se presenta el marco legal por el cuál las empresas tienen que regirse en Europa y en España en cuestiones de divulgación de información de sostenibilidad. Este cumplimiento legislativo pone en valor el compromiso adquirido con la sostenibilidad corporativa, ya que, como se verá a continuación, no solo implica hacer

públicas políticas internas si no la involucración activa en acciones dentro del marco ESG que supongan un impacto positivo tanto en el propio negocio como en el entorno.

3. ¿Cómo se tiene que publicar la información no financiera en el IBEX 35?

Como este trabajo se centra en analizar la información no financiera (INF) publicada por las empresas del IBEX 35, a continuación se introduce de forma conceptual que es el IBEX 35, quienes lo componen y cuál es su relevancia a nivel nacional.

Por otra parte, se contextualiza cual es la vigente legislación sobre divulgación en sostenibilidad a escala europea y en el ámbito español que afecta a sus miembros, así como, en qué consisten los estándares internacionales comúnmente aceptados en los que se basa la publicación de INF.

3.1. El IBEX 35

El IBEX 35 es el índice bursátil de referencia en España compuesto por los "35 valores más líquidos del mercado español" (BME, 2022). Las compañías que recoge el índice son aquellas que "cotizan en el Sistema de Interconexión Bursátil Español y en el mercado continuo de las cuatro bolsas españolas (Madrid, Barcelona, Bilbao y Valencia)" (Cerem, 2022).

La elección de las empresas que componen el IBEX 35 la realiza por un grupo de expertos conocido como Comité Técnico Asesor. Este grupo analiza qué empresas cumplen mejor con los requisitos de capitalización, liquidez y volumen de negociación, estableciendo el promedio de estos indicadores (Cerem, 2022). Cada empresa tiene un peso distinto dentro del índice en función del tamaño de su capitalización bursátil en comparación con el resto, para influir en el comportamiento del índice de manera que refleje la realidad del tejido empresarial español.

El IBEX 35 está formado por las empresas más importantes y grandes de la economía española. Dentro del índice, las empresas pertenecen a los sectores de petróleo y energía, infraestructura, bienes de consumo, servicios de consumo, servicios bancarios, servicios inmobiliarios y tecnología y telecomunicaciones (BME, s.f.).

Esté índice es una representación del desempeño de la economía española por varios motivos. Por una parte, contiene a las empresas más importantes del país pertenecientes

a los sectores nombrados previamente, reflejando diversificación empresarial dentro del propio índice (BME, 2022). A su vez, los 35 valores suponen un 90% del efectivo negociado en la Bolsa (Cerem, 2022), dando una imagen bastante aproximada del desempeño total. Se puede afirmar que el índice refleja el panorama financiero de las principales empresas para la economía española, por lo que también se puede considerar un indicador tentativo sobre su desarrollo y evolución (BME, 2022).

3.2. Marco regulatorio en información no financiera

España es parte de la Unión Europea (UE) desde junio de 1986 (Ministerios de Asuntos Exteriores, Unión Europea y Cooperación, s.f.) lo que implica que está bajo el marco regulatorio de la Unión. Esto supone el alineamiento legislativo nacional con el europeo, pero solo se va a profundizar en este trabajo en aquellas implicaciones relacionadas con la regulación en información no financiera.

En este caso, la divulgación de información no financiera es tratada por la UE mediante la aprobación de directivas. Las directivas establecen los objetivos mínimos que tienen que cumplir los Estados miembros en distintas materias, sin especificar los medios necesarios para su consecución (Comisión Europea, s.f.). Los países son los encargados de redactar la legislación que transponga el contenido de la directiva europea al marco regulatorio nacional, incorporando todo lo mencionado como obligatorio y con la libertad de incluir los aspectos voluntarios o no mencionados, siempre que estén alineados con el objetivo de la directiva (Comisión Europea, s.f.).

Esto lleva a afirmar que en el IBEX 35, la directiva por la que se rige el reporte de información no financiera es la Directiva sobre Información No Financiera (en inglés referida como NFRD), aprobada en 2014, plasmada en la Ley 11/2018 de Información no Financiera y Diversidad. A continuación, se explica el contenido de la directiva y cómo recoge sus directrices el Derecho Nacional.

Cabe mencionar la entrada en vigor en enero de 2023 de la Directiva de Reporte en Sostenibilidad Corporativa (en inglés referida como CSRD), que sustituye paulatinamente a la NFRD en sus funciones como directiva de información no financiera en la UE. A efectos de mayo de 2023, el Derecho Español no recoge las implicaciones de

esta nueva directiva ya que su trasposición se dará una vez se apruebe un proyecto de Ley en el ámbito nacional (Ministerio de Asuntos Económicos y Transformación Digital, 2023).

La información no financiera analizada más adelante corresponde al año fiscal 2021 estando bajo el paraguas de la NFRD y de la Ley 11/2018. A continuación, se profundiza en estas dos normas alineadas con el alcance de este trabajo de fin de grado, ya que las implicaciones de la CSRD todavía no se reflejan en la información no financiera de 2021.

3.2.1. Directiva de Reporte No-Financiero (NFRD)

En octubre de 2014, el Parlamento Europeo aprueba la Directiva 2014/95/UE sobre información no financiera y diversidad sobre empresas (Deloitte, 2017). Supone la creación de un marco obligatorio para todos los Estados miembros respondiendo a la necesidad de alinear los esfuerzos en materia de sostenibilidad por parte de las empresas.

La directiva supone dos grandes implicaciones en materia de transparencia sobre información no financiera. Por una parte, obliga a hacer públicas las políticas de diversidad relacionadas con la gobernanza empresarial así como su método de implementación y resultado. Por otra parte, se establece la figura del Estado de Información No Financiera (EINF), pudiendo estar integrado en el Informe Anual o presentarse de forma independiente por parte de la empresa.

La obligación de redactar un EINF es para aquellas empresas que sean de interés público o que cuenten con más de 500 empleados (Directiva 2014/95/UE). Este debe incluir información relacionada con el impacto de las actividades de la empresa en materia de ESG: protección del medioambiente, responsabilidad social y trato a empleados, respeto de los derechos humanos y lucha contra de la corrupción. También explicita la publicación de todas aquellas políticas internas, procesos de *due diligence*, riesgos de la actividad empresarial asociados o indicadores de desempeño, alineados con el marco ESG (Deloitte, 2017).

Como se ha mencionado brevemente, la directiva supone un marco legal sobre el cual los Estados miembros establecen su propia legislación garantizando su cumplimiento, pero

pudiendo ir un paso más allá. Según Deloitte, un 80% de los Estados miembros ha transpuesto la directiva en leyes más estrictas (2017), siendo España uno de ellos. Estas modificaciones se materializan en el aumento del alcance de la normativa, el establecimiento de sanciones para aquellas empresas no cumplidoras o la introducción de la obligatoriedad de verificar por un tercero la información reportada.

3.2.2. Divulgación de Información No Financiera en España: Ley 11/2018

En línea con la Directiva 2014/95/UE, en diciembre de 2018 entra en vigor la Ley 11/2018 por la que se modifica el marco regulatorio en España en materia de información no financiera y diversidad. En este caso, la trasposición de ley amplía el alcance de algunos ámbitos al igual que hace más estrictos otros. A continuación se indican las especificaciones relevantes que recoge la Ley española:

- *Número de sociedades afectadas*: la Ley 11/2018 amplía el número de empresas obligadas a publicar el EINF. Se obliga a reportar a las empresas que o (i) tengan más de 250 empleados o (ii) sean de interés público o (iii) tengan un activo superior 20 millones o un importe neto de la cifra anual de negocios superior a 40 millones de euros (Garrigues, 2018).
- *Contenido del EINF*: el Derecho nacional aumenta el contenido que debe de ser incluido en el EINF. Se incorpora información adicional alineada con el compromiso de la empresa con la sociedad y su contribución al desarrollo sostenible, así como una descripción más detallada del modelo de negocio (Garrigues, 2018).
- *Verificación externa*: el reglamento español incluye la obligatoriedad de que un tercero audite y verifique la información presentada en el EINF, el cual deberá emitir un informe a incluir en el EINF (Garrigues, 2018).

Dentro del artículo publicado por Garrigues (2018) se hace referencia a otras puntualizaciones recogidas por la Ley relacionadas con el número de mujeres en consejos de dirección, las facultades indelegables del consejo, la transparencia en políticas de diversidad, así como la obligación de elaborar un EINF para el Consejo Estatal de Responsabilidad Social de las Empresas (CERSE).

La Ley española recomienda el uso de diferentes estándares internacionales para estructurar los informes no financieros. A continuación, se enumeran cuatro de los estándares de publicación más usados a nivel internacional, siendo *Global Reporting Initiative* (GRI) el más aceptado y utilizado en España.

3.3. Estándares Internacionales de Referencia

Debido a la falta de homogeneidad en la divulgación de información no financiera y en los indicadores que tienen que publicar las empresas en materia de ESG, la mayoría de los informes de sostenibilidad se adscriben a distintos estándares internacionales mencionados a continuación.

Según un informe de Deloitte, un 100% de las empresas españolas que publican la información no financiera lo hacen bajo algún marco internacional (2017). Al final, los estándares internacionales sirven para verificar por un tercero que la información difundida está alineada con la legislación actual, y ofrece homogeneidad a un proceso complejo y en el que influyen muchos matices diferentes (Deloitte, 2017). Dentro de los estándares internacionales destacan:

- *Global Reporting Initiative (GRI)*: conjunto de indicadores que informan sobre los impactos económicos, sociales y medioambientales de las empresas. Dentro de estos se presentan tanto los impactos positivos como los negativos, contribuyendo con la identificación de las temáticas materiales en sostenibilidad para las empresas (GRI, s.f.).
- *Pacto Mundial de Naciones Unidas (Global Compact)*: conjunto de diez principios para incluir en las estrategias empresariales en materia de derechos humanos, trabajo, medioambiente y lucha contra la corrupción (Pacto Mundial de las Naciones Unidas, s.f.).
- *Marco Integral de Reporting Integrado (IIRC):* marco teórico que facilita el reporte de información integrado de las empresas para poder tener una asignación de capital más efectiva alineada con la estabilidad financiera y el desarrollo sostenible (IIRC, s.f.).

- *Sustainability Accounting Standards (SASB)*: conjunto de normas que identifican el grupo de cuestiones medioambientales, sociales y de gobernanza más relevantes para los resultados financieros de cada sector. Actualmente, está disponible en un total de 77 sectores (SASB, s.f.).

Estos estándares no son solo un marco de divulgación de información no financiera, sino una herramienta que ayuda a focalizar los recursos y capacidades de una empresa en materia de sostenibilidad.

Guiarse por estándares internacionales es una práctica voluntaria que aporta veracidad a la información difundida por las empresas hasta ahora. Acogerse a ellos supone un compromiso adicional con la transparencia en información no financiera y con la sostenibilidad de las actividades de cada empresa.

4. *Text Mining*

El *text mining* es un campo de estudio enfocado en la extracción de información y conocimiento relevante a partir de grandes cantidades de datos textuales. Se trata de un proceso intensivo en el que un usuario interactúa con un conjunto de documentos utilizando distintas herramientas de análisis con el objetivo de extraer información de patrones observados en el texto (Feldman & Sanger, 2006).

El *text mining* nace como respuesta al aumento exponencial de información no estructurada generada a diario y a la falta de recursos y capacidades para poder procesarla e interpretarla. Esta información no estructurada no presenta ningún tipo de esquema o modelo de datos prestablecido (Telefónica, s.f.), creando la necesidad de realizar un procesamiento de la información previo a utilizar cualquier algoritmo de *text mining*. Con ello se convierte la información no estructurada a un formato intermedio con la suficiente estructura para posteriormente poder emplear técnicas de análisis (Feldman & Sanger, 2006). Esta disciplina se nutre de distintas técnicas de análisis destacando las siguientes (Kalra & Kumar, 2013):

- La recuperación de la información (*Information Retrieval*) es el proceso por el cual se obtienen datos de un texto en función de distintos parámetros definidos por el usuario (Kalra & Kumar, 2013). Su objetivo es reducir la dimensión y tamaño del texto y facilitar su posterior análisis. Este proceso agrupa técnicas como la tokenización o el *steeming*, que ayudan a simplificar y agrupar la información dentro del texto en el proceso inicial de *text mining* (IBM, s.f.)
- El procesamiento de lenguaje natural (*Natural Language Processing*) consta de un conjunto de técnicas que facilitan la interpretación del contenido del texto mediante el análisis de la estructura y gramática de las palabras y oraciones que conforman el texto (Kalra & Kumar, 2013). Dentro de estas técnicas se encuentran la síntesis, el etiquetado gramatical (*POS tagging*), la categorización de textos o el análisis de sentimiento, los cuales permiten a los ordenadores comprender y procesar el lenguaje humano en forma escrita (IBM, s.f.).
- La extracción de información (*Information Extraction*) consiste en la obtención de patrones o datos relevantes buscados en varios documentos (IBM, s.f.). Dentro

de la extracción de información se encuentran técnicas como la selección de atributos, la extracción de características o la identificación de entidades que ayudan a seleccionar las características relevantes de los textos para reducir la dimensionalidad de los datos y facilitar su análisis posterior (IBM, s.f.).

- La minería de datos (*Data Mining*) se puede describir como la búsqueda de patrones y extracción de información de conjuntos de datos masivos (Kalra & Kumar, 2013). Esta técnica se usa tanto para analizar datos estructurados y no estructurados, así como para identificar nueva información, tendencias en los datos y predecir futuros comportamientos (Kalra & Kumar, 2013). El *text mining* es un campo de la *data mining* centrado en ordenar y analizar datos no estructurados.

Dentro del *text mining* se encuentran muchas otras técnicas además, de las previamente mencionadas, pero fuera del alcance de este trabajo de fin de grado. Todas ellas contribuyen al objetivo de la aplicación del *text mining*: facilitar la transformación de datos masivos no estructurados a estructurados y poder identificar patrones relevantes y extraer conclusiones significativas (IBM, s.f.).

4.1. Notación y terminología en *text mining*

Dentro de un proceso de *text mining*, se utiliza la siguiente terminología para referirse a las distintas unidades de información (Blei et al., 2003):

- Palabra: unidad básica de datos discretos, que no proporciona información relevante al estudiarla de forma aislada
- Documento: secuencia de N palabras, que puede darse en forma de artículo, informe, libro, etc.
- *Corpus*: conjunto de documentos objeto de estudio

A lo largo de este trabajo de fin de grado, se utilizan de manera frecuente los términos nombrados para describir diversos aspectos al referirse al funcionamiento de un proceso de *text mining*, explicar la técnica de análisis seleccionada y describir el análisis realizado en la base de datos elegida.

4.2. Proceso de análisis en *text mining*

Como se ha mencionado previamente, el *text minng* incorpora distintas técnicas de tratamiento y análisis de información textual durante su aplicación. A continuación, se explican las etapas más frecuentes dentro de un proceso genérico de *text mining* (Allahyari et al., 2017; Kalra & Kumar, 2013):

1. El primer paso en *text mining* implica obtener los documentos que serán objeto de estudio. Esto se puede lograr de dos maneras: utilizando una base de datos existente que contenga los documentos relevantes o creando manualmente la base de datos mediante la descarga y agrupación de los textos de interés de diversas fuentes en un *corpus*. La elección depende de la disponibilidad de una base de datos adecuada y de los requisitos del proyecto.
2. El pre-procesamiento del *corpus* consiste en la limpieza y simplificación de los documentos identificados de cara a facilitar su posterior análisis. Al estar tratando con datos no estructurados, esta etapa es crucial para poder reducir la complejidad de análisis y acercarse a una base de datos lo más homogéneos en estructura posible. El pre-procesamiento es considerado como la parte más importante y costosa de la *text mining*, ya que su resultado tiene implicaciones a lo largo de todo el proceso. Esta parte incluye, entre otras, las siguientes técnicas:
    - Tokenización: divide el *corpus* en palabras o frases denominándolos tokens.
    - Eliminación de patrones gramaticales: elimina signos de puntuación o exclamación dentro del *corpus*.
    - Limpieza del *corpus*: reduce toda información no necesaria ni requerida dentro de un *corpus* como la eliminación de imágenes o la normalización de texto presentado en tabla, gráficos u otro tipo de formatos.
    - *Steeming*: reduce los tokens identificados a su raíz.
    - Etiquetado gramatical: clasifica los tokens en función de su categoría gramatical.
    - Lematización: identifica la categoría morfológica de cada token en función de su significado.

3. Una vez limpiado el *corpus*, la generación de atributos supone la creación de un documento de texto que incluya los tokens identificados previamente, así como su frecuencia absoluta o relativa. Se puede realizar mediante dos enfoques:
    - *Bag of Words*: presenta cada documento como un conjunto de tokens, teniendo en cuenta la frecuencia de cada uno en el texto.
    - Modelo de espacio vectorial: transforma los documentos en vectores numéricos, incorporando la importancia (ponderación) de cada token dentro del documento.
4. La selección de atributos utiliza la información proporcionada de cada token, para establecer criterios y priorizar en qué atributos se centra el análisis. Se basa en la asunción de que los datos contienen características relevantes, las cuales hay que reducir.
5. Una vez obtenida una base de datos estructurada, formada por documentos, atributos y tokens, se realiza el análisis a través de *data mining*. Dentro de esta fase, el usuario decide qué algoritmo incorporar en función de las necesidades y objetivos de su análisis y las características de su base de datos. Las técnicas más conocidas son:
    - Clasificación: asigna categorías pre-establecidas a las palabras que conforman un texto con el objetivo de poder predecir la temática del documento en cuestión
    - *Clustering*: agrupa textos de forma automática en clústeres en función de cómo de parecido es el contenido o significado de cada documento.
    - *Topic Modeling*: establece las temáticas subyacentes del conjunto de textos en función de la repetición de palabras y su significado.
6. Finalizado el proceso de *data mining*, se evalúan los resultados obtenidos y el grado de cumplimiento con las hipótesis planteadas. Se extraen conclusiones de los datos obtenidos y, también, se considera su uso como datos de entrada en otro potencial proceso de *data mining*.

El *text mining* es un proceso iterativo, el cual se adapta a las necesidades del texto utilizado y los problemas que puedan surgir. Realizar un pre-procesamiento claro y eficiente es clave para el resto del análisis, ya que de no darse esto, cualquier resultado o

conclusión obtenida no va a reflejar la realidad del problema. Por ello, es fundamental destacar que el pre-procesamiento, la generación de atributos y su selección no necesariamente siguen un enfoque lineal, ya que se ajustan a las características específicas de los datos analizados.

4.3. Aplicaciones del *text mining* en la empresa

El *text mining* ha impactado de forma transversal en distintos procesos de todas las industrias que trabajan con información no estructurada, haciéndolos más automáticos y eficientes. A continuación, se nombran algunas de las aplicaciones del text *mining* en el día a día de una empresa:

- Gestión de riesgos: El *text mining* puede proporcionar información sobre las tendencias del mercado y la evolución del sector, y en consecuencia, utilizarse en la gestión de riesgos para intentar anteponerse a los cambios de tendencia. Gestionar los riesgos de forma eficiente, puede llevar a aumentar la confianza de los clientes.
- Atención al cliente: Al pedir opiniones al usuario, las empresas obtienen información textual sobre la experiencia de cliente. La combinación de utilizar herramientas de análisis de texto junto con sistemáticas de retroalimentación como encuestas a clientes permite identificar debilidades en su *customer journey* y mejorarlo de forma rápida y precisa. El *text mining* puede ofrecer a las empresas una forma de clasificar las reclamaciones más importantes de los clientes, permitiendo priorizar la respuesta a distintos problemas y aumentar la satisfacción del cliente.
- Filtrado de *curriculums*: El uso del *text mining* en los departamentos de recursos humanos supone la automatización de procesos intensivos en tiempo y dedicación como puede ser la lectura de *curriculums*. Durante una revisión manual de un currículum, se buscan errores, credenciales educativas, experiencia laboral y otra información personal. Estos datos pueden extraerse automáticamente como primer paso del proceso de filtrado de currículos.
- Aumento de la productividad: Introducir el *text mining* en los procesos que requieran de un análisis exhaustivo de información textual, implica el incremento

de su efectividad y eficiencia. La automatización de la lectura de información puede suponer una reducción de costes y gasto en tiempo muy beneficioso para las empresas, ya que consiguen el mismo objetivo de manera más eficiente y destinando menos recursos.

En definitiva, el *text mining* se ha convertido en una herramienta indispensable para empresas de todos los sectores, brindando ventajas competitivas y mejorando la toma de decisiones, la satisfacción del cliente y la eficiencia operativa. Su capacidad para procesar y comprender grandes volúmenes de texto de manera automatizada y precisa es un activo valioso en el mundo empresarial actual.

5. *Topic Modeling*: Asignación Latente de Dirichlet

Este trabajo de fin de grado utiliza el *topic modeling* como técnica de análisis de texto. Este se puede materializar en distintos tipos de algoritmos que llevan a identificar temáticas de forma no supervisada, es decir, sin previa especificación de estas temáticas. La técnica seleccionada para este trabajo es el modelo de Asignación Latente de Dirichlet, en ingles *Latent Dirichlet Allocation model* (modelo LDA), explicada a continuación.

La Asignación Latente de Dirichlet, introducida por Blei et al. (2003), es un modelo probabilístico generativo utilizado en el procesamiento de lenguaje natural para extraer temáticas latentes en un *corpus* de texto. Se basa en un enfoque bayesiano jerárquico de tres niveles y se considera una de las técnicas más populares en el campo del *topic modeling*.

En el contexto del modelo LDA cada documento se considera como una colección de palabras sin un orden específico (enfoque *bag of words*). Estas palabras son utilizadas para generar una combinación de temas basada en probabilidades asignadas. A su vez, cada tema se representa como una distribución de palabras en el *corpus*. Esto implica que las palabras estrechamente relacionadas con el tema principal de un documento tienen una mayor probabilidad de aparecer en él (Blei et al., 2003). Este enfoque nos permite modelar la estructura latente y subyacente del *corpus*, identificando los temas que lo componen y cómo se distribuyen a lo largo de los documentos seleccionados.

El modelo LDA se define, entre otras formas, por la siguiente ecuación (Blei et al., 2003):

$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) = p(\theta \mid \alpha) \prod p(z_n \mid \theta)\, p(w_n \mid z_n, \beta)$$

Esta ecuación representa la distribución, *theta* (θ), conjunta de cada tema identificado a través de los *z* temas totales y *w* palabras influenciadas por distintos parámetros que ajustan el resultado del modelo, siendo N el número de topics al que nos vamos a referur como K :

1. El parámetro *alpha* (α) marca la cantidad de temas que se identifican en cada documento. Un valor alto de *alpha* implica que los documentos tienden a abarcar

diversas temáticas, mientras que un valor bajo supone la presencia pocas temáticas principales en cada documento.

2. Por otro lado, el parámetro *beta* (β) controla la diversidad de palabras dentro de cada tema. Un valor alto de *beta* indica que los temas contienen una amplia variedad de palabras relevantes, mientras que un valor bajo sugiere que cada tema se compone principalmente de unas pocas palabras dominantes.
3. Por último, K es el número de temas que van a ser obtenidos del *corpus*, una vez ha sido aplicado el modelo LDA.

El uso de los parámetros en el método LDA depende de la elección específica de la técnica utilizada. Sin embargo, el parámetro K es común a todas las metodologías y es de particular importancia. A continuación, se busca comprender su significado, cómo se selecciona su valor y por qué es crucial llevar a cabo este proceso de manera adecuada.

La elección del número de temas (K) es una decisión crítica en el uso de LDA. Establecer valor alto de K puede llevar a la definición de temas demasiado específicos mientras que un valor bajo supone tópicos simples e insignificantes en el cómputo global.

Como en cualquier algoritmo que requiere la elección del valor de un parámetro de forma manual, existen métodos para optimizar y determinar el número adecuado de temáticas. Estas técnicas ayudan a encontrar un equilibrio entre tener un número suficiente de temáticas para capturar la diversidad del *corpus* y evitar un exceso de temas que puedan llevar a una interpretación confusa.

Para el desarrollo de este trabajo de fin de grado, se han seleccionado las siguientes tres métricas en la optimización del número de temas del modelo LDA:

1. La métrica de densidad (*density*) mide el número de palabras asignadas a un tema en particular. Esta métrica es importante para determinar la diferencia entre los temas identificados por el modelo. El objetivo es minimizar el valor de la densidad, lo que significa que se busca tener una asignación precisa de palabras a cada tema. Esta métrica, propuesta por Cao et al. (2009), se basa en el concepto de densidad utilizado en *clustering*. En el caso de *topic modeling*, se busca

minimizar la diferencia dentro de cada tema y maximizarla entre temas para obtener una separación clara y significativa.

2. La divergencia intra-tópico (*within-topic divergence*) evalúa la coherencia y calidad de los temas generados. Esta métrica, propuesta por Arun et al. (2010), se centra en medir la cohesión semántica de cada tema identificado. Se analiza qué tan coherentes y relacionadas están las palabras dentro de un tema en comparación con el *corpus* general. Una baja divergencia intra-tópico indica que las palabras dentro de un tema están altamente relacionadas y tienen una coherencia semántica sólida, lo que contribuye a una mejor interpretación y comprensión del tema.

3. La divergencia inter-tópico (*across-topic divergence*) evalúa la separación entre los temas generados. Esta métrica, propuesta por Deveaud et al. (2014), se enfoca en medir la diversidad temática entre los textos seleccionados. Evalúa qué tan diferentes y separados están los temas entre sí. Una alta divergencia inter-tópico indica que los temas son distintos y se abordan diferentes aspectos del *corpus*, lo que puede ser beneficioso para representar la diversidad y riqueza de los temas presentes en los datos.

Estas métricas permiten una evaluación integral y sólida del *topic modeling*, asegurando que los tópicos generados sean precisos, coherentes y diversos. Determinar el valor óptimo de K a introducir en el modelo LDA, supone encontrar el valor en el que se minimice la densidad y la divergencia intra-tópico, y se maximice la divergencia inter-tópico. Al optimizar la elección de temáticas, se logra obtener una representación significativa y comprensible de los datos, lo que facilita su interpretación y análisis en diversas aplicaciones.

Una vez optimizado K, se aplica el modelo sobre la base de datos especificando el número de temas deseados. El proceso de LDA sigue una forma iterativa compuesta por los siguientes pasos (Silge & Robinson, 2021):

1. Asignación aleatoria de un tema a cada palabra en los documentos del *corpus*.
2. Cálculo de la probabilidad de que una palabra aparezca en el tema asignado y de que un tema aparezca en un documento.

3. Optimización de las probabilidades, ajustando los valores para maximizar la coherencia y la representatividad de los temas. Este proceso se repite durante un número específico de iteraciones, que se establece previamente en función del grado de precisión que busque el usuario.
4. Obtención de un listado de temas que incluye las distribuciones y probabilidades asociadas a las palabras y las temáticas. Destaca *gamma* (γ), que mide la distribución de cada palabra por cada tema, y *theta* (θ) que representa como se distribuyen los temas a través de los documentos. El alcance de este trabajo se limita a el uso de *theta* para identificar los temas más relevantes (Jones, T. W., 2019).

Una vez comprendido el propósito y funcionamiento del modelo LDA, y vista la parte explicativa del trabajo de fin de grado, se pasará a implementar los modelos descritos. A continuación, se describe el proceso de análisis y obtención de resultados buscando comprender mejor el contenido de los informes en materia de sostenibilidad de las empresas durante el ejercicio 2021.

6. Aproximación a la información no financiera del IBEX 35 mediante *text mining* y *topic modeling*

Este trabajo de fin de grado analiza cuáles son los grandes temas abarcados por la información no financiera publicada por las empresas del IBEX 35 sobre el ejercicio 2020-2021, dado que es la última información disponible. Se busca comprobar si las empresas del IBEX 35 están distanciándose de la dedicación exclusiva a la dimensión medioambiental y filantrópica, para establecer un modelo de negocio que integre transversalmente la sostenibilidad.

Este análisis se ha realizado en Rstudio, ya que proporciona las librerías y funciones necesarias para analizar información masiva en texto y aplicar los algoritmos seleccionados de *topic modeling*. Todos los procesos y conclusiones obtenidos han sido realizados en esta herramienta, mediante el código incluido en el anexo 2 de este trabajo.

A continuación, se describe el proceso realizado, desde la búsqueda y agrupación de los informes seleccionados bajo una misma base de datos textuales hasta la obtención y análisis de los temas latentes gracias al modelo LDA.

6.1. Creación y descripción de la base de datos

Actualmente, no existe una base de datos que agrupe la información pública de todas las empresas del IBEX 35 en materia de sostenibilidad, más allá de los propios informes en sostenibilidad de cada empresa alineados con la Ley 11/2018. Dado que este trabajo de fin de grado busca analizar la información no financiera (INF) del IBEX 35 de forma conjunta, se ha creado una base de datos de texto que cumpla con esta función.

El anexo 1 recoge en una tabla las empresas del IBEX 35; todas han sido incluidas en este análisis. Se diferencian siete sectores: Materiales Básicos, Industria y Construcción; Servicios de Consumo, Tecnología y Telecomunicaciones, Petróleo y Energía, Servicios Financieros, Servicios Inmobiliarios y Bienes de Consumo. Entre estos destacan empresas como ACS, AENA, Telefónica, Iberdrola, Banco Santander, Inmobiliaria Colonial o Inditex.

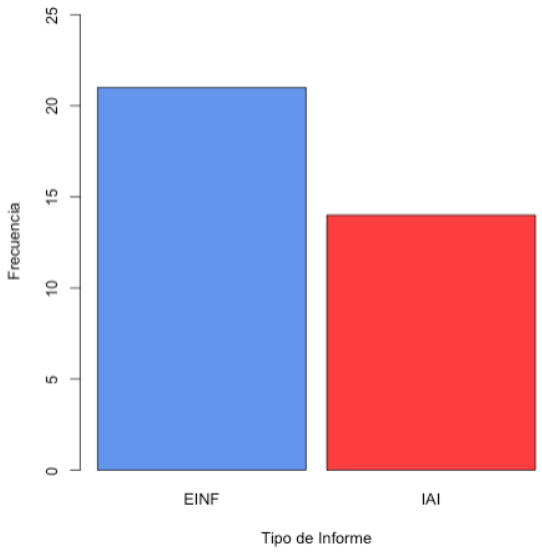
La base de datos se crea en Rstudio mediante la carga y lectura del contenido de cada informe. Estos se obtienen mediante la búsqueda y descarga manual de los Estados de Información no Financiera (EINF) incluidos en el apartado de "Accionistas e Inversores" en la página oficial de cada empresa. Dentro de este proceso, surge la problemática de que no todas las empresas incluidas publican la INF bajo un formato homogéneo.

Las empresas no están obligadas por la regulación actual a presentar la INF bajo un formato específico, existiendo distintas formas de publicarla. En este caso, los distintos formatos identificados se han agrupado en dos bloques para facilitar el análisis:

1. El primer bloque se ha denominado EINF e incluye los EINFs, memorias o informes de sostenibilidad disponibles de las empresas del IBEX.
2. El segundo, denominado Informe Anual Integrado (IAI), recoge los informes anuales e informes de gestión consolidada de aquellas empresas que no presentan la INF con entidad propia.

La Figura 1 muestra la distribución de formatos de los informes recogidos. Un 60% de las empresas publican la información en sostenibilidad de manera independiente, mientras que un 40% lo hace de manera integrada en su informe anual.

Figura 1: *Distribución de formato en el IBEX 35 de publicación de la INF*



Fuente: Elaboración Propia

En el alcance de este trabajo no entra valorar cual es el formato adecuado de presentación de la INF. Si que es importante destacar, de cara a los resultados del análisis, que tener la INF integrada junto con el resto de información en el Informe Anual Integrado (IAI), contamina la base de datos y su razón de ser en materia de sostenibilidad. Esto da mayor importancia a la fase de pre-procesamiento y limpieza del *corpus*, buscando homogeneizar el contenido de los informes y eliminar el efecto que pueda tener sobre el análisis la existencia de información puramente de negocio.

6.2. Obtención, limpieza y filtrado del *corpus*

El siguiente paso consiste en unificar los informes en Rstudio mediante la función corpus() para obtener el *corpus* sobre el que se realiza el análisis, formado por 35 filas, que representan cada uno de los informes incluidos en la base de datos. La Figura 2 muestra la estructura del *corpus*, incluyendo por cada línea el contenido textual de los informes. Solo recoge los seis primeros por predeterminaciones de Rstudio.

Figura 2: *Contenido del corpus en Rstudio*

```
Corpus consisting of 35 documents.
text1 :
"                                                principales ..."

text2 :
"informe anual integrado        2021 1              informe d..."

text3 :
"informe de gestión consolidado 2021  extracto correspondient..."

text4 :
"informe global 2021 análisis de la actividad empresarial, fi..."

text5 :
"m   e   m   o   r   i   a   d   e   s   o   s   t   e   n   ..."

text6 :
"m   e   m   o   r   i   a   d   e   s   o   s   t   e   n   ..."

[ reached max_ndoc ... 29 more documents ]
```

Fuente: Elaboración Propia; salida de Rstudio

Una vez obtenido, se realiza la tokenización del *corpus* dividiendo cada documento en unidades más pequeñas de información. En este caso, la tokenización busca obtener de

manera individual las palabras que contiene cada informe seleccionado, para poder realizar un análisis más detallado y estructurado del contenido.

Limpiar el *corpus* supone eliminar las palabras que no aporten significado y que puedan influir en los resultados de este trabajo. En este caso, se utilizan las siguientes técnicas de limpieza del *corpus* mediante la función tokens_remove():

- Eliminación de signos de puntuación, exclamación e interrogación, números, enlaces web, letras independientes.
- Eliminación de palabras vacías (*stopwords*), es decir, palabras poco significativas que no aportan información relevante. Estas palabras suelen ser artículos, pronombres y conjunciones.
- Eliminación de términos característicos. En este caso, se han eliminado manualmente un total de 255 palabras. Estas están relacionadas, entre otros temas, con el ámbito geográfico y sector de las empresas del IBEX o con términos frecuentes sobre el ámbito operativo empresarial. La Tabla 1 muestra ejemplos de palabras eliminadas dentro de los ámbitos mencionados.

Tabla 1: *Muestra de palabras eliminadas del corpus*

Ámbito Geográfico	Ámbito Sectorial	Ámbito Operativo
España	Construcción	Gestión
Portugal	Energías renovables	Resultados
Madrid	Acero inoxidable	Estrategia
Barcelona	Banca privada	Cuentas
Valencia	Gestión aeroportuaria	Auditoría

Fuente: Elaboración Propia

Una vez realizado el primer filtrado, común a todos los procesos de *text mining*, se usa el enfoque *bag of words.* Este transforma el texto no estructurado en una matriz término-documento (MTD), representando cada fila un documento del *corpus* y cada columna

una palabra. La MTD contiene información sobre la frecuencia de cada palabra identificada en el *corpus* en cada documento.

La Figura 3 muestra la matriz MTD obtenida por Rstudio. Se recogen las frecuencias de los diez primeros términos identificados en los seis primeros documentos. La matriz contiene 35 filas, una por cada informe analizado, y 59.530 columnas, representando una palabra diferente por columna.

Figura 3: *Muestra del contenido de la MTD obtenida en Rstudio*

	features									
docs	principales	cifras	magnitudes	operativas	financieras	reexp	cifra	negocios	beneficio	bruto
text1	117	1	4	38	42	38	26	58	93	17
text2	36	7	0	1	6	0	5	0	7	1
text3	106	4	0	10	10	0	7	13	6	0
text4	55	35	0	2	14	0	4	52	12	1
text5	69	19	0	2	9	0	35	112	12	3
text6	45	15	0	1	7	0	17	42	11	2

[reached max_ndoc ... 29 more documents, reached max_nfeat ... 59,520 more features]

Fuente: Elaboración propia; salida de Rstudio

Las dimensiones de la matriz muestran que el número de palabras identificadas es de 59.530, suponiendo una alta dimensionalidad del *corpus*, pudiendo ser consecuencia de que el *corpus* recoge la misma palabra en diferentes formas. En la Figura 3 se observa que "cifras" y "cifra" han sido identificados como términos independientes, cuando en realidad tienen el mismo significado y podrían contabilizarse como una única palabra. Realizar *steeming* sobre la matriz ayudaría a reducir su tamaño, debido a la identificación de las raíces de cada palabra.

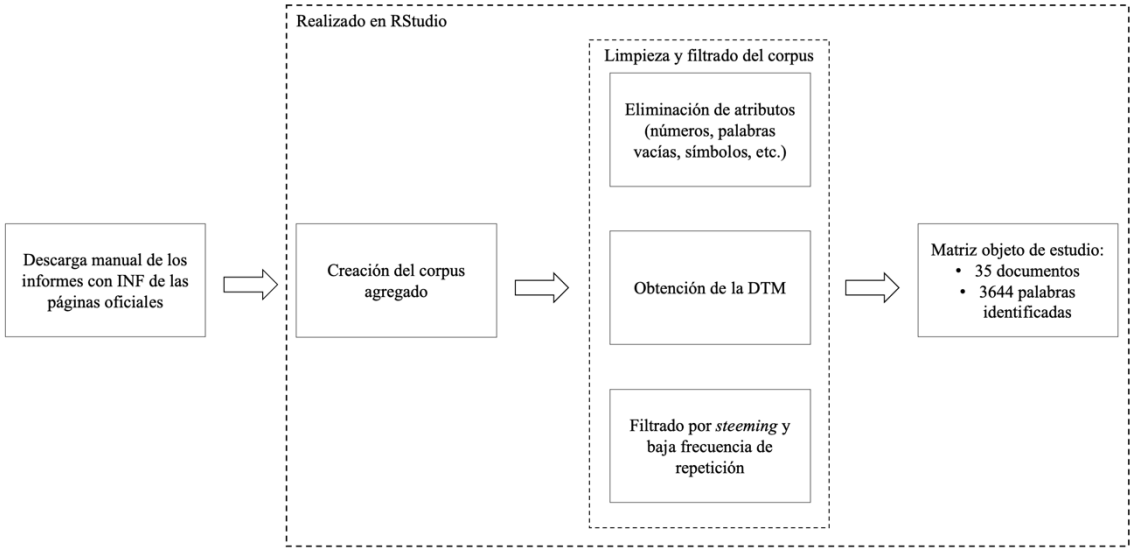
El uso de *steeming* elimina sufijos y prefijos de cada palabra para obtener solo su raíz o forma base. Su uso hace que palabras como "sostenibilidad" o "sostenibles" se reduzcan a la forma base "sostenible". La aplicación de la función dfm_worstem() lleva a una reducción considerable del número de palabras identificadas, obteniendo 28.088.

Para concluir el pre-procesamiento, así como la generación y selección de atributos que van a ser analizados, se eliminan palabras con frecuencias muy bajas. Esto se sustenta en la necesidad de reducir el ruido que suponen las palabras con poca frecuencia, ya que no aportan información suficientemente significativa

Por ello, se decide eliminar todas aquellas palabras con una frecuencia inferior a 35, buscando reducir el tamaño del *corpus* y que las palabras a analizar sean representativas del contenido de los informes. De esta manera, se obtiene una matriz MTD con 3.644 palabras, teniendo unas dimensiones asumibles de cara a realizar un análisis exploratorio de términos frecuentes y, más tarde, la implementación del modelo LDA para obtener las temáticas subyacentes del *corpus*.

La Figura 4 representa el flujo de trabajo realizado hasta finalizar el pre-procesamiento del *corpus*, resaltando qué partes han sido realizadas en Rstudio.

Figura 4: *Flujo de trabajo de la obtención de la base de datos textuales*



Nota: incluye los procesos descritos en el capítulo 6.1 y 6.2. Fuente: Elaboración propia

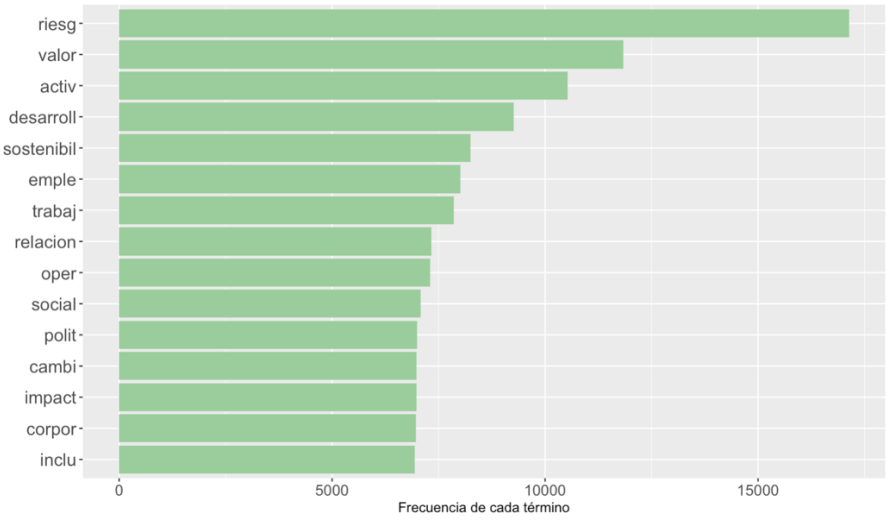
6.3. Análisis exploratorio del *corpus*

El conjunto de datos sobre el que se realiza el análisis es un *corpus* reducido de 3.644 palabras con múltiples subtemas subyacentes que serán obtenidos más adelante mediante la implementación del modelo LDA. La realización de un análisis exploratorio previo en base a la frecuencia de las palabras da una imagen inicial de los posibles temas a identificar por el modelo.

6.3.1. Análisis de términos más frecuentes

La Figura 5 muestra los 15 términos más frecuentes en orden descendente, así como su frecuencia. Cabe recordar que cada término representa la forma base de varias palabras, agrupando su significado bajo este. Cada término individual proporciona poca información sobre las grandes temáticas del corpus, aunque se identifican principalmente dos que están claramente relacionadas con temas ESG (e.g., "sostenibil", "social") y operacionales (e.g., "riesg", "activ"), entre otros. Los temas, idealmente, van a consistir en la combinación de palabras relacionadas que son habituales a lo largo del texto.

Figura 5: *Términos más frecuentes en la INF del IBEX35*



Fuente: Elaboración propia

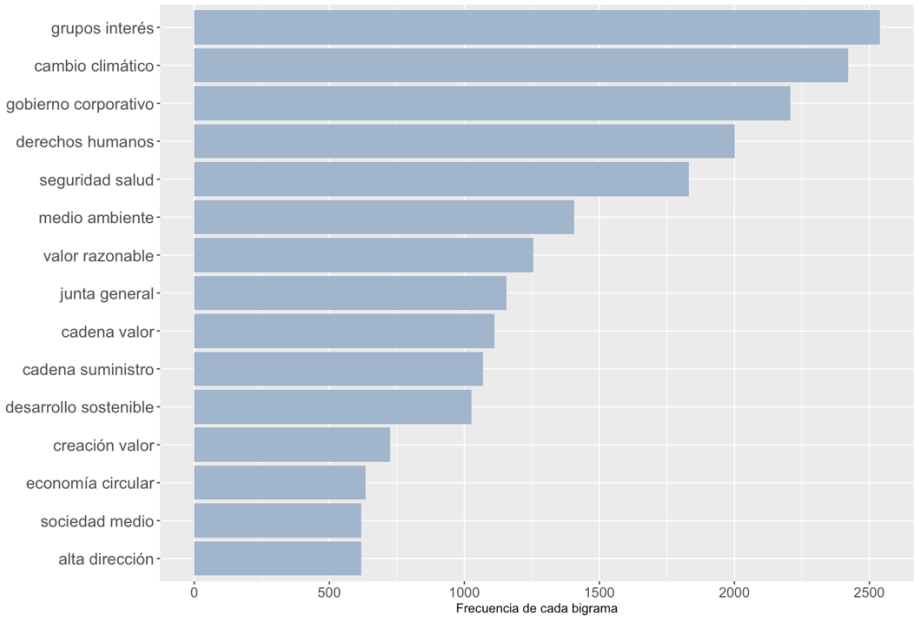
Al observar la Figura 5, destaca la alta frecuencia del término "riesg", superando ampliamente al segundo término, "valor". Esta diferencia significativa no es coincidencia, ya que la generación de valor a partir de la sostenibilidad requiere una gestión efectiva de los riesgos en el modelo de negocio de las empresas. Además, se puede apreciar la presencia de términos relacionados con temas sociales como "emple", "trabaj", "inclu" o "social", lo cual se alinea con la creciente importancia que la sostenibilidad social está adquiriendo en las empresas españolas**.** Es sorprendente observar que solo hay un término directamente relacionado con la sostenibilidad

medioambiental, “cambi”, siendo esta la dimensión sobre la que tradicionalmente se divulga más información (Transcendent, 2022).

6.3.2. Análisis de bigramas más frecuentes

En paralelo, se han construido bigramas, es decir, grupos de dos palabras que aparecen juntas a lo largo del *corpus*. El análisis de bigramas permite capturar combinaciones de palabras que tienen un significado conjunto y proporcionan una mayor comprensión de los temas abordados en la INF pública. Este análisis exploratorio se ve limitado porque los bigramas solo se pueden obtener a partir de la matriz MTD original, y no a partir de la matriz objeto de estudio.

Figura 6: *Bigramas más frecuentes en la INF del IBEX35*



Fuente: Elaboración propia

A pesar de esta situación, la Figura 6 muestra los bigramas más repetidos a lo largo de la base de datos, dando una idea más completa sobre la esencia del texto que la frecuencia de cada término mostrada en la Figura 5. En este caso, las grandes temáticas identificadas a simple vista son sostenibilidad medioambiental (e.g., “cambio climático”) y social (e.g., “derechos humanos”, “seguridad salud”, “economía circular”), gobernanza (e.g., “grupos

interés", "gobierno corporativo") y operaciones (e.g., "valor razonable", "cadena suministro").

En los términos y los bigramas más frecuentes, se observa la presencia de la sostenibilidad social. Esto confirma que la dimensión social en ESG ha adquirido una importancia especial durante el año 2021 a nivel empresarial, y es probable que haya experimentado un aumento en comparación con el año anterior.

Resulta destacable la presencia de bigramas relacionados con gobernanza, lo cual es interesante, dado que suele ser el aspecto menos mencionado dentro de los tres criterios ESG. Esta tendencia puede ser atribuida a que un 40% de las empresas de la base de datos presenta la INF dentro de su IAI. En dicha información suelen incorporarse datos sobre la gobernanza operativa y ejecutiva de la empresa, así como prácticas dirigidas a los accionistas, no pudiendo afirmar con certeza que sea una temática con entidad propia.

6.3.3. Análisis sectorial de términos y bigramas más frecuentes

Al obtener la base de datos, se identifican siete sectores a los que pertenecen las empresas del IBEX. Con el objetivo de tener una comprensión general de los términos más comunes en cada sector y poder realizar comparaciones entre ellos, se lleva a cabo un proceso de filtrado de documentos en basa a la actividad operativa de cada empresa. El filtrado se aplica tanto al *corpus* simplificado que contiene los textos de los documentos, como al objeto que contiene los bigramas generados previamente.

La Tabla 2 contiene los cinco términos más frecuentes en la INF de 2021 de cada sector del IBEX. Si tenemos en cuenta los 5 términos más frecuentes en el *corpus* ("riesg", "valor", "activ", "desarroll", "sostenibl") visualizados en la Figura 6, el único presente en el top 5 de todos los sectores es "riesg". Es posible que los riesgos derivados de la integración de la sostenibilidad en el negocio, no se limita a un solo sector, sino que es una tendencia transversal al IBEX.

A su vez, "valor" está entre los términos más comunes en cuatro de los siete sectores identificados. Se asienta con más fuerza la idea de que la sostenibilidad, aun suponiendo

un riesgo para muchos modelos de negocio, de ser integrada y gestionada de manera correcta, puede suponer la creación de valor adicional en el negocio sin comprometerlo.

Tabla 2: *Términos más frecuentes por sector en la INF del IBEX35*

Mat. Básicos	Serv. de Consumo	Tec, y Telecom	Petróleo y Energía	Serv. Financieros	Serv. Inmobiliarios	Bienes de Consumo
riesgo	riesgo	valor	sostenibilidad	riesgo	valor	trabajo
desarrollo	empleo	riesgo	riesgo	valor	riesgo	riesgo
actividad	política	consolidado	valor	cliente	sociedad	impacto
impacto	trabajo	base	electricidad	actividad	actividad	desarrollo
trabajo	seguridad	actividad	desarrollo	financiero	ejercicio	sostenible

Fuente: Elaboración propia

Viendo los términos más frecuentes en Materiales Básicos, Servicios de Consumo, Tecnología y Telecomunicaciones, y Servicios Financieros, destaca la atención significativa hacia la gestión de riesgos y la creación de valor a través operaciones. Esto indica que estas industrias son conscientes de los posibles riesgos en materia ESG asociados con sus actividades y están tomando medidas para mitigarlos. Además, están enfocadas en aprovechar oportunidades de desarrollo y generar valor en sus respectivos mercados alineados con la sostenibilidad.

En el caso específico del sector de Petróleo y Energía, se observa un cambio de paradigma en cuanto a la consideración de la sostenibilidad. Los términos más repetidos, como "sostenibilidad" y "riesgo", indican que las empresas de este sector están dejando de ver la sostenibilidad como algo negativo y están integrándola en sus estrategias y operaciones. Esto sugiere que están reconociendo la importancia de la sostenibilidad como un elemento clave para generar mayor valor a largo plazo.

La Tabla 3 sigue el mismo razonamiento que la Tabla 2, pero identificando los bigramas más repetidos por sector. Analizar la frecuencia de los bigramas más repetidos por sector proporciona una imagen más completa sobre posibles temáticas subyacentes del *corpus*. En este caso, no hay ningún bigrama que sea común a los siete sectores, pero se identifican

grupos de bigramas comparables: cuestiones medioambientales (e.g., “cambio climático”, “medio ambiente”, “emisiones gei”) y cuestiones sociales (e.g., “derechos humanos”, “salud seguridad”).

Tabla 3: *Bigramas más frecuentes por sector en la INF del IBEX35*

Mat. Básicos	Serv. de Consumo	Tec, y Telecom	Petróleo y Energía	Serv. Financieros	Serv. Inmobiliarios	Bienes de Consumo
desarrollo sostenible	derechos humanos	medio ambiente	grupos interés	gobierno corporativo	sociedad dominante	grupos interés
cambio climático	grupos interés	cadena valor	cambio climático	valor razonable	gobierno corporativo	cadena suministro
derechos humanos	cambio climático	sociedad medio	derechos humanos	grupos interés	junta general	derechos humanos
grupos interés	salud seguridad	personas sociedad	sostenibilidad cambio	cambio climático	grupos interés	impacto positivo
prácticas buen	cuestiones sociales	gobernanza personas	creación valor	económico riesgos	emisiones gei	cambio climático

Fuente: Elaboración propia

Si se profundiza por sectores, se encuentra que:

- En Materiales Básicos y Servicios de Consumo, destacan los bigramas “desarrollo sostenible”, “derechos humanos”, “salud seguridad” y “grupos interés” pudiendo indicar una atención significativa en la dimensión social focalizada en impactos en empleados y grupos de interés externos.
- En Tecnología y Telecomunicaciones, la alta frecuencia “medio ambiente” y “cadena valor” puede suponer un aumento de atención hacia los impactos ambientales y la gestión transparente de las partes interesadas.
- Las empresas de Petróleo y Energía, a través de “cambio climático”, “derechos humanos” y “creación valor”, pueden estar mostrando una creciente preocupación por abordar los desafíos ambientales y sociales asociados con la industria energética
- En Servicios Financieros, se pone énfasis en la valoración financiera y la gestión adecuada de los riesgos económicos, mediante la repetición de “valor razonable” y “económico riesgos”.

- Mientras que en el sector de Servicios Inmobiliarios, destacan las menciones a "gobierno corporativo" y "junta general" indicando una posible relevancia de la gobernanza en este sector, y también, destaca la aparición de "emisiones gei" (gases efecto invernadero).
- Finalmente, en Bienes de Consumo, resalta "cadena suministro", "grupos interés" y "derechos humanos" pudiendo referirse a la gestión efectiva de la cadena de valor integrando las necesidades de los grupos de interés.

El análisis de los términos y los bigramas proporciona una visión inicial de los temas clave presentes en el *corpus* de INF. No es ninguna sorpresa las menciones repetidas a aspectos medioambientales, ya que es la dimensión más "popular" dentro de la sostenibilidad, pero destacan las múltiples apariciones de términos sociales a través de los sectores, tanto a nivel de palabras como a nivel de bigramas.

Sin embargo, para obtener una comprensión más profunda y precisa de los temas, es necesario aplicar el modelo LDA al *corpus*. Este enfoque permite descubrir temas específicos alineados con los identificados, pero con mayor aproximación a la temática real de la INF publicada por las empresas del IBEX 35.

6.4. Implementación del modelo LDA

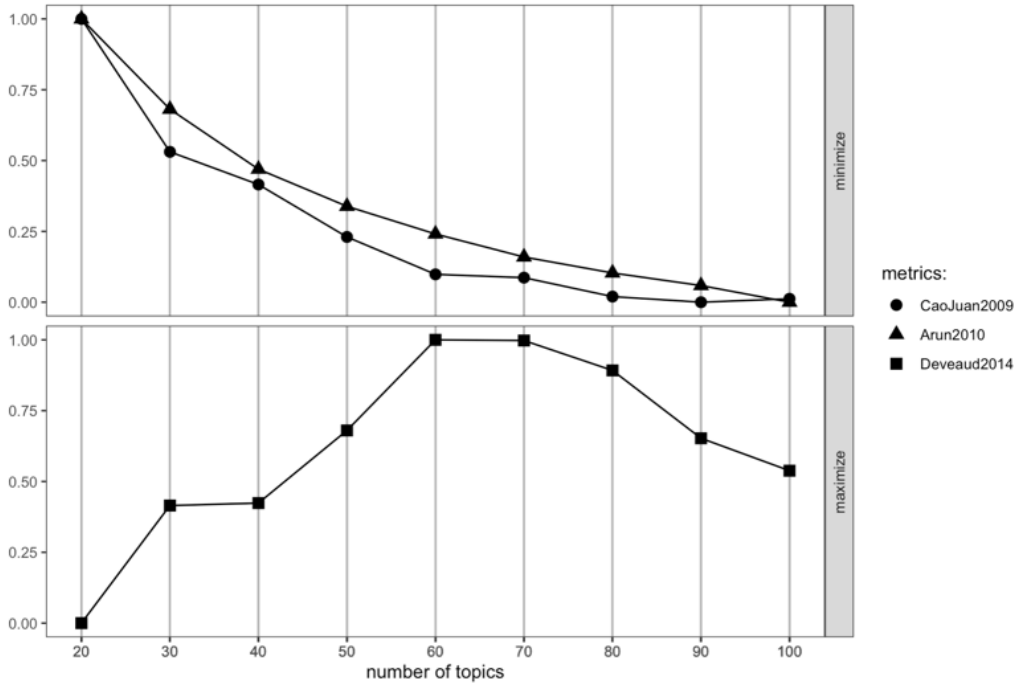
Previo a la implementación del modelo LDA, se han eliminado los términos del *corpus* con frecuencia mayor a 9.000. Una palabra que ha sido altamente mencionada a lo largo de un texto no aporta un valor diferencial significativo. En este aspecto se sigue el mismo razonamiento por el que, al inicio en el pre-procesamiento, se eliminan palabras mencionadas muy pocas veces al no ser relevantes. Se busca evitar un listado de temas muy similares centrados alrededor de los términos más repetidos como son "riesgo", "valor" o "activo".

Dentro del funcionamiento del modelo LDA planteado en el capítulo 5, se menciona que el número de temas a obtener se establece de manera manual. Se puede introducir un número de manera aleatoria y asumir el riesgo de que no se ajuste bien a las necesidades del *corpus*, o utilizar distintas métricas para optimizar el parámetro K.

En este caso, se busca encontrar el valor de K que minimice la densidad y la divergencia inter-tópico, al tiempo que maximice la divergencia entre tópicos. Para lograr esto, se utiliza la función FindTopicsNumber(). Esta función calcula estas métricas para diferentes valores de K, que van desde 20 hasta 100, con incrementos de 10. El objetivo es obtener el valor óptimo de K que permita obtener resultados más claros y distintos entre los diferentes tópicos identificados en el análisis.

La Figura 7 representa el resultado obtenido de la optimización de los tres parámetros. Según el criterio previamente establecido, el mejor resultado de *topic modeling* se obtiene dividiendo el *corpus* en 70 temas latentes.

Figura 7: *Optimización del valor de K*[1]



Fuente: Elaboración propia

El resultado de la aplicación del modelo LDA es un listado de 70 temas que representan las temáticas subyacentes que hay en la INF del 2021 de las empresas del IBEX. Cada tema contiene términos que se relacionan en el texto de forma frecuente. A continuación,

[1] Las variables mostradas en la leyenda representan a las métricas presentadas en el capítulo 5: densidad, divergencia intra-tópico y divergencia inter-tópico

se estudia cuáles son los temas más relevantes en el *corpus*, así como cuales son los temas más frecuentes por cada documento.

6.5. Análisis de resultados

Revisar los temas obtenidos por el modelo LDA no es asumible en ningún análisis de *topic modeling* cuando se obtiene tal escala de temas, en este caso 70. Es por ello que se utiliza el valor de *theta* (**θ**) para calcular qué temas son los más frecuentes dentro del corpus, y basar nuestro análisis en los cinco temas más recurrentes a lo largo de los informes en INF del IBEX en 2021.

La Tabla 4 representa los cinco temas con mayor valor de θ entre las 70 temáticas obtenidas del *corpus* mediante el modelo LDA. Dentro de cada tema hay un número extenso de términos, pero por facilitar el análisis se han enunciado aquellos más representativos y prioritarios dentro de cada uno.

Tabla 4: *Temas con mayor valor de theta (θ) obtenidos en el modelo LDA*

Tema 1	retribución, presidente, ejecutivo, sociedad, modelo, etc.
Tema 2	sostenibilidad, análisis, compañía, impacto, persona, cambio, etc.
Tema 3	seguro, empleo, trabajo, persona, social, cliente, etc.
Tema 4	sostenibilidad, operaciones, trabajo, emisión, impacto, dato, etc.
Tema 5	compañía, empleo, compromiso, sostenibilidad, emisión, etc.

Fuente: Elaboración propia

De los cinco temas seleccionados, se ha hecho un análisis individual de que contenidos o aspectos relacionados con la sostenibilidad y operatividad del negocio puede englobar cada uno.

El primer tema engloba el modelo de retribución ligado a la sostenibilidad. Con palabras como “retribución” o “ejecutivo”, se materializa el compromiso actual de incorporar objetivos ESG dentro de la retribución variable. Las recomendaciones de buen gobierno de la Comisión Nacional de Mercados y Valores (CNMV) expresan que esta se debe

establecer en base a criterios financieros y no financieros, siempre promoviendo la sostenibilidad y la rentabilidad de la empresa en el largo plazo (CNMV, 2015). La aparición de este tema puede estar unida a la presencia de información financiera dentro del *corpus*, producto de las memorias anuales incluidas.

El segundo habla sobre la sostenibilidad y la gestión del impacto. La presencia de términos como 'sostenibilidad', "impacto", "cambio" o "análisis", se alinean con la nueva forma que está tomando la integración de la sostenibilidad actualmente, que es la medición y gestión del impacto, estando presente en la agenda ESG de las empresas de 2021.

La siguiente temática se relaciona con bienestar y seguridad laboral. En este tema, la presencia de "empleo", "persona" o "social" se alinean con lo observado en el análisis exploratorio sobre la sostenibilidad social. Mientras que la medioambiental ha sido la dimensión ESG predominante, en los últimos años la dimensión social está ganando peso y asentando su importancia en las prioridades de las empresas.

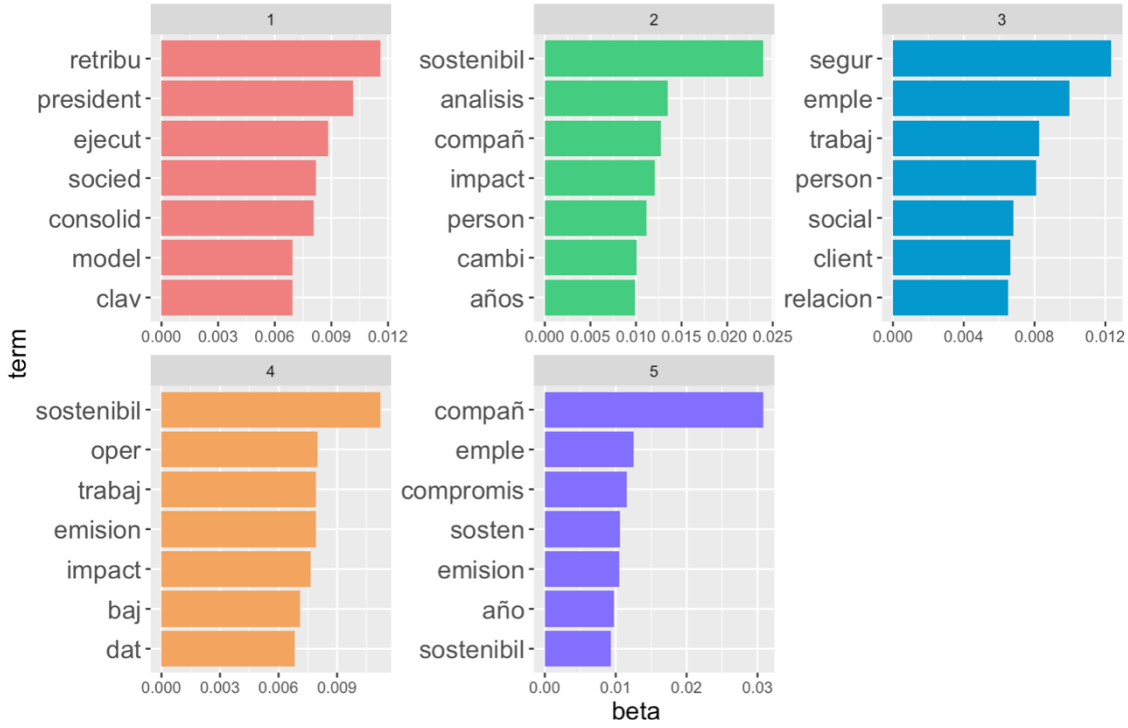
El cuarto habla sobre la sostenibilidad de las operaciones. Este tema se alinea con el segundo, pero con foco en la empresa a través de palabras como "operaciones", "trabajo" o "dato". Se introduce el matiz de las operaciones que no es sorpresa dado que se está tratando INF sobre las empresas, de la cual el 40% contiene también información sobre el resto de sus operaciones y líneas de negocios.

Por último, destaca el compromiso empresarial con la sostenibilidad. Términos como "compromiso", "emisiones" o "empleo", hacen referencia a posibles objetivos que se marcan las empresas en temática ambiental, social y de buen gobierno. En el 2021, el 54% de las empresas del IBEX realizaron compromisos públicos en materia de diversidad, comunidades, empleados o clientes (Transcendent, 2022). El establecimiento y cumplimiento de objetivos cuantitativos es, actualmente, una de las mayores palancas de integración de la sostenibilidad en el día a día de las empresas.

La Figura 8 muestra cuál es la frecuencia relativa de los siete términos más mencionados por tema. Dentro de los temas analizados, el uno y tres tienen una distribución de palabras medianamente equitativa, sin que predomine una sobre las otras de manera evidente.

Mientras que en los temas dos, cuatro y cinco; encontramos palabras con valores significativamente más altos que los demás términos, dando lugar a una alta influencia del término dentro del tema.

Figura 8: *Frecuencia relativa de cada término en los temas seleccionados*



Fuente: Elaboración propia

Por lo tanto, se puede concluir que el uso del modelo LDA para obtener las tendencias de la INF es un proceso satisfactorio pero complejo, que requiere de un alto nivel de detalle por parte del usuario. Aunque se han identificado temas que se alinean con lo observado durante la fase exploratoria, es necesario tener un profundo conocimiento sobre sostenibilidad y la situación empresarial actual para poder contextualizar los temas más allá de su mera identificación y realmente aportar valor con este análisis. Dicho esto, los temas obtenidos indican que las empresas españolas están alejándose del enfoque tradicional de Responsabilidad Social Corporativa mencionado en el capítulo 2 de este trabajo, y se están centrando en la Sostenibilidad Corporativa a través de sus operaciones, la gestión de impacto y el bienestar de su personal.

7. Conclusiones

Este trabajo concluye tras haber realizado un análisis exhaustivo del contenido de la información no financiera del IBEX 35 correspondiente al ejercicio 2021, tanto a escala global como sectorial. Además, se han identificado los temas subyacentes en la base de datos analizada mediante una técnica de *topic modeling*, en este caso, el modelo LDA (Asignación Latente de Dirichlet).

En línea con los objetivos de este trabajo, se ha empezado contextualizando la situación actual en el mundo corporativo y regulatorio en materia de sostenibilidad, pasando a estudiar la estructura y aplicaciones de un proceso de *text mining* en el mundo empresarial. Posteriormente, se ha realizado un análisis exploratorio sobre la información no financiera obtenida y agrupada para, por último, terminar implementando el modelo LDA sobre una base de datos uniforme.

Para poder obtener datos y conclusiones que garantizasen la veracidad de los resultados, se han utilizado fuentes fiables como aquellas pertenecientes a Naciones Unidas, Ministerios de España, la Unión Europea, páginas oficiales de las empresas incluidas en la base de datos o informes de consultoras de referencia en materia de sostenibilidad.

A lo largo de este trabajo, se ha evidenciado cómo la falta de estándares de cumplimiento obligatorio a la hora de reportar información ESG en la Directiva 2014/95/UE y en la Ley 11/2018, dificulta la comparación del grado de madurez de la sostenibilidad entre empresas y la identificación de temáticas relevantes. En el ejercicio 2020-2021, el 60% de las empresas del IBEX publicaron su información no financiera de manera específica, mientras que el resto lo integró en su informe anual.

La necesidad de establecer un formato único también se refleja en los resultados del análisis exploratorio del apartado de *text mining*. Es importante destacar que la presencia de términos relacionados con la operatividad empresarial y el ámbito financiero entre los más comunes puede explicarse por el enfoque integrado que adoptan el 40% de las empresas del IBEX al divulgar su información no financiera, lo cual puede generar una representación no del todo precisa de las palabras más frecuentes.

Sin embargo, la Directiva 2022/2464, que reemplazará a la actual directiva sobre información de sostenibilidad (NFRD), introduce cambios sobre el formato de publicación. Esta nueva directiva establece la obligación de integrar los contenidos del informe específico de sostenibilidad en el informe de gestión anual (Directiva 2022/2464). A partir del ejercicio 2022-2023, la publicación de INF reflejará esta modificación, lo que facilitará su análisis y comparabilidad.

En el cuerpo del análisis exploratorio, se ha obtenido una visión aproximada de los temas relevantes en la agenda ESG del IBEX. La presencia predominante de palabras como “riesgo”, “valor” y “activo” en el *corpus* sugiere que las empresas están abandonando la percepción de la sostenibilidad como una amenaza para su negocio, y están adoptando una perspectiva más optimista al considerarla como una oportunidad de generar valor a través de sus actividades empresariales. Además, la aparición de estos términos sugiere que existe una tendencia hacia la integración de la gestión de riesgos ESG dentro de la cadena de valor, con el objetivo de maximizar la actividad económica empresarial.

A su vez, se ha llevado a cabo un análisis de términos y bigramas a escala global y sectorial, y se ha observado que hay un tema transversal en ambos niveles de análisis relacionado con la sostenibilidad social. La presencia de términos como "empleo", "cliente", "derechos humanos", "salud y seguridad" o "grupos de interés" refleja la creciente tendencia de otorgar mayor importancia al bienestar de las personas, ya que son el motor de las empresas. Este enfoque puede ser resultado del impacto que tuvo la pandemia del COVID-19 en el cuidado de los empleados, así como la necesidad de establecer empresas que generen valor para las personas.

Es sorprendente que las menciones relacionadas con el medio ambiente se vean reducidas a términos como "cambio climático", "medio ambiente" y "emisiones de gases de efecto invernadero". Históricamente, esta dimensión ha recibido una mayor atención por parte de las empresas, pero en este análisis su presencia es menor en comparación con los aspectos sociales.

Finalmente, la obtención de los temas subyacentes de la información no financiera del IBEX mediante la optimización del modelo LDA lleva a varias ideas principales:

1. El uso de *topic modeling* facilita la identificación de temáticas relevantes en la información no financiera. Sin embargo, es importante tener conocimientos previos en sostenibilidad para contextualizar adecuadamente estos temas. Además, la mezcla de formatos de publicación de la información puede influir en los resultados, por lo que es necesario interpretarlos teniendo en cuenta esta consideración.
2. Los temas identificados reflejan un cambio en la percepción de la sostenibilidad, alejándose de un enfoque meramente filantrópico de la Responsabilidad Social Corporativa y centrándose en la integración de la sostenibilidad en el negocio. Tres de los cinco temas prioritarios están relacionados con la interacción de los criterios ESG con las operaciones de las empresas, estando los otros dos relacionados con la retribución ligada a sostenibilidad y el bienestar y seguridad laboral.
3. La sostenibilidad social se posiciona como una prioridad para las empresas en las tres dimensiones ESG. La identificación de un tema específico relacionado con aspectos sociales confirma la idea recurrente a lo largo del trabajo de que el compromiso con el aspecto social es uno de los pilares importantes para las empresas.

Finalmente, se puede destacar que con el análisis realizado y los datos obtenidos, han surgido conclusiones interesantes como la necesidad de homogenizar el formato de presentación en INF, la creciente importancia de la sostenibilidad social en el ámbito nacional o el funcionamiento acertado de *topic modeling* sobre información textual con alta dimensionalidad. De esta manera, se concluye que la integración transversal de la sostenibilidad en el modelo de negocio es prioritario para las empresas, garantizando la continuidad de sus operaciones y la creación de valor adicional, apalancándose en las dimensiones ESG.

Basándose en las conclusiones de este trabajo, se pueden identificar varias líneas de investigación futuras. En primer lugar, sería interesante realizar un análisis similar una vez entre en vigor la Directiva 2022/2464 y se publiquen los informes de gestión que contengan la INF de las empresas. El objetivo sería comparar los temas identificados en

este estudio con los obtenidos después de la implementación de la regulación, para determinar si ha facilitado la estandarización y comparabilidad de la información.

En segundo lugar, sería relevante plantear este análisis en diferentes ámbitos geográficos para examinar cómo varían las prioridades en función del país. Aunque en el ámbito europeo se esperarían similitudes debido a la aplicación de la Directiva, en países como Estados Unidos, con una regulación menos estricta, es probable que se observen diferencias significativas en los información divulgada.

Estas líneas de investigación permitirían ampliar el conocimiento sobre la evolución de la integración de la sostenibilidad en las empresas, el impacto de la regulación en la identificación de temas relevantes y la comparación de las prioridades en diferentes contextos geográficos.

8. Bibliografía

Acciona. (2022). *Memoria de Sostenibilidad: 2021.* [Archivo PDF]. https://mediacdn.acciona.com/media/q3ihslvb/memoria-de-sostenibilidad-acciona-2021.pdf

Acciona Energía. (2022). *Memoria de Sostenibilidad: 2021*. [Archivo PDF]. https://procoazrbolsast1.blob.core.windows.net/media/vaaj0yw2/memoria-de-sostenibilidad-2021-acciona-energia.pdf

Acerinox. (2022). *Estado de Información No Financiera (EINF).* [Archivo PDF]. https://www.acerinox.com/export/sites/acerinox/.content/galerias/galeria-descargas/Informe-Anual-Integrado.pdf

ACS. (2022). *Informe Integrado del Grupo ACS: 2021.* [Archivo PDF]. https://www.grupoacs.com/ficheros_editor/File/03_accionistas_inversores/03_informe_anual/2021/INFORME%20INTEGRADO%202021.pdf

AENA. (2022). *Informe de Gestión Consolidado:2021. Extracto correspondiente al Estado de Información no Financiera (EINF).* [Archivo PDF]. https://www.aena.es/es/corporativa/rc/balance-sostenibilidad/estado-de-la-informacion-no-financiera.html

Allahyari, M., Assefi, M., Gutierrez, J., Kochut, K., Pouriyeh, S., Safaei, S. y Trippe, E. (28 de julio de 2017). A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. *KDD Bigdas*.

Amadeus. (2022). *Informe Global 2021. Análisis de la actividad empresarial, financiera y de sostenibilidad*. [Archivo PDF]. https://corporate.amadeus.com/documents/es/recursos/informacion-empresarial/documentos-corporativos/informes-globales/2021/informe-global-de-amadeus-2021.pdf

Andreu, A., y Fernández, J.L. (Diciembre 2011). De la RSC a la sostenibilidad corporativa: una evolución necesaria para la creación de valor. *Ediciones Deusto*.

https://www.researchgate.net/publication/297757785_De_la_RSC_a_la_sostenibilidad_corporativa_una_evolucion_necesaria_para_la_creacion_de_valor#fullTextFileContent

ArcelorMittal. (2022). *Informe de Sostenibilidad 2021.* [Archivo PDF]. https://spain.arcelormittal.com/wp-content/uploads/informe-sostenibilidad-2021.pdf

Arun, R., Suresh, V., Veni Madhavan, C.E., y Narasimha Murthy, M.N. (2010). On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. *Lecture Notes in Computer Science*.

Banco Sabadell. (2022). *Estado de Información no Financiera del ejercicio anual terminado el 31 de diciembre de 2021.* [Archivo PDF]. https://www.grupbancsabadell.com/corp/files/1454451075811/einf_bs_2021_cas.pdf?bsb=RmlsZV9DLTE0NTQ0NTEwNzU4MTEtMTM3NDA5ODA3OTg5NQ

Banco Santander. (2022). *Informe Anual 2021.* [Archivo PDF]. https://www.santander.com/content/dam/santander-com/es/documentos/informe-financiero-anual/2021/ifa-2021-informe-financiero-anual-consolidado-es.pdf

Bankinter. (2022). *Estado de la Información no Financiera Consolidado 2021*. [Archivo PDF]. https://www.bankinter.com/file_source2/webcorporativa/estaticos/pdf/informacion-corporativa/banca-sostenible/informes-anuales/EINFC%20consolidado_2021_informa_auditoria.pdf

BBVA. (2022). *Informe Anual 2021*. [Archivo PDF]. https://accionistaseinversores.bbva.com/wp-content/uploads/2022/03/Informe-Anual-2021.pdf

Blei, D., Ng, A.Y., y Jordan, M.I. (Marzo de 2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research*. 3. 993-1022.

BME. (s.f.). Precios. https://www.bolsasymercados.es/bme-exchange/es/Mercados-y-Cotizaciones/Acciones/Mercado-Continuo/Precios/IBEX-35-ES0SI0000005

BME. (Abril 2022). *Normas técnicas para la composición y cálculo de los índices de Sociedad de Bolsas*. [Archivo PDF]. https://www.bolsasymercados.es/bme-exchange/docs/docsSubidos/Indices/Regulacion/Normas_Indices_IBEX_esp.pdf

CaixaBank. (2022). *Informe de Gestión Consolidado*. [Archivo PDF]. https://www.caixabank.com/deployedfiles/caixabank_com/Estaticos/PDFs/Accionistasinversores/Informacion_economico_financiera/InformeGestionConsolidado_interactivo_alta_CAS.pdf

Cao, J., Xia, T., Li, J., Zhang, Y. y Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*. 72(7-9), 1775-1781.

Carroll, A.B. (1991). The pyramid of corporate social responsibility: Toward the moral management of organizational stakeholders. *Business Horizons*, 34 (4).

Cellnex. (2022). *Informe Anual Integrado. Informe de gestión Consolidado, Cuentas Anuales Consolidadas.* [Archivo PDF]. https://informeanualintegrado2021.cellnextelecom.com/files/PDF_Consolidado_ESP31.12.2021.pdf

Comisión Europea. (s.f.). *Tipos de legislación de la UE*. https://commission.europa.eu/law/law-making-process/types-eu-law_es

Comisión Mundial sobre el Medio Ambiente y el Desarrollo. (1987). Nuestro futuro común. *Fondo de Cultura Económica*. [Archivo PDF]. https://sustainabledevelopment.un.org/content/documents/5987our-common-future.pdf

Comisión Nacional del Mercado de Valores. (Febrero de 2015). Código de buen gobierno de las sociedades cotizadas, revisado en junio 2020. [Archivo PDF]. https://www.cnmv.es/DocPortal/Publicaciones/CodigoGov/CBG_2020.pdf

Convención Marco de las Naciones Unidas sobre el Cambio Climático. (1998). *Protocolo de Kioto*. https://unfccc.int/es/kyoto_protocol

Deloitte. (Diciembre 2017). *¿Hacia dónde se dirige España en materia de información no financiera? Análisis de las implicaciones para las empresas españolas.* [Archivo PDF]. https://www.dirse.es/wp-content/uploads/2018/01/180110-Directiva_Informacion_no_Financiera_Web.pdf

Deloitte. (14 de diciembre de 2021). *Qué son los criterios ESG y para qué sirven. Factores ambientales, sociales y de buen gobierno se cuelan dentro de los balances financieros*. [Archivo PDF]. https://www2.deloitte.com/es/es/blog/sostenibilidad-deloitte/2021/que-son-criterios-esg-para-que-sirven.html

Deveaud, R., SanJuan, E. y Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique*. 17, 61-84.

Directiva 2014/95/UE del Parlamento Europeo y Consejo, de 22 de octubre de 2014, por la que se modifica la Directiva 2013/24/UE en lo que respecta a la divulgación de información no financiera e información sobre diversidad por parte de determinadas grandes empresas y determinados grupos. Diario Oficial de la Unión Europea L nº 330/2, de 15 de noviembre de 2014.

Directiva 2022/2464 del Parlamento Europeo y Consejo, de 14 de octubre de 2022, por la que se modifica el reglamento (UE) n.º 537/2014, la Directiva 2004/109/CE, la Directiva 2006/43/CE y la Directiva 2013/34/UE, por lo respecta a la presentación de información sobre sostenibilidad por parte de las empresas. Diario Oficial de la Unión Europea L nº 322/15, de 16 de diciembre de 2022.

Enagás. (2022). *Informe Anual 2021.* [Archivo PDF]. https://www.enagas.es/content/dam/enagas/es/ficheros/sala-de-comunicacion/publicaciones/informe-anual/INFORME%20ANUAL%202021_ENAGAS.pdf

Endesa. (2022). *Estado de Información no Financiera y Sostenibilidad 2021*. [Archivo PDF]. https://www.endesa.com/content/dam/enel-es/home/inversores/infoeconomicafinanciera/resultadosfinancieros/documentos/2021/fy/estado-de-informacion-no-financiera-y-sostenibilidad-2021.pdf

Feldman, R., y Sanger, J. (2006). The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge: Cambridge University Press.

Ferrovial. (2022). *Informe de Gestión.* [Archivo PDF]. https://informeanualintegrado2021.ferrovial.com/wp-content/uploads/sites/7/2022/02/ferrovial-informe-anual-integrado-2021-informe-de-gestion.pdf

Fluidra. (2022). *Informe Integrado 2021.* [Archivo PDF]. https://www.fluidra.com/uploads/media/default/0001/04/a902ef2e12c9018415bae35f98c28a506b3c80d9.pdf

Garrigues. (31 de diciembre de 2018). *Publicada la ley en materia de información no financiera y diversidad en España.* https://www.garrigues.com/es_ES/noticia/publicada-la-ley-en-materia-de-informacion-no-financiera-y-diversidad-en-espana

Global Reporting Initiative. (s.f.). *GRI Standards Spanish Translations.* https://www.globalreporting.org/how-to-use-the-gri-standards/gri-standards-spanish-translations/

Grifols. (2022). *Informe Anual Integrado y de Sostenibilidad 2021.* [Archivo PDF]. https://www.grifols.com/documents/3625622/3683813/integrated-report-2021-es.pdf/9242303b-ef57-a76d-d043-804647d36420?t=1651214384342

IAG. (2022). *Estado de Información no Financiera consolidado 2021.* [Archivo PDF]. https://www.iairgroup.com/~/media/Files/I/IAG/AGM%202022/Estado%20de%20informacin%20no%20financiera%20consolidado.pdf

Iberdrola. (2022). *Estado de Información no Financiera. Informe de sostenibilidad.* [Archivo PDF]. https://www.iberdrola.com/documents/20125/1606413/jga22_IA_InformeSostenibilidad2021.pdf

Iberdrola. (s.f.). *ESG, ¿cómo realizar inversiones sostenibles y responsables?* https://www.iberdrola.com/compromiso-social/criterios-esg

IBM. (s.f.). *What is text mining?* https://www.ibm.com/topics/text-mining

Inditex. (2022). *Estado de Información no Financiera 2021.* [Archivo PDF]. https://static.inditex.com/annual_report_2021/es/documentos/estado-de-informacion-no-financiera-2021.pdf

Indra. (2022). *Informe de Sostenibilidad 2021.* [Archivo PDF]. https://www.indracompany.com/sites/default/files/indra_informe_sostenibilidad_es_2021_0.pdf

Inmobiliaria Colonial. (2022). *Informe Integrado Anual 2021.* [Archivo PDF]. https://www.inmocolonial.com/sites/default/files/uploaded-files/2022-06/COLONIAL_2021_ESP_WEB.pdf?_gl=1*d1rb5f*_up*MQ..*_ga*NTAxMjMxMTc3LjE2ODA1MzI4NTQ.*_ga_HGQ4EV0Y42*MTY4MDUzMjg1NC4xLjEuMTY4MDUzMjg3NC4wLjAuMA

International Integrated Reporting Council. (s.f.). *The IIRC: About us.* https://www.integratedreporting.org/the-iirc-2/

Kalra, P. y Kumar, L. (Marzo de 2013). *Text Mining: Concepts, Process and Applications. Journal of Global Research in Computer Science*. 4(3). https://www.researchgate.net/publication/277160258_TEXT_MINING_CONCEPTS_PROCESS_AND_APPLICATIONS

Laboratorios Rovi. (2022). *Estado de Información no Financiera 2021.* [Archivo PDF]. https://www.rovi.es/sites/default/files/informe_de_verificacion_y_estado_de_informacion_no_financiera_grupo_rovi_2021.pdf

Ley 11/2018 de 28 de diciembre, por la que se modifica el Código de Comercio, el texto refundido de la Ley de Sociedades de Capital aprobado por el Real Decreto Legislativo 1/2010, de 2 de julio, y la Ley 22/2015, de 20 de julio, de Auditoría de Cuentas, en materia de información no financiera y diversidad (BOE núm.167, de 14 de julio de 1998).

Logista. (2022). 2021. *Informe Anual Integrado.* [Archivo PDF]. https://www.logista.com/content/dam/documents/logista-corporate/corporate-governance/annual-integrated-reports/es/IAI%202021_ES_web.pdf

Mapfre. (2022). *Informe Integrado 2021.* [Archivo PDF]. https://www.mapfre.com/media/accionistas/2022/06-informe-integrado-2021.pdf

Meadows, D. H., Meadows, D. L., Randers, J., y Behrens, W. W. (1972). The limits to growth: a report for the Club of Rome's project on the predicament of mankind. *Universe Books*. http://www.ask-force.org/web/Global-Warming/Meadows-Limits-to-Growth-Short-1972.pdf

Meliá Hotels. (2022). *Informe de gestión 2021.* [Archivo PDF]. https://www.meliahotelsinternational.com/es/ourCompany/Documents/Hist%C3%B3ricoInforme/Informe_de_Gesti%C3%B3n_2021_Melia_Hotels_International.pdf

Merlin Properties. (2022). *Memoria de Sostenibilidad: 2021*. [Archivo PDF]. https://ir.merlinproperties.com/wp-content/uploads/2022/05/MP_RSC_2021_CAS.pdf

Ministerios de Asuntos Exteriores, Unión Europea y Cooperación. (s.f.). *España y la Unión Europea.* https://www.exteriores.gob.es/es/PoliticaExterior/Paginas/EspanaUE.aspx

Ministerios de Derechos Sociales y Agenda 2030. (s.f.). *Conoce la Agenda.* https://www.mdsocialesa2030.gob.es/agenda2030/conoce_la_agenda.htm

Naciones Unidas. (s.f.). *Objetivos y metas de desarrollo sostenible.* https://www.un.org/sustainabledevelopment/es/sustainable-development-goals/

Naturgy. (2022). *Informe de Sostenibilidad y Estado de Información no Financiera 2021.* [Archivo PDF]. https://stpropwebcorporativangy.blob.core.windows.net/uploads/2023/02/INFORME_SOSTENIBILIDAD_ESP-63f393a094b92.pdf

Organización de las Naciones Unidas. (1972). *Conferencia de las Naciones Unidas sobre el Medio Humano.* https://www.un.org/es/conferences/environment/stockholm1972

Organización de las Naciones Unidas. (1992). *Agenda 21.* https://www.un.org/esa/sustdev/documents/agenda21/spanish/a21_summary_spanish.pdf

Organización de las Naciones Unidas. (1992). *Conferencia de las Naciones Unidas sobre Medio Ambiente y Desarrollo, Río de Janeiro, Brasil, 3 a 14 de junio de 1992.* https://www.un.org/es/conferences/environment/rio1992

Organización de las Naciones Unidas. (2015). *Acuerdo de París.* https://www.un.org/es/climatechange/paris-agreement

Organización de las Naciones Unidas. (2015). *Transformar nuestro mundo: la Agenda 2030 para el Desarrollo Sostenible.* https://www.un.org/sustainabledevelopment/es/development-agenda/

Pacto Mundial de las Naciones Unidas (s.f.). *Los Diez Principios.* https://www.pactomundial.org/que-puedes-hacer-tu/diez-principios/

Pacto Mundial de las Naciones Unidas. (2022). *Comunicando el progreso 2022: Renovando las reglas del reporte empresarial*. [Archivo PDF]. https://divem.accem.es/wp-content/uploads/2020/05/Comunicando_el_Progreso_2022_Pacto_Mundial_de_la_ONU_ESP.pdf

Purvis, B., Mao, Y. y Robinson, D. (2019). Three pillars of sustainability: in search of conceptual origins. *Sustain Sci*. 14, 681–695.

Ramsey, J.L. (2015), On Not Defining Sustainability. *J Agric Environ Ethics*. 28, 1075–1087.

Red Eléctrica. (2022). *Informe de Sostenibilidad 2021.* [Archivo PDF]. https://www.redeia.com/sites/webgrupo/files/publication/2022/06/downloadable/Informe_Sostenibilidad_2021.pdf

Repsol. (2022). *Informe de Gestión Integrado.* [Archivo PDF]. https://www.repsol.com/content/dam/repsol-corporate/es/accionistas-e-inversores/informes-anuales/2021/informe-gestion-integrado-2021.pdf

Sacyr. (2022). *Informe Integrado de Sostenibilidad 2021.* [Archivo PDF]. https://www.sacyr.com/documents/63048160/198771338/informe_integrado_2021-web.pdf/e411fd86-f1d6-5f30-76b5-7dff814fb48c?1.3

Silge, J. y Robinson, D. (6 de abril de 2021). *Text Mining with R. A tidy approach.* https://www.tidytextmining.com/index.html.

Solaria. (2022). *Informe de Sostenibilidad 2021.* [Archivo PDF]. https://solariaenergia.com/wp-content/uploads/Memoria-de-sostenibilidad-2021-1.pdf

Sustainability Accounting Standards Boards (s.f.). *About SASB.* https://www.sasb.org/about/

Telefónica. (2022). *Estado de Información no Financiera 2021.* [Archivo PDF]. https://www.telefonica.com/es/wp-content/uploads/sites/4/2022/03/liderar-ejemplo-2021.pdf

Telefónica Tech. (s.f.). *AI of Things: Datos no-estructurados.* https://aiofthings.telefonicatech.com/recursos/datapedia/datos-no-estructurados

Jones, T. W. (15 de noviembre de 2019). *textmineR. 3. Topic Modeling.* https://www.rtextminer.com/articles/c_topic_modeling.html

Transcendent. (2022). *Evolución de los objetivos medioambientales y sociales en las empresas cotizadas*. [Archivo PDF]. https://transcendent.es/wp-content/uploads/2022/12/Evolucion_de_los_objetivos_medioambientales_y_sociales_en_las_empresas_cotizadas.pdf

Unicaja. (2022). *Estado de Información no Financiera Consolidado.* [Archivo PDF]. https://www.unicajabanco.com/content/dam/unicaja/unicaja-corporacion/documentos-corporacion/rsc/Informe_RSC_2021.pdf

United Nations Global Compact. (s.f.). *Social Sustainability.* https://unglobalcompact.org/what-is-gc/our-work/social

United Nations Statistics Division. (2021). *Marco de indicadores mundiales para los Objetivos de Desarrollo Sostenible y metas de la Agenda 2030 para el Desarrollo Sostenible.* https://unstats.un.org/sdgs/indicators/Global%20Indicator%20Framework%20after%202020%20review_Spa.pdf

9. Anexos

9.1. **Anexo 1**: Empresas del IBEX 35 incluidas en el análisis

Ticker	*Compañía*	*Sector*
ACS	ACS	Mat. Básicos, Industria y Construcción
ACX	Acerinox	Mat. Básicos, Industria y Construcción
AENA	AENA	Servicios de Consumo
AMS	Amadeus	Tecnología y Telecomunicaciones
ANA	Acciona	Mat. Básicos, Industria y Construcción
ANE	Acciona Energía	Petróleo y Energía
BBVA	BBVA	Servicios Financieros
BKT	Bankinter	Servicios Financieros
CABK	Caixabank	Servicios Financieros
CLNX	Cellnex	Tecnología y Telecomunicaciones
COL	Inmoviliaria Colonial	Servicios Inmobiliarios
ELE	Endesa	Petróleo y Energía
ENG	Enagás	Petróleo y Energía
FDR	Fluidra	Mat. Básicos, Industria y Construcción
FER	Ferrovial	Mat. Básicos, Industria y Construcción
GRF	Grifols	Bienes de Consumo
IAG	IAG	Servicios de Consumo
IBE	Iberdrola	Petróleo y Energía
IDR	Indra	Tecnología y Telecomunicaciones
ITX	Inditex	Bienes de Consumo
LOG	Logista	Servicios de Consumo
MAP	Mapfre	Servicios Financieros
MEL	Meliá Hotels	Servicios de Consumo
MRL	Merlin Properties	Servicios Inmobiliarios
MTS	ArcelorMittal	Mat. Básicos, Industria y Construcción
NTGY	Naturgy	Petróleo y Energía
RED	Red Electrica	Petróleo y Energía
REP	Repsol	Petróleo y Energía
ROVI	Laboratorios Rovi	Bienes de Consumo
SAB	Banco Sabadell	Servicios Financieros
SAN	Santander	Servicios Financieros
SCRY	Sacyr	Mat. Básicos, Industria y Construcción
SLR	Solaria	Petróleo y Energía
TEF	Telefónica	Tecnología y Telecomunicaciones
UNI	Unicaja	Servicios Financieros

Fuente: Elaboración propia mediante salida de Rstudio

9.2. Anexo 2: Código empleado en Rstudio

```
## Empezamos borrando el entorno de trabajo para no arrastrar nada a nuestro análisis
rm(list = ls())

## Carga de librerías ya instaladas
library(rstudioapi)
library(quanteda)
library (readtext)
library(ggplot2)
library(gridExtra)
library(dyplr)
library(textmineR)
library(topicmodels)
library(ldatuning)
library(tidyverse)

## Establecer el entorno de trabajo
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))
getwd()

## Establecer las rutas en las que se va a basar casi todo el análisis
EINF_folder_path <- paste(getwd(),'/EINFs OG', sep='')
EINFs_list <- list.files(EINF_folder_path)

                    ##### OBTENCIÓN DEL CORPUS #####
## Creación de un data frame que contiene nombre, ticker y tipo de report utilizado para la generación del corpus
Company <- c('ACS', 'Acerinox','AENA', 'Amadeus', 'Acciona', 'Acciona Energía', 'BBVA', 'Bankinter', 'Caixabank', 'Cellnex', 'Inmoviliaria Colonial', 'Endesa', 'Enagás', 'Fluidra', 'Ferrovial', 'Grifols', 'IAG', 'Iberdrola', 'Indra', 'Inditex', 'Logista', 'Mapfre', 'Meliá Hotels', 'Merlin Properties', 'ArcelorMittal', 'Naturgy', 'Red Electrica', 'Repsol', 'Laboratorios Rovi', 'Banco Sabadell', 'Santander', 'Sacyr', 'Solaria', 'Telefónica', 'Unicaja')
```

```
Tickers <- sub("_.*", "", EINFs_list)
Sector <- c('Mat. Básicos, Industria y Construcción','Mat. Básicos, Industria y Construcción', 'Servicios de Consumo', 'Tecnología y Telecomunicaciones','Mat. Básicos, Industria y Construcción', 'Petróleo y Energía', 'Servicios Financieros', 'Servicios Financieros', 'Servicios Financieros', 'Tecnología y Telecomunicaciones', 'Servicios Inmobiliarios', 'Petróleo y Energía', 'Petróleo y Energía','Mat. Básicos, Industria y Construcción','Mat. Básicos, Industria y Construcción', 'Bienes de Consumo', 'Servicios de Consumo',  'Petróleo y Energía', 'Tecnología y Telecomunicaciones', 'Bienes de Consumo', 'Servicios de Consumo', 'Servicios Financieros', 'Servicios de Consumo','Servicios Inmobiliarios', 'Mat. Básicos, Industria y Construcción', 'Petróleo y Energía', 'Petróleo y Energía', 'Petróleo y Energía', 'Bienes de Consumo', 'Servicios Financieros', 'Servicios Financieros',  'Mat. Básicos, Industria y Construcción', 'Petróleo y Energía', 'Tecnología y Telecomunicaciones','Servicios Financieros' )

## Para uniformar los tipos de reports obtenidos, se van a categorizar en Informes Anuales Integrados (IAI) y en Estado de Información no Financiera (EINF)
Report <- sub(".*_", "", EINFs_list)
Report <- sub(".pdf", "", Report)
for (i in 1:length(Report)){
  if(Report[i]=='MSOS' | Report[i]=='ISOS'){
    Report[i]='EINF'
  } else if (Report[i]!='EINF'){
    Report[i]='IAI'
  }}

## Data frame con información sobre compañía, ticker, sector y tipo de report que se incluye en el análisis
info <- cbind(Tickers, Company, Sector, Report)

## Imprimimos un gráfico que contenga la distribución de formatos de INF
inf <- data.frame(info)
```

```
barplot(table(inf$Report), width= 0.5,space=0.1,main = 'Formato de presentación de INF', xlab = "Tipo de Informe", ylab = "Frecuencia",
        col=c('cornflowerblue', 'brown1'), ylim=c(0, max((table(inf$Report))*1.2)))

## Generación del objeto que contiene los reports correspondientes a sostenibilidad
EINFs <- c()
for (i in 1:length(EINFs_list)){
  doc<-readtext(paste(EINF_folder_path,EINFs_list[i], sep='/'))
  EINFs <- append(EINFs, doc$text)
  print(EINFs_list[i])
}

## Creación del corpus e incorporación de variables compañía y tipo de report
Corpus_EINFs <-corpus(tolower(EINFs))
sum <- summary(Corpus_EINFs) # resumen del corpus
sum$Text <- Tickers
CorpusSummary<- sum
CorpusSummary


                          ##### LIMPIEZA DEL CORPUS #####
tokens_EINFs <- quanteda::tokens(Corpus_EINFs, remove_punct = TRUE,
                                 remove_symbols = TRUE,
                                 remove_numbers = TRUE,
                                 remove_url = TRUE,
                                 remove_separators = TRUE)

tokens_EINFs<-tokens_remove(tokens_EINFs, pattern = "\\p{P}+") ## elimina los símbolos de puntuación
tokens_EINFs<-tokens_remove(tokens_EINFs, pattern = tokens_keep(tokens_EINFs,pattern = "\\b[0-9]+\\b")) ## elimina los números
tokens_EINFs<-tokens_remove(tokens_EINFs, pattern = "www.*") ## elimina links
```

```
tokens_EINFs<-tokens_remove(tokens_EINFs, pattern = letters) ## elimina las letras aisladas
tokens_EINFs<-tokens_remove(tokens_EINFs, pattern = stopwords("spanish"))

## Hacemos una eliminación de palabras que probablemente se repitan a lo largo de los corpus debido a la naturaleza de la temática o palabras que probablemente se repitan, como los meses del año

## Lista de ciudades donde operan las empresas del IBEX35 de forma mayoritaria
ciudades <- c("madrid", "barcelona", "valencia", "sevilla", "bilbao", "zaragoza", "málaga", "murcia", "palma de mallorca", "valladolid", "córdoba", "alicante", "vigo", "gijón", "hospitalet de llobregat", "a coruña", "vitoria-gasteiz", "granada","elche", "oviedo", 'españa')
## Lista de países donde operan las empresas del IBEX35 de forma mayoritaria
paises <- c("España", "Portugal", "Reino Unido", "Estados Unidos", "México", "Brasil", "Chile", "Colombia", "Perú", "Argentina", "Francia", "Alemania","Italia", "Países Bajos", "Suiza", "Suecia", "Noruega", "China", "India", "Japón", "Australia", "Canadá", "Singapur", "Emiratos Árabes Unidos", "Arabia Saudita")
## Lista de palabras comunes en relación con el mundo empresarial, que posiblemente hayan sido mencionadas en los IAI
business_gloss <- c( "grupo","consolidado", "integrado", "negocio", "consejo", "administración", "informe", "gestión", "resultados", "estrategia","objetivos", "indicadores", "progreso", "desempeño", "evaluación", "acciones", "implementación", "eficiencia", "efectividad", "mejora", "logros", "metas", "anual", "anexo", "total", "anexos", "cuentas", 'anuales', 'ex', 'nd', "planificación", "seguimiento","equipo", "colaboración", "liderazgo", "consejeros","organización", "plan", "implementar", "evaluar", "mejorar", "optimizar","alcanzar", "analizar", "estrategias", "auditoría","objetivos", "metodología", "información", "rendimiento", "reporte", "iniciativa", "efectividad", "eficacia", "pág","productividad", "rendición", "medición", "eficiencia", "control", "mejoras", 'millones',"s.a", 'euros', 'así','gri', 'ee.uu', 'melia')
company_gloss <- c("Ingeniería","Construcción","Infraestructuras", "Energías renovables", "Proyectos", "Acero inoxidable", "Metalurgia","Industria
```

```
siderúrgica","Productos laminados", "Exportación","Aeropuertos","Gestión aeroportuaria","Transporte aéreo","Conectividad","Sistemas de reserva","Tecnología para viajes", "Gestión de agua", "Energía limpia", "Banca", "Inversiones", "Créditos", "Gestión de activos", "Banca privada", "Banca digital", "Infraestructuras de telecomunicaciones", "Torres de comunicación", "Operador de infraestructuras", "Conectividad", "Antenas", "Bienes raíces", "Inversión inmobiliaria", "Propiedades comerciales", "Oficinas y locales", "Hoteles","Turismo", 'piscinas')

## Eliminamos las 254 palabras del corpus de tokens
words_remove <- tolower(c(Company,Tickers, Sector, ciudades, paises, business_gloss, company_gloss))
tokens_EINFs <- tokens_remove(tokens_EINFs, pattern = tolower(c(quanteda::tokens(words_remove))))
summary(tokens_EINFs)

## Una vez hecho la primera limpieza del corpus, vamos a pasar a construir la matriz dtm, la cual va a ayudar en este filtrado de palabras y limpieza de la base de datos

##### OBTENCIÓN MATRIZ DTM #####
## Construcción de la matriz dtm en base al corpus que hemos obtenido con los reports del IBEX 35
dtm_EINFs <- dfm(tokens_EINFs)
dim(dtm_EINFs) ## la matriz presenta unas dimensiones de 35x59531 -> reponde a los 35 informes y a casi 60000 tokens identificados
topfeatures(dtm_EINFs, 100, decreasing=TRUE) ## palabras más frecuentes
topfeatures(dtm_EINFs, 100, decreasing=FALSE) ## palabras menos fecuentes
topfeatEINF <- topfeatures(dtm_EINFs, 10, groups = docnames(dtm_EINFs)) # términos más frecuentes por documento (revisar como se presenta el orden)

## Seguimos eliminando términos
dtm_EINFs1<-dfm_remove(dtm_EINFs, pattern = c("[0-9]+(?:st| st|nd| nd|rd| rd|th| th|s)","\\b[a-zA-Z]\\b" ), valuetype = "regex")
```

```
dtm_EINFs1<-dfm_remove(dtm_EINFs1, pattern = '*-*') # eliminamos todas aquellas palabras que contengan guiones
dim(dtm_EINFs1) # se reduce a 50740 términos

## Utilizamos steeming
dtm_EINFs2 <- dfm_wordstem(dtm_EINFs1, language = "spanish")
dim(dtm_EINFs2) # se reduce a 28089 términos (casi la mitad de las features que teniamos originalmente)

## Eliminamos todas aquellas palabras que no se hayan mencionado más de 35 veces a lo largo del corpus
dtm1 <- dfm_trim(dtm_EINFs2, min_termfreq = 35)
dim(dtm1) # Obtenemos un total de 4745 palabras resultantes
topfeatures(dtm1, 100, decreasing=TRUE) ## palabras más frecuentes
topfeatures(dtm1, 100, decreasing=FALSE) ## palabras menos frecuentes

                                ##### ANÁLISIS EXPLORATORIO #####
## Visualizamos cuales son los top15 términos más frecuentes en todo el corpus tras la realización de un filtrado más amplio
c <- topfeatures(dtm1, 15, decreasing = TRUE)
b <- data.frame(nam = names(c), freq = topfeatures(dtm1, 15, decreasing = TRUE))
ggplot(b, aes(x = reorder(nam, +freq), y = freq )) + geom_bar(stat = "identity", fill="darkseagreen3")+ xlab("") +coord_flip() +
  theme(legend.position = "none") +  labs(title = "Términos más frecuentes en la INF del IBEX35 ", y='Frecuencia de cada término') +
  theme(plot.title = element_text(size= 16, hjust = 0.5, margin = margin(b = 16))) +
  theme(axis.text.y = element_text(size = 14), axis.text.x = element_text(size = 12))


## Introducimos el concepto de ngramas para ver que combinaciones de dos y tres palabras son las más repetidas a lo largo del corpus
ngramas2 <- quanteda::tokens_ngrams(tokens_EINFs, n = 2, concatenator = ' ')
```

```
dtm_ng2 <- dfm(ngramas2)
dim(dtm_ng2)
topfeatures(dtm_ng2, 100, decreasing=TRUE) ## palabras más frecuentes

dtm_ng2<-dfm_remove(dtm_ng2, pattern = c("[0-9]+(?:st| st|nd| nd|rd| rd|th| th|s)","\\b[a-zA-Z]\\b" ), valuetype = "regex")
dtm_ng2<-dfm_remove(dtm_ng2, pattern = '*-*') # eliminamos todas aquellas palabras que contengan guiones
dim(dtm_ng2)

#Eliminamos todas aquellas palabras que no se hayan mencionado más de 30 veces a lo largo del corpus
dtmng2 <- dfm_trim(dtm_ng2, min_termfreq = 100)
topfeatures(dtm_ng2, 100, decreasing=TRUE)
ngelim <- c('piscina wellness','largo plazo', 'años años', 'centros comerciales', 'melia international', 'conoce gobernanza', 'mejores prácticas', 'colaboramos generar', 'consejero delegado', 'hombres mujeres', 'retribución variable')
dtmng2<-dfm_remove(dtmng2, pattern = ngelim, valuetype = "regex")
dim(dtmng2) # Obtenemos un total de 620 bigramas resultantes

## Visualizamos cuales son los top15 términos más frecuentes en todo el corpus tras la realización de un filtrado más amplio
c <- topfeatures(dtmng2, 16, decreasing = TRUE)
c[['seguridad salud']] <- c[['seguridad salud']] + c[['salud seguridad']] #juntamos sus frecuencias debido a que tienen el mismo significado
c <- subset(c, !names(c)=='salud seguridad')
b <- data.frame(nam = names(c), freq = c)
ggplot(b, aes(x = reorder(nam, +freq), y = freq, fill = freq)) + geom_bar(stat = "identity", fill="lightsteelblue3")+ xlab("") +coord_flip() +
  theme(legend.position = "none") +  labs(title = "Bigramas más frecuentes en la INF del IBEX35 ",  y='Frecuencia de cada bigrama') +
  theme(plot.title = element_text(size= 16, hjust = 0.5, margin = margin(b = 16))) +
```

```
  theme(axis.text.y = element_text(size = 14), axis.text.x = element_text(size = 12))
```

Para finalizar el análisis exploratorio, vamos a ver cómo cambia la frecuencia de términos más repetidos por sector identificado. Recogemos la variable de sector presentada previamente, para establecer la subdivisión de dtms a construir ahora

```
sectores <- unique(Sector) # obtenemos que hay 7 sectores por lo que obtendremos los n-gramas más repetidos por sector
a1 <- list()
a2 <- list()

for (i in 1:length(sectores)){
  lista <- which(Sector == sectores[i])
  nombres_einfs <- dtmng2@docvars$docname_[lista]
  dtms1 <- dfm_subset(dtm1, dtm1@docvars$docname_ %in% nombres_einfs )
  dtms2 <- dfm_subset(dtmng2, dtmng2@docvars$docname_ %in% nombres_einfs )
  a1 <- append(a1, topfeatures(dtms1, 5, decreasing=TRUE))
  a2 <- append(a2, topfeatures(dtms2, 5, decreasing=TRUE))
}
topgrams <- split(a1, rep(1:7, each = 5, length.out = 35))
topngrams <- split(a2, rep(1:7, each = 5, length.out = 35))
ptable1 <- matrix(nrow=5, ncol=7)
ptable2 <- matrix(nrow=5, ncol=7)
for (i in 1:length(topngrams)) {
  ptable1[, i] <-names(topgrams[[i]])
  ptable2[, i] <-names(topngrams[[i]])
}
ptable1 <- data.frame(ptable1)
colnames(ptable1) <- sectores
ptable2 <- data.frame(ptable2)
colnames(ptable2) <- sectores
```

OBTENCIÓN DEL MODELO LDA

```
## Eliminamos palabras con frecuencias mayores a 9000 y cambiamos el tipo de objeto
dtm1 <- dfm_trim(dtm1, max_termfreq=9000)
topfeatures(dtm1, 100, decreasing=TRUE)
dtm1lda <- as(dtm1, "dgCMatrix")
colnames(dtm1lda)
dim(dtm1lda)

## Buscamos el valor óptimo de K
result <- FindTopicsNumber(
  dtm1lda,
  topics = seq(from = 20, to = 100, by = 10),
  metrics = c("CaoJuan2009", "Arun2010", "Deveaud2014"),
  method = "VEM",
  control = list(seed = 123),
  mc.cores = 2L,
  verbose = TRUE)
FindTopicsNumber_plot(result)

## Obtenemos el contenido de los temas en función del valor de K obtenido
lda_mod <- LDA(dtm1lda,
               k = 70,
               control = list(seed=123))
terms(lda_mod)
trms <- t(terms(lda_mod, k=7))

## Obtener el resultado y probabilidades de los temas
tmResult <- posterior(lda_mod)
theta <- tmResult$topics
topicProportions <- colSums(theta)



## Obtenemos los temas con mayor theta par posteriormente visualizarlos
```

```
a <- sort(topicProportions, decreasing = TRUE)
positions <- names(a)
lda.terms <- as.data.frame(topicmodels::terms(lda_mod, 7), stringsAsFactors = FALSE)
lda.main.tops <- lda.terms[,as.integer(positions)]
lda.main.tops <- lda.main.tops[,(1:5)]
prob.top <- tidytext::tidy(lda_mod, matrix='beta')
prob.term <- prob.top %>%
  group_by(topic) %>%
  slice_max(beta, n = 7) %>%
  ungroup() %>%
  arrange(topic, -beta)

## Obtenemos los top temas y los visualizamos a continuación
top5 <- as.integer(positions[1:5])
mytop5 <- rbind(prob.term[as.integer(prob.term$topic)==top5[1],],
prob.term[as.integer(prob.term$topic)==top5[2],],
prob.term[as.integer(prob.term$topic)==top5[3],],
prob.term[as.integer(prob.term$topic)==top5[4],],
prob.term[as.integer(prob.term$topic)==top5[5],])
mytop5$topic <- c(4,4,4,4,4,4,4,3,3,3,3,3,3,3,1,1,1,1,1,1,1,2,2,2,2,2,2,2,5,5,5,5,5,5,5,5)
mytop5 %>%
  mutate(term = tidytext::reorder_within(term, beta, topic)) %>%
  ggplot(aes(beta, term, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  tidytext::scale_y_reordered() +
  scale_fill_manual(values = c("lightcoral", "seagreen3", "deepskyblue3", "sandybrown",
"slateblue1")) +
  theme(axis.text.y = element_text(size = 14),
        axis.title = element_text(size = 14))
```