Alberto Flores Pastor

ANÁLISIS DE SIMILITUD SEMÁNTICA DE TWEETS SOBRE SOSTENIBILIDAD UTILIZANDO MODELOS BERT PRE-ENTRENADOS Y TÉCNICAS DE GENERACIÓN DE EMBEDDINGS

TRABAJO FIN DE GRADO

Dirigido por Dr. Antonio Moreno

Doble Grado Biotecnología e Ingeniería Informática

UNIVERSITAT ROVIRA I VIRGILI

Tarragona

2023

RESUMEN

El proyecto se enfoca en el análisis exhaustivo de un conjunto específico de tweets recopilados durante el período comprendido entre 2018 y principios de 2022, los cuales provienen de instituciones y oficinas turísticas que utilizan el idioma español. El propósito principal consiste en identificar de manera precisa aquellos tweets que abordan la temática de la sostenibilidad, y agruparlos en función del subtema que traten.

Para llevar a cabo este análisis, se empleará el modelo de lenguaje BERT. Como parte del proceso, se seleccionarán frases de referencia específicas relacionadas con la sostenibilidad, las cuales servirán como criterios fundamentales para evaluar si los tweets en cuestión abordan o no esta temática, y permitirán clasificarlos en los diferentes subtemas.

Una vez que se haya realizado la recopilación de los tweets, se procederá a calcular los embeddings de los tweets y de las frases de referencia con BERT. Esto permitirá evaluar la similitud entre los tweets y las frases de referencia, determinando si los tweets abordan la temática de la sostenibilidad. Asimismo, esta comparación también permitirá asignar a cada tweet el subtema al que se asemeja con mayor similitud.

Posteriormente, se llevará a cabo un análisis exhaustivo para evaluar la precisión y exhaustividad de los resultados obtenidos. Además, se investigará la frecuencia con la que se menciona la sostenibilidad en los tweets analizados, y se identificarán los subtemas más relevantes tanto a nivel de comunidad autónoma como en el conjunto total de tweets recopilados.

RESUM

El projecte s'enfoca a l'anàlisi exhaustiva d'un conjunt específic de tweets recopilats durant el període comprès entre el 2018 i el principi del 2022, els quals provenen d'institucions i oficines turístiques que utilitzen l'idioma espanyol. El propòsit principal consisteix a identificar de manera precisa aquells tweets que aborden la temàtica de la sostenibilitat, i agrupar-los en funció del subtema que tractin.

Per dur a terme aquesta anàlisi, es farà servir el model de llenguatge BERT. Com a part del procés, se seleccionaran frases de referència específiques relacionades amb la sostenibilitat, les quals serviran com a criteris fonamentals per avaluar si els tweets en qüestió aborden o no aquesta temàtica, i permetran classificar-los en els diferents subtemes.

Un cop feta la recopilació dels tweets, es procedirà a calcular els embeddings dels tweets i de les frases de referència amb BERT. Això permetrà avaluar la similitud entre els tweets i les frases de referència, determinant si els tweets aborden la temàtica de la sostenibilitat. Així mateix, aquesta comparació també permetrà assignar a cada tweet el subtema a què s'assembla amb més similitud.

Posteriorment, es durà a terme una anàlisi exhaustiva per avaluar la precisió i exhaustivitat dels resultats obtinguts. A més, s'investigarà la freqüència amb què s'esmenta la sostenibilitat als tweets analitzats, i s'identificaran els subtemes més rellevants tant a nivell de comunitat autònoma com al conjunt total de tweets recopilats.

ABSTRACT

The project focuses on the exhaustive analysis of a specific set of tweets collected during the period between 2018 and the beginning of 2022, which come from institutions and tourist offices that use the Spanish language. The main purpose is to accurately identify those tweets that address the topic of sustainability, and group them based on the subtopic they address.

To carry out this analysis, the BERT language model will be used. As part of the process, specific reference phrases related to sustainability will be selected, which will serve as fundamental criteria to evaluate whether or not the tweets address the theme of sustainability and will allow them to be classified into the different sub-themes.

Once the tweets have been collected, the embeddings of the tweets and the reference phrases will be calculated with BERT. This will allow evaluating the similarity between the tweets and the reference phrases, determining if the tweets address the theme of sustainability. This comparison will also allow us to assign each tweet to the sub-theme to which it most closely resembles.

Subsequently, a comprehensive analysis will be carried out to assess the precision and completeness of the results obtained. In addition, the frequency with which sustainability is mentioned in the analyzed tweets will be investigated, and the most relevant sub-themes will be identified both at the autonomous community level and in the total set of tweets collected.

ÍNDICE

ÍNDICE TABLAS

ÍNDICE FIGURAS

1 INTRODUCCIÓN

En este capítulo introductorio, se aborda el tema de la interacción entre las redes sociales y el turismo, con un enfoque específico en la sostenibilidad. Se explorará cómo las redes sociales han influido en la difusión de información y la promoción de destinos turísticos, y cómo se abordan los aspectos de sostenibilidad en esta plataforma.

Además, se hablará sobre la motivación y objetivos de este proyecto, los cuales radican en la necesidad de comprender y evaluar la comunicación de la sostenibilidad en el turismo a través de las redes sociales. Finalmente, se explicará la estructura que se ha seguido para desarrollar este documento.

1.1 Interacción entre las redes sociales, el turismo y la sostenibilidad

1.1.1 Las redes sociales y el turismo

En la actualidad, el turismo se ha convertido en uno de los sectores económicos más importantes a nivel mundial, y España destaca como uno de los destinos turísticos más populares. En este contexto, las TIC[1] han desempeñado un importante papel en diversas industrias, y el sector turístico no ha sido una excepción.

La evolución de la tecnología, especialmente con la aparición de internet, ha revolucionado la industria del turismo al brindar nuevas oportunidades y desafíos. Las páginas web se han convertido en una herramienta fundamental para la promoción y venta de servicios turísticos. Los blogs especializados permiten a los viajeros compartir sus experiencias y recomendaciones, creando una comunidad virtual donde se intercambian información valiosa sobre destinos, alojamientos, actividades y más (Valeri & Baggio, 2021).

Sin embargo, uno de los mayores cambios que ha experimentado el sector turístico con la llegada de la tecnología es la aparición de la Web 2.0 o Web Social. Esta nueva generación de internet ha permitido a los usuarios no solo consumir información, sino también generar contenido y participar activamente en la creación de comunidades en línea.

Plataformas como TripAdvisor, Facebook y Twitter han desempeñado un papel fundamental en esta transformación. TripAdvisor, por ejemplo, ha democratizado la información turística al permitir que los viajeros compartan sus opiniones y reseñas sobre hoteles, restaurantes y atracciones. Esto ha brindado a los turistas acceso a una amplia gama de comentarios y recomendaciones de otros viajeros, lo que les ayuda a tomar decisiones informadas (Compagnone & Fiorentino, 2018; Xiang, 2018).

Facebook y Twitter han facilitado la interacción entre los turistas y las empresas turísticas, permitiendo una comunicación directa y en tiempo real. Los viajeros pueden seguir las páginas de destinos, hoteles o aerolíneas en estas redes sociales, recibir actualizaciones sobre promociones, eventos o noticias relevantes, e incluso interactuar directamente con las marcas a través de comentarios o mensajes privados. En este sentido, las redes sociales han demostrado ser un medio crucial para el sector turístico, ya que permiten a las organizaciones establecer una mayor conexión con los usuarios, generar compromiso y crear una relación emocional con sus marcas. Los turistas suelen utilizar estas plataformas para compartir sus experiencias de

[1]Tecnologías de la Información y las Comunicaciones

viaje, buscar información y leer opiniones de otros viajeros, lo que las convierte en una fuente valiosa de referencia (Sánchez Jimenéz, 2018).

El avance de las Tecnologías de la Información ha impulsado el crecimiento del turismo independiente, donde los viajeros eligen organizar y gestionar sus propios viajes en lugar de depender de agencias de viajes. Esto ha sido posible gracias a la accesibilidad a la información que proporcionan las tecnologías, permitiendo a los turistas encontrar información, reservar y pagar por servicios en línea. Los viajeros buscan tanto información oficial proporcionada por las empresas, como información generada por otros turistas en las redes sociales, donde comparten opiniones, experiencias y recomendaciones (Ortiz & González Sánchez, 2014).

En los últimos años, las TIC han tenido un impacto significativo en el sector turístico, especialmente en las funciones de marketing y distribución. Estas tecnologías han permitido la personalización en masa y el acceso a nichos de mercado en diferentes áreas geográficas, lo que ha llevado a las empresas turísticas a integrar estas herramientas en sus estrategias de marketing (Mariani et al., 2016). La competitividad en el mercado turístico se ha vuelto más exigente debido a la constante innovación de hardware, software y el desarrollo de las redes, lo que requiere que las organizaciones turísticas sean dinámicas y se mantengan actualizadas (Hernández-Méndez et al., 2016; Stokłosa et al., 2019).

En este sentido, las redes sociales desempeñan un papel fundamental al brindar a las empresas turísticas una plataforma para promocionar sus productos y servicios de manera dinámica y atractiva. A través de la creación de contenido visual, como imágenes y videos, y la interacción con los usuarios, las empresas pueden generar un mayor interés y compromiso por parte de los turistas. Las redes sociales permiten una comunicación directa y en tiempo real, lo que facilita la atención al cliente, la respuesta a consultas y la resolución de problemas de manera rápida y efectiva.

Además, las redes sociales ofrecen a las empresas turísticas la oportunidad de aprovechar el poder del marketing de influencia. Los influencers, con un gran número de seguidores y credibilidad en línea, pueden promocionar destinos, hoteles, actividades y servicios turísticos a través de contenido auténtico y persuasivo. Esta colaboración con influencers puede aumentar la visibilidad de la marca, llegar a nuevos públicos y generar confianza en los usuarios (Guerreiro et al., 2019).

1.1.2 Sostenibilidad

En los últimos años, ha surgido una creciente preocupación por los impactos negativos que el turismo puede tener en el medio ambiente, la sociedad y la cultura de los destinos visitados. El aumento en el flujo de turistas y el desarrollo turístico descontrolado han llevado a la sobreexplotación de recursos naturales, la degradación del entorno, la pérdida de biodiversidad y la alteración de los ecosistemas. Además, la llegada masiva de turistas puede generar tensiones sociales, como el aumento en el costo de vida para los residentes locales, la desculturización y la pérdida de identidad cultural.

En respuesta a esta problemática, el turismo sostenible ha emergido como una prioridad para los destinos turísticos españoles. El turismo sostenible se refiere a la práctica de planificar, desarrollar y gestionar el turismo de manera que se maximicen los beneficios económicos, minimizando al mismo tiempo los impactos negativos en el medio ambiente y en las comunidades locales. El objetivo es lograr un equilibrio entre el crecimiento económico, la

protección del medio ambiente y la preservación de la cultura y el patrimonio (Ruhanen et al., 2019).

En términos ambientales, el turismo sostenible busca minimizar la huella ecológica de las actividades turísticas. Esto implica la adopción de prácticas de gestión ambiental que promuevan la conservación de los recursos naturales, la protección de los ecosistemas frágiles, la reducción de la contaminación y la eficiencia en el uso de energía y agua. También se fomenta el turismo de bajo impacto, que se centra en la apreciación y la conservación de la biodiversidad y los paisajes naturales.

En el ámbito social, el turismo sostenible busca generar beneficios para las comunidades locales, promoviendo la participación activa de los residentes en el desarrollo turístico y asegurando que compartan equitativamente los beneficios económicos. Se busca también preservar la identidad cultural de las comunidades, respetando sus tradiciones, valores y prácticas. Esto implica fomentar el turismo responsable, en el cual los visitantes interactúen de manera respetuosa y se promueva el intercambio cultural enriquecedor (Y. Guo et al., 2019).

Además de los aspectos ambientales y sociales, el turismo sostenible también aborda la dimensión económica del desarrollo turístico. Se busca maximizar los beneficios económicos para las comunidades locales a través de la creación de empleo, la generación de ingresos y la promoción de oportunidades empresariales sostenibles. Esto implica fomentar la participación de las empresas locales en la cadena de valor del turismo, promoviendo la compra de productos y servicios locales y apoyando la economía local (Pan et al., 2018).

Un aspecto clave del turismo sostenible es la planificación y gestión adecuada de los destinos turísticos. Esto implica la identificación de áreas de desarrollo turístico adecuadas y la implementación de políticas y regulaciones efectivas para garantizar un desarrollo ordenado y sostenible. La participación activa de los actores locales, incluidas las comunidades, las autoridades locales y las empresas turísticas, es fundamental en este proceso. La colaboración y la cooperación entre los diferentes actores son necesarias para garantizar la integración de los principios de sostenibilidad en todas las etapas del desarrollo turístico.

Además, el turismo sostenible promueve la educación y la sensibilización tanto entre los turistas como entre la población local. Se busca informar a los turistas sobre las prácticas sostenibles y promover comportamientos responsables durante su visita, como el respeto por el medio ambiente, la cultura y las tradiciones locales. Asimismo, se fomenta la sensibilización de la población local sobre la importancia de la sostenibilidad y se les involucra en la toma de decisiones relacionadas con el turismo (Kapera, 2018).

Es importante destacar que el turismo sostenible no solo beneficia a los destinos turísticos y a las comunidades locales, sino que también ofrece una experiencia enriquecedora para los propios turistas. Viajar de manera sostenible permite a los turistas conectar de manera más auténtica con el entorno natural y cultural, experimentar la hospitalidad local y contribuir positivamente al desarrollo de los destinos visitados. El turismo sostenible ofrece la oportunidad de disfrutar de experiencias significativas y memorables, al tiempo que se asegura la conservación y preservación de los recursos para las generaciones futuras (Etzion, 2018; Higgins-Desbiolles, 2018).

1.1.3 Sostenibilidad en redes sociales

El uso de las redes sociales como herramienta de comunicación en el ámbito de la sostenibilidad turística ha adquirido una relevancia significativa en los últimos años. Estas plataformas digitales ofrecen un espacio dinámico y accesible para difundir mensajes relacionados con prácticas sostenibles, proyectos de conservación y concienciación sobre la importancia de preservar el medio ambiente y las comunidades locales.

Una de las ventajas más destacadas de las redes sociales es su capacidad para llegar a una amplia audiencia, incluyendo turistas, residentes locales, empresas turísticas y otros actores del sector. A través de publicaciones, imágenes y videos compartidos en estas plataformas, se pueden mostrar ejemplos concretos de prácticas sostenibles, experiencias en destinos respetuosos con el entorno y actividades turísticas responsables. Estas visualizaciones ayudan a crear una imagen positiva del destino y a despertar el interés de los turistas comprometidos con la sostenibilidad (Della Corte et al., 2019; MacKenzie & Gannon, 2019).

Asimismo, las redes sociales fomentan la colaboración entre diferentes actores del sector turístico. Las empresas, organizaciones no gubernamentales, instituciones gubernamentales y comunidades locales pueden unir fuerzas a través de estas plataformas para promover iniciativas conjuntas en pro de la sostenibilidad. Se pueden compartir conocimientos, recursos y experiencias, lo que facilita el desarrollo de proyectos de turismo comunitario, la protección del medio ambiente y la implementación de prácticas sostenibles en todas las etapas de la cadena de valor del turismo.

No obstante, el uso de las redes sociales también plantea desafíos en la comunicación de aspectos asociados a la sostenibilidad. La gran cantidad de información disponible en estas plataformas puede dificultar que los turistas encuentren y accedan a contenido relevante sobre sostenibilidad. Además, la veracidad y autenticidad de la información pueden ser cuestionadas, ya que las redes sociales permiten la libre expresión y la generación de contenido por parte de los usuarios (Varshney & Vishwakarma, 2021).

Para superar estos desafíos, es fundamental que los destinos turísticos españoles implementen estrategias de gestión de la reputación en las redes sociales. Esto implica monitorear activamente los comentarios, opiniones y críticas de los usuarios, y responder de manera transparente y constructiva (Gai et al., 2023). La comunicación bidireccional, basada en la confianza y la honestidad, es crucial para establecer una relación sólida con los turistas y generar compromiso con la sostenibilidad.

El uso estratégico de las redes sociales en la promoción de la sostenibilidad turística ha demostrado ser una forma efectiva de generar conciencia y fomentar la acción colectiva. Estas plataformas ofrecen la oportunidad de crear una comunidad en línea de personas interesadas en temas relacionados con la sostenibilidad y el turismo responsable. A través de grupos, hashtags y campañas temáticas, es posible unir a individuos, organizaciones y empresas con objetivos comunes, compartiendo información, ideas y buenas prácticas.

Además, las redes sociales permiten la interacción directa y el diálogo entre los actores del sector turístico y la comunidad en general. Los destinos turísticos pueden utilizar estas plataformas para obtener retroalimentación instantánea de los turistas y los residentes locales, lo que les permite adaptar sus estrategias de sostenibilidad de acuerdo a las necesidades y expectativas de sus públicos (Lian et al., 2020). Asimismo, se pueden organizar encuestas,

debates y foros de discusión en línea, generando una mayor participación y compromiso con las cuestiones relacionadas con la sostenibilidad turística.

La viralización de contenido relacionado con la sostenibilidad en las redes sociales también puede tener un impacto significativo en la conciencia pública. Cuando los mensajes sobre prácticas sostenibles, proyectos de conservación o casos de éxito se vuelven virales, se multiplican las oportunidades de difusión y generación de impacto. Estos contenidos pueden inspirar a otros destinos turísticos a implementar medidas sostenibles, atraer la atención de los medios de comunicación y motivar a los turistas a elegir opciones más responsables durante sus viajes.

Por otro lado, las redes sociales también han facilitado la creación de comunidades de influencers y microinfluencers dedicados a promover la sostenibilidad en el turismo. Estas personas, con una audiencia comprometida, utilizan sus plataformas para compartir consejos, recomendaciones y experiencias relacionadas con prácticas responsables de viaje. Su influencia puede ser poderosa para inspirar cambios de comportamiento y promover la adopción de prácticas más sostenibles entre sus seguidores(Guerreiro et al., 2019).

Sin embargo, es importante destacar que el uso de las redes sociales en la comunicación de la sostenibilidad turística requiere una gestión adecuada y ética. La autenticidad y la transparencia son valores fundamentales para generar confianza y credibilidad en el contenido compartido (Varshney & Vishwakarma, 2021). Las organizaciones turísticas deben asegurarse de verificar la veracidad de la información antes de compartirla y evitar la sobreexplotación o apropiación indebida de temas delicados relacionados con la sostenibilidad.

Por todo ello, se llevará a cabo un estudio completo y riguroso sobre el análisis de la comunicación de aspectos asociados a la sostenibilidad por destinos turísticos españoles a través de redes sociales. A través de este análisis, se busca identificar los temas más debatidos tanto por los ayuntamientos como por las oficinas de turismo sobre la comunicación de sostenibilidad, así como determinar la frecuencia con la que se abordan.

1.2 Motivación y objetivos

La motivación detrás de este proyecto se basa en la creciente importancia de la sostenibilidad en el contexto del turismo y la necesidad de comprender cómo se aborda este tema en las redes sociales en español. En los últimos años, la sostenibilidad se ha convertido en un factor clave en la toma de decisiones de los turistas, así como en la estrategia de desarrollo de destinos turísticos y empresas del sector. Para comprender mejor las percepciones y opiniones de la comunidad en línea sobre la sostenibilidad en el turismo, es fundamental analizar los tweets en español y extraer información valiosa sobre los subtemas más relevantes y las tendencias actuales.

Objetivo global: Evaluar la comunicación de los diferentes aspectos de sostenibilidad a través de las redes sociales por parte de los ayuntamientos u oficinas de turismo de destinos turísticos.

Tareas específicas:

- Recopilar una amplia cantidad de tweets en español relacionados con el turismo y la sostenibilidad: El primer objetivo es obtener una muestra representativa de tweets en español que aborden temas relacionados con el turismo y la sostenibilidad. Esto implicará la recolección de una gran cantidad de tweets a través de diversas fuentes, como Twitter API y otras herramientas de búsqueda y recopilación de datos.
- Utilizar el modelo BERT para identificar tweets relacionados con la sostenibilidad: El siguiente objetivo es aplicar el modelo BERT para analizar los tweets y determinar cuáles de ellos tratan específicamente sobre sostenibilidad en el turismo. BERT es un modelo de aprendizaje automático avanzado que tiene la capacidad de comprender el contexto y el significado de las palabras en un texto, lo que nos permitirá identificar de manera precisa los tweets relevantes para nuestro análisis.
- Realizar un análisis detallado de los subtemas y tendencias sobre sostenibilidad en el turismo: Una vez identificados y clasificados los tweets relevantes, se llevará a cabo un análisis exhaustivo para comprender mejor cómo se aborda la sostenibilidad en las redes sociales en español. Se explorarán los subtemas más recurrentes, las opiniones y tendencias más destacadas, y se identificarán los aspectos clave relacionados con la sostenibilidad en el turismo.

1.3 Estructura del documento

En el capítulo que estamos viendo, se aborda la interacción entre las redes sociales, el turismo y la sostenibilidad. Se explora la relevancia de las redes sociales en el ámbito turístico y se introduce el concepto de sostenibilidad, así como su aplicación en las redes sociales. También se detallan la motivación y los objetivos del proyecto, y se proporciona una visión general de la organización del documento.

El próximo capítulo, titulado "Modelos Avanzados de Tratamiento del Lenguaje Natural", se adentra en los conceptos fundamentales de los modelos de procesamiento del lenguaje natural, como los embeddings y en particular el modelo BERT. Se analizan diferentes variantes de BERT y se realiza una comparación para seleccionar el modelo más adecuado para el proyecto.

Seguidamente, se tratará el capítulo denominado "Diseño del Sistema". En él se detalla el proceso de diseño y desarrollo del sistema utilizado para el análisis de los tweets. Se explican aspectos como la recopilación de los tweets, la selección de frases y conceptos de referencia, así como técnicas como la similitud coseno y la elección del threshold.

El Capítulo 4, "Resultados", presenta los hallazgos obtenidos a partir del análisis de los tweets sobre sostenibilidad. Se abordan dos aspectos principales: la frecuencia de los tweets relacionados con la sostenibilidad y el análisis de subtemas tanto a nivel de cuenta, comunidad autónoma y a nivel global.

En el Capítulo 5, "Conclusiones y Trabajo Futuro", se exponen las conclusiones derivadas del proyecto, destacando los principales resultados y hallazgos. Además, se plantean posibles líneas de trabajo futuro para profundizar en el análisis de la sostenibilidad en el turismo a través de las redes sociales.

El último capítulo, "Autoevaluación", permite realizar una evaluación crítica del trabajo realizado, destacando los conocimientos y habilidades adquiridos a lo largo del proyecto

Finalmente, se incluyen las Referencias bibliográficas y los Anexos, donde se proporcionan detalles adicionales como el listado de cuentas analizadas, los conceptos de referencia utilizados, el código implementado y el análisis de subtemas por cuenta.

En conclusión, las redes sociales han transformado la industria turística al permitir a los turistas compartir opiniones, buscar información y conectarse con las empresas turísticas. Por otro lado, la sostenibilidad es una preocupación creciente en el turismo, buscando minimizar los impactos negativos en el medio ambiente y las comunidades locales. Las redes sociales juegan un papel crucial al promover prácticas sostenibles, proyectos de conservación y colaboración entre los actores del sector turístico para fomentar la sostenibilidad. Ante esta situación, surge la elaboración de este proyecto para tratar de comprender cómo se aborda este tema en las redes sociales en español. Y para tratar este problema se usará el modelo avanzado para el tratamiento del lenguaje natural conocido como BERT.

2. MODELOS AVANZADOS EN EL TRATAMIENTO DEL LENGUAJE NATURAL

Este capítulo, se adentra en los modelos utilizados en el procesamiento del lenguaje natural, centrándose en los embeddings y el modelo BERT[2]. Se explorarán diversas variantes como sBERT, BioBERT, DistilBERT, CamemBERT, BERTO y mBERT, evaluando sus propias características y aplicaciones. Además, se llevará a cabo una comparación exhaustiva de los modelos BERT comentados para seleccionar el más adecuado. Se detallará el modelo elegido y se describirá el dataset utilizado y las bases del modelo escogido.

2.1 Embeddings y BERT

El NLP[3] es un campo de estudio que busca permitir a las máquinas comprender y procesar el lenguaje humano de una manera similar a cómo lo hacen los seres humanos. Esto implica desafíos significativos debido a la naturaleza compleja y ambigua del lenguaje natural. Los modelos de lenguaje tradicionales se basaban en representaciones simples de palabras, como los "one-hot vectors", que no capturaban la riqueza semántica y contextual del lenguaje (Liu et al., 2020).

Aquí es donde entran los emeddings. Los embeddings son representaciones numéricas de palabras o frases en un espacio vectorial continuo. Estas representaciones capturan información semántica y sintáctica de las palabras, lo que permite a los NLP entender y trabajar con el significado de las palabras en lugar de tratarlas como meros símbolos (Almeida & Xexéo, 2019). Los embeddings se generan mediante técnicas de aprendizaje automático, como el aprendizaje no supervisado o el aprendizaje profundo, a partir de grandes cantidades de texto.

Uno de los enfoques más destacados para generar embeddings de palabras es el algoritmo Word2Vec, que utiliza redes neuronales para aprender representaciones vectoriales de palabras en función de su contexto (Adewumi et al., 2022). Estos embeddings capturan similitudes y relaciones entre palabras, lo que permite realizar operaciones semánticas interesantes, como la analogía.

BERT, por otro lado, es un modelo de lenguaje basado en la transformación bidireccional de codificadores. BERT revolucionó el campo del NLP cuando fue presentado por Google en 2018. A diferencia de los modelos de lenguaje tradicionales, que solo consideran el contexto anterior o posterior de una palabra, BERT tiene en cuenta ambos contextos simultáneamente. Esto le permite capturar relaciones más ricas y complejas entre las palabras y producir representaciones de mayor calidad (Koroteev, 2021).

El entrenamiento de BERT es un proceso que implica grandes cantidades de datos textuales no etiquetados. Utiliza tareas de preentrenamiento, como la predicción de palabras enmascaradas y la predicción de la siguiente oración, para aprender representaciones contextualizadas de las palabras (Rogers et al., 2020). Este entrenamiento previo le permite aprender una comprensión profunda del lenguaje y luego se puede ajustar o fine-tunear para tareas específicas, como la clasificación de sentimientos, la extracción de información o la traducción automática. Una de las principales ventajas de BERT es su capacidad para capturar el contexto y la ambigüedad en el lenguaje.

[2] Bidirectional Encoder Representations from Transformers

[3] procesamiento del lenguaje natural

La capacidad de BERT para capturar la ambigüedad y el contexto del lenguaje ha llevado a mejoras significativas en una amplia gama de tareas de NLP. Por ejemplo, en la clasificación de sentimientos, BERT puede comprender el matiz y la connotación de una oración para determinar si es positiva, negativa o neutral. En la extracción de información, BERT puede identificar entidades y relaciones en el texto con mayor precisión (Alsentzer et al., 2019; Kanade et al., 2020).

Además, los embeddings y BERT han sido fundamentales en la traducción automática, donde se ha demostrado que mejoran la calidad y la fluidez de las traducciones generadas (Shimanaka et al., 2019). También han impulsado avances en la generación de texto, permitiendo la producción de textos más coherentes y contextuales.

Uno de los casos de uso más comunes de BERT y en el cual se basa este trabajo es en la comparación de frases, donde el objetivo es determinar la similitud o relación semántica entre dos oraciones. La comparación de frases con BERT implica la tarea de determinar la similitud o la relación semántica entre dos oraciones (Koroteev, 2021; Reimers & Gurevych, 2019; Tahmid et al., 2020). Este enfoque es útil en aplicaciones como la búsqueda de respuestas, la clasificación de duplicados de texto y la recuperación de información, donde es necesario encontrar oraciones que sean semánticamente similares a una consulta dada.

El proceso de comparación de frases con BERT consta de varias etapas. El primer paso en la generación de embeddings es la tokenización. BERT utiliza una técnica llamada WordPiece, que divide las oraciones en subunidades más pequeñas, conocidas como tokens. Estos tokens pueden representar palabras completas o partes de palabras. Estos tokens luego se procesan en el modelo BERT para generar las representaciones contextualizadas (Song et al., 2020).

Después de la tokenización, se agrega un token especial [CLS][4] al comienzo de la primera oración y un token especial [SEP][5] entre las dos oraciones. Estos tokens especiales permiten a BERT distinguir entre las oraciones y realizar tareas específicas, como la clasificación de pares de oraciones.

A continuación, se procesa la secuencia de tokens a través de múltiples capas de transformadores en la arquitectura de BERT. Los transformadores son bloques de construcción clave que permiten a BERT capturar la relación entre las palabras en función del contexto. Estas capas transformadoras permiten que BERT sea capaz de comprender tanto el contexto anterior como posterior de cada palabra en una oración (Zhan et al., 2020).

Durante el procesamiento de las capas transformadoras, BERT emplea una técnica fundamental llamada “masked attention” (atención enmascarada). Esta técnica permite al modelo dirigir su atención a diferentes partes de las oraciones a medida que captura las relaciones semánticas entre los tokens. La atención enmascarada es una variante de la atención utilizada en los modelos de Transformer, la arquitectura en la que se basa BERT.

En el contexto de BERT, la atención enmascarada se aplica durante la fase de entrenamiento. Consiste en ocultar o “enmascarar” aleatoriamente ciertos tokens en una oración antes de presentarla al modelo. De esta manera, BERT no tiene acceso directo a la información de esos tokens enmascarados durante el entrenamiento, lo que lo obliga a aprender a inferir su significado basándose en el contexto proporcionado por los tokens circundantes (Clark et al., 2019).

[4] classification

[5] separator

Al utilizar la atención enmascarada, BERT puede capturar relaciones tanto a nivel local como a nivel global dentro de una oración. Al realizar múltiples cálculos de atención simultáneamente, conocidos como atención multi-cabeza, BERT es capaz de capturar una variedad de patrones de relación entre los tokens. Cada cabeza de atención se especializa en capturar un aspecto particular de las relaciones, lo que permite una representación más rica y completa de las interacciones semánticas en el texto (Huang et al., 2019).

Una vez que las oraciones se han procesado a través de las capas transformadoras, se obtienen representaciones contextualizadas para cada token. Estas representaciones capturan la información semántica y contextual de cada palabra en función del contexto global de las oraciones.

Para la comparación de frases, los embeddings más relevantes son los que se obtienen de los tokens [CLS], que representan la oración completa o el par de oraciones completo. Estos embeddings agregados capturan la esencia semántica de las oraciones y se utilizan para calcular la similitud entre ellas (Kim et al., 2021).

Una forma común de calcular la similitud entre los embeddings es utilizando la similitud coseno. Este cálculo mide la similitud direccional entre dos vectores y devuelve un valor entre -1 y 1. Un valor cercano a 1 indica una similitud alta, mientras que un valor cercano a -1 indica una similitud baja (Reimers & Gurevych, 2019). Las etapas descritas anteriormente se pueden ver esquematizadas en la *Figura 1*.

Otra técnica popular es utilizar las representaciones de las oraciones para entrenar un modelo adicional, como una red neuronal, que pueda clasificar la similitud entre las oraciones de manera más precisa (Li et al., 2020).

Es importante destacar que BERT se entrena en grandes cantidades de datos no etiquetados y se fine-tunea en tareas específicas con datos etiquetados. Esto le permite adaptarse a dominios y tareas particulares y mejorar su rendimiento en la comparación de frases.

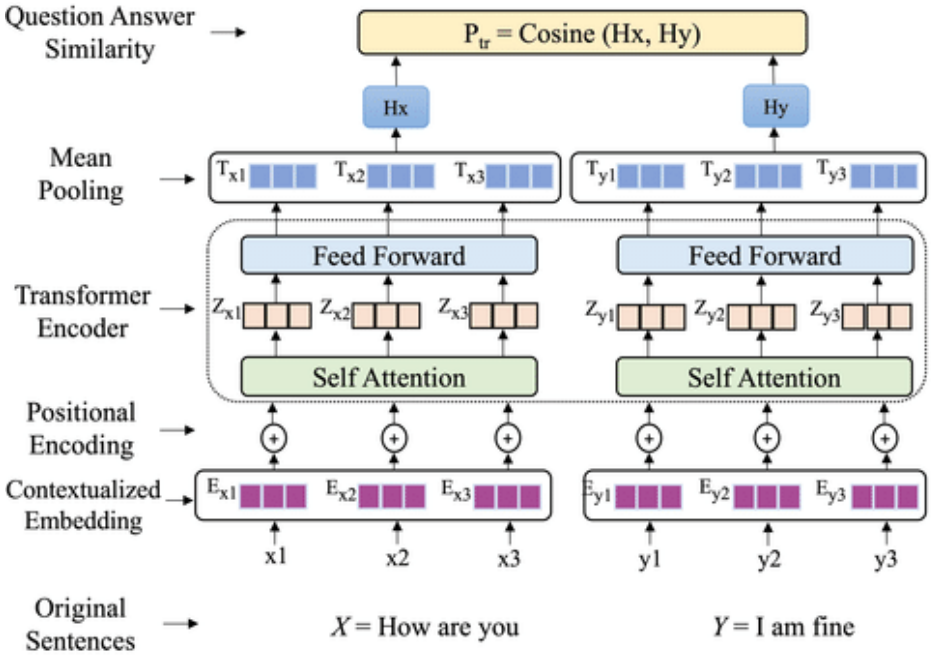

Figura 1: Esquema de las etapas en la comparación de oraciones con BERT. Extraído de: (Tahmid et al., 2020)

2.2 Modelos planteados

Durante la elaboración de este proyecto, se llevó a cabo un exhaustivo análisis y prueba de varios modelos de BERT disponibles para el procesamiento del lenguaje en español. El objetivo principal fue encontrar el modelo más adecuado y eficiente para la tarea específica de identificar y analizar tweets relacionados con la sostenibilidad.

Se consideraron modelos como BETO, CamemBERT, mBERT, sBERT, DistilBERT y BioBERT. Cada uno de estos modelos fue evaluado en términos de su rendimiento en tareas de clasificación y similitud de frases, así como en su capacidad para capturar características semánticas y sintácticas relevantes para el análisis de sostenibilidad. A continuación, se explicará en detalle cada uno de ellos, con sus ventajas e inconvenientes.

2.2.1 sBERT

sBERT, o Sentence-BERT, es un enfoque que se centra específicamente en la generación de representaciones de alta calidad para oraciones o textos cortos. A diferencia de BERT y DistilBERT, que se enfocan en el procesamiento de texto a nivel de palabra, sBERT está diseñado para capturar la semántica de oraciones completas y generar representaciones vectoriales densas y significativas.

El principio fundamental detrás de sBERT es utilizar una técnica de entrenamiento conocida como siameses, que emplea redes neuronales siamesas para aprender a comparar y medir la similitud semántica entre pares de oraciones. Durante el entrenamiento, el modelo aprende a mapear oraciones semánticamente similares a puntos cercanos en el espacio vectorial y oraciones semánticamente diferentes a puntos más distantes. Este proceso de aprendizaje permite que SBERT capture información semántica y contextos relevantes para la comparación de oraciones (Wang & Kuo, 2020).

Una de las ventajas principales de sBERT es su capacidad para generar representaciones vectoriales de alta calidad que reflejan la similitud semántica entre oraciones. Estas representaciones densas y significativas pueden ser utilizadas para tareas como recuperación de información, búsqueda semántica, agrupamiento de textos y clasificación de similitud. sBERT ha demostrado un rendimiento sobresaliente en comparación con otros enfoques en una variedad de conjuntos de datos y desafíos de evaluación.

Además, sBERT es altamente eficiente en términos de tiempo de inferencia. Una vez que el modelo está entrenado, puede generar representaciones vectoriales para oraciones de manera rápida y escalable, lo que lo hace adecuado para su aplicación en sistemas en tiempo real.

Sin embargo, una posible desventaja de sBERT es que su rendimiento y calidad de representación dependen en gran medida de la calidad y tamaño del conjunto de datos de entrenamiento. Es importante tener suficientes ejemplos de pares de oraciones etiquetadas con su similitud o diferencia semántica para que el modelo pueda aprender patrones significativos y generalizables (S. Guo et al., 2022).

2.2.2 BioBERT

Un modelo que podría tener relación con la naturaleza es BioBERT. Aunque BioBERT no está específicamente diseñado para el idioma español, es un modelo basado en BERT que ha sido preentrenado en un gran corpus de texto biomédico y científico en inglés (van Aken et al., 2021).

BioBERT se ha entrenado en una amplia gama de textos relacionados con la biología, medicina y ciencias de la vida, lo que le permite capturar conocimientos y terminología específica de estos campos. Esto lo convierte en un modelo especialmente útil para tareas de procesamiento del lenguaje natural relacionadas con la naturaleza, como análisis de textos científicos, reconocimiento de entidades biomédicas, clasificación de artículos científicos, extracción de información en el dominio de la biología, entre otros.

La ventaja de utilizar BioBERT en tareas relacionadas con la naturaleza es que el modelo ha sido entrenado con una gran cantidad de información y conocimientos en el campo biomédico y científico. Esto le permite comprender conceptos específicos, terminología técnica y relaciones entre entidades en este dominio, lo cual es de gran utilidad para analizar y comprender textos relacionados con la naturaleza y las ciencias de la vida.

Sin embargo, es importante tener en cuenta que BioBERT está preentrenado en inglés y puede no capturar la especificidad del idioma español en términos de terminología y contextos particulares. Si bien puede ser utilizado como punto de partida para tareas relacionadas con la naturaleza en español, es posible que se requiera un ajuste adicional o una adaptación al idioma específico para obtener los mejores resultados.

2.2.3 DistilBERT

DistilBERT es una versión compacta y comprimida de BERT. Fue desarrollado por los investigadores de Hugging Face como una alternativa más liviana y eficiente en términos de recursos computacionales, manteniendo un rendimiento comparable en muchas tareas de procesamiento del lenguaje.

La idea detrás de DistilBERT es destilar el conocimiento del modelo BERT original en una versión más pequeña, eliminando ciertos componentes que no son esenciales para el rendimiento general del modelo. Esto incluye la reducción del número de capas y la disminución de la dimensión oculta (Sanh et al., 2019).

Una de las principales ventajas de DistilBERT radica en su huella de memoria más pequeña y su menor carga computacional en comparación con BERT. Esto permite tiempos de ejecución más rápidos y un uso más eficiente de los recursos computacionales, lo que resulta beneficioso en aplicaciones en las que se requiere un procesamiento rápido y en entornos con restricciones de recursos.

A pesar de su tamaño reducido, DistilBERT ha demostrado ser altamente eficaz en una amplia gama de tareas de procesamiento del lenguaje, incluyendo clasificación de texto, generación de texto, extracción de información y tareas de comprensión del lenguaje. Aunque puede haber

una ligera disminución en el rendimiento en comparación con BERT en algunas tareas más complejas y especializadas, en general, DistilBERT ofrece un equilibrio entre eficiencia y rendimiento adecuado para muchas aplicaciones prácticas (Cañete et al., 2022).

Es importante tener en cuenta que DistilBERT logra su compresión mediante la técnica de destilación del conocimiento, lo que implica entrenar el modelo utilizando el modelo más grande y luego transferir ese conocimiento a una versión más pequeña. Esto permite que DistilBERT beneficie de las representaciones aprendidas por el modelo más grande y mantenga un rendimiento competitivo en una variedad de tareas.

2.2.4 CamemBERT

CamemBERT es una variante de BERT que ha sido adaptada y preentrenada específicamente para el idioma español. Al igual que BERT, CamemBERT utiliza una arquitectura de red neuronal de transformer bidireccional y ha sido entrenado en grandes cantidades de texto en español (Nozza et al., n.d.).

La principal ventaja de CamemBERT es su capacidad para comprender y generar representaciones de texto en español con mayor precisión y especificidad. Al haber sido entrenado en un corpus específico en español, el modelo es capaz de capturar patrones lingüísticos y semánticos específicos del idioma, lo que le permite realizar tareas de procesamiento del lenguaje natural de manera más efectiva.

Además, CamemBERT también tiene en cuenta las particularidades del español, como las variaciones dialectales y las construcciones gramaticales propias del idioma. Esto permite que el modelo capture las sutilezas y matices del español, lo que genera un mejor rendimiento en tareas específicas para el idioma.

Otra ventaja de CamemBERT es su capacidad para realizar transferencia de aprendizaje. Esto significa que el modelo preentrenado en español puede ser utilizado como punto de partida para tareas de procesamiento del lenguaje en español, incluso con conjuntos de datos más pequeños. Esto ocasiona un ahorro significativo de tiempo y recursos computacionales, ya que el modelo no necesita ser entrenado desde cero para cada tarea específica (DeGenaro & Kalita, 2022).

Sin embargo, una desventaja de CamemBERT es que su disponibilidad y recursos pre-entrenados pueden ser limitados en comparación con BERT o modelos más ampliamente utilizados. Esto se debe a que el enfoque se centra específicamente en el español y puede haber menos recursos disponibles en comparación con idiomas más populares o ampliamente estudiados.

2.2.5 BETO

BETO es un modelo basado en BERT que ha sido adaptado y preentrenado específicamente para el procesamiento del lenguaje en español. Al igual que BERT, BETO utiliza una arquitectura de transformer bidireccional y ha sido entrenado en grandes cantidades de texto en español.

La principal ventaja de BETO es su capacidad para comprender y generar representaciones de texto en español de manera precisa y efectiva. Al haber sido entrenado en un corpus extenso en español, el modelo puede capturar las características lingüísticas, gramaticales y semánticas propias del idioma, lo que le permite realizar diversas tareas de procesamiento del lenguaje natural con alto rendimiento (*BETO: Spanish BERT. Transformer Based Models Are Creating… | by Elvis | DAIR.AI | Medium*, n.d.).

BETO ha demostrado ser especialmente útil en tareas como clasificación de texto, extracción de información, análisis de sentimientos y resumen de texto en español. Su capacidad para capturar las sutilezas y matices del idioma español le brinda una ventaja en comparación con modelos más generales que no están específicamente adaptados al español.

Además, BETO también ofrece la posibilidad de realizar transferencia de aprendizaje. Esto significa que el modelo preentrenado en español puede ser utilizado como punto de partida para tareas específicas de procesamiento del lenguaje en español, lo que ahorra tiempo y recursos computacionales al no requerir entrenamiento desde cero para cada tarea (Plaza-del-Arco et al., 2021).

Sin embargo, una limitación de BETO es que, al igual que otros modelos preentrenados basados en BERT, requiere una gran cantidad de recursos computacionales y memoria para su uso eficiente. Esto puede ser un desafío en entornos con recursos limitados o cuando se trabaja con conjuntos de datos masivos.

2.2.6 mBERT

mBERT, o multilingüe BERT, es una versión de BERT que ha sido entrenada en múltiples idiomas, lo que le permite comprender y generar representaciones de texto en varios idiomas diferentes. A diferencia de BERT, que fue entrenado principalmente en inglés, mBERT ha sido entrenado en una amplia gama de idiomas, lo que lo convierte en una opción valiosa para aplicaciones multilingües (Gutiérrez-Fandiño et al., 2021).

La principal ventaja de mBERT es su capacidad para manejar y comprender texto en varios idiomas. Esto se logra a través del entrenamiento en un corpus multilingüe masivo, donde el modelo aprende patrones lingüísticos compartidos y características semánticas en diferentes idiomas. Esto permite que mBERT capture conocimientos y estructuras lingüísticas comunes, lo que a su vez facilita la transferencia de aprendizaje entre idiomas.

Otra ventaja de mBERT es que puede ser utilizado como un modelo de "lenguaje cero". Esto significa que, en lugar de entrenar un modelo separado para cada idioma, mBERT puede ser utilizado directamente para tareas de procesamiento del lenguaje en diferentes idiomas sin necesidad de ajustes adicionales. Esto ahorra tiempo y recursos computacionales, ya que un solo modelo puede ser utilizado para múltiples idiomas.

Sin embargo, mBERT también presenta algunas desventajas. Debido a que ha sido entrenado en múltiples idiomas, el tamaño del modelo es mayor en comparación con BERT u otros modelos monolingües. Esto puede requerir más recursos computacionales y memoria para su implementación y uso (Muller et al., 2020).

Además, aunque mBERT puede comprender y generar representaciones de texto en varios idiomas, no siempre alcanza el mismo rendimiento que los modelos entrenados específicamente para un solo idioma. En algunos casos, modelos monolingües pueden superar a mBERT en tareas específicas en un idioma en particular. Esto se debe a que los modelos monolingües pueden capturar mejor los matices y características idiomáticas específicas.

2.3 Comparación de los modelos BERT

En la *Tabla 1* se puede ver la comparativa entre los diferentes modelos que se han estudiado y descrito anteriormente. Además, se incluye el modelo "hiiamsid/sentence_similarity_spanish_es". Este será el seleccionado y su explicación se detallará en el próximo apartado.

Modelo	Ventajas	Desventajas
sBERT	Captura la semántica de las oraciones mediante la creación de vectores de embeddings que representan el significado de las frases.	Requiere el uso de pares de oraciones etiquetados para el entrenamiento, lo que puede limitar la disponibilidad de datos etiquetados adecuados.
BioBERT	Preentrenado en corpus biomédicos, lo que lo hace adecuado para tareas relacionadas con la biologia y la salud.	Requiere un conjunto de datos grande y específico para el afinamiento o “fine-tuning” en tareas de dominio biomédico.
DistilBERT	Versión más pequeña y comprimida de BERT, que mantiene una calidad de rendimiento comparable con una huella de memoria y carga computacional reducidas.	Puede tener un rendimiento ligeramente inferior en comparación con el modelo completo de BERT en ciertas tareas complejas y conjuntos de datos más grandes.
CamemBERT (español)	Diseñado para el procesamiento del lenguaje español, capturando las características y sutilezas del idioma.	Requiere recursos computacionales más altos debido a su tamaño y complejidad.
BERTO	Lenguaje preentrenado para el español, lo que permite una mejor comprensión del idioma en tareas de procesamiento del lenguaje.	Puede requerir más recursos computacionales para su entrenamiento y uso debido a su tamaño y complejidad.
mBERT	Modelo multilingüe preentrenado que admite múltiples idiomas, incluido el español.	La comprensión contextual puede ser limitada en comparación con modelos específicos del idioma.
hiiamsid/sentence similarity spanish es	Capaz de analizar frases contextualizando al español, mapeando oraciones a un espacio vectorial de 768 dimensiones.	En algunos casos, puede haber dificultad para interpretar adecuadamente el contexto de ciertas frases en español.

Tabla 1: Resumen de los varios modelos planteados y estudiados durante este trabajo

2.4 Modelo escogido

Tras llevar a cabo diferentes pruebas con los modelos de BERT descritos anteriormente, se concluyó que ninguno de ellos cumplía satisfactoriamente con los requisitos y objetivos establecidos para este proyecto en particular. A pesar de sus ventajas y características destacadas, ninguno de los modelos evaluados logró ofrecer resultados precisos y consistentes en la tarea de análisis de similitud de oraciones relacionadas con la sostenibilidad.

Después de considerar cuidadosamente los resultados obtenidos y revisar las opciones disponibles, se tomó la decisión de escoger el modelo "hiiamsid/sentence_similarity_spanish_es" como la mejor opción para abordar la problemática planteada en este proyecto. Este es un modelo de Sentence Transformers, el cual mapea oraciones y párrafos a un espacio vectorial denso de 768 dimensiones y se puede utilizar para tareas como agrupamiento o búsqueda semántica. El modelo del cual surge es del modelo "dccuchile/bert-base-spanish-wwm-cased" y del dataset STSB-Multi-MT (*Hiiamsid/Sentence_similarity_spanish_es · Hugging Face*, n.d.).

2.4.1 Modelo base

El modelo "dccuchile/bert-base-spanish-wwm-cased" es una versión basada en BERT que ha sido adaptada específicamente para el idioma español. Fue desarrollado por el Departamento de Ciencias de la Computación de la Universidad de Chile (DCC). Fue entrenado utilizando una gran cantidad de datos en español, incluyendo textos de diversos dominios y géneros, como noticias, libros, artículos científicos y páginas web. Este proceso de entrenamiento se llevó a cabo utilizando técnicas de aprendizaje automático supervisado, donde el modelo aprende a capturar las relaciones semánticas y sintácticas en el texto mediante la predicción de palabras enmascaradas y la clasificación de pares de oraciones (Teixeira et al., 2022).

La arquitectura de este modelo se basa en transformers, que es una arquitectura de red neuronal que utiliza mecanismos de atención para capturar las dependencias contextuales en el texto. Estos mecanismos de atención permiten que el modelo se centre en partes relevantes del texto durante la representación de las palabras y capture las relaciones entre ellas. Además, el modelo es "cased", lo que significa que distingue entre mayúsculas y minúsculas, lo que da como resultado una representación más precisa y detallada del texto en español.

Este modelo tiene un tamaño similar a BERT-Base y se entrenó utilizando la técnica de Whole Word Masking (WWM). El modelo utiliza un vocabulario de aproximadamente 31k subpalabras BPE construidas con SentencePiece y se entrenó durante 2 millones de pasos.

La técnica de WWM se utiliza durante el entrenamiento de este modelo. Garantiza que, al enmascarar un token específico, si el token corresponde a una subpalabra en una oración, entonces todos los tokens contiguos que conforman la misma palabra también se enmascaran. Esto significa que, en lugar de enmascarar tokens individuales, todos los tokens correspondientes a una palabra se enmascaran a la vez. La tasa general de enmascaramiento se mantiene constante (*SPANISH PRE-TRAINED BERT MODEL AND EVALUATION DATA*, n.d.).

Una vez entrenado, el modelo "dccuchile/bert-base-spanish-wwm-cased" se puede utilizar para diversas tareas de procesamiento del lenguaje natural en español. Estas tareas incluyen clasificación de texto, extracción de información, generación de texto, entre otras. Al aplicar el modelo a un conjunto de datos, se generan representaciones contextuales para cada palabra en el texto, lo que permite capturar tanto la información local como la información global del contexto.

En el contexto de tu proyecto, este modelo se puede utilizar para analizar y clasificar los tweets recopilados de instituciones y oficinas turísticas en español. Al comparar las representaciones de los tweets con las frases de referencia relacionadas con la sostenibilidad, se puede determinar la similitud entre ellos y clasificarlos en diferentes subtemas de sostenibilidad. Esto permite obtener una comprensión más profunda de cómo se aborda la sostenibilidad en los tweets y qué temas son más relevantes en el contexto de las instituciones y oficinas turísticas.

2.4.2 Dataset

El dataset STSB-Multi-MT, también conocido como Semantic Textual Similarity Benchmark Multi-MT, es un conjunto de datos utilizado para evaluar la similitud semántica entre oraciones en varios idiomas. Fue creado como una extensión multilingüe del conjunto de datos STS Benchmark. El propósito de este dataset es proporcionar una evaluación rigurosa y exhaustiva de la similitud textual en diferentes idiomas, lo que resulta relevante para aplicaciones de procesamiento de lenguaje natural.

El dataset STSB-Multi-MT fue construido y anotado por un equipo de investigadores profesionales. Se recolectaron pares de oraciones en varios idiomas, cubriendo una amplia variedad de géneros y dominios, incluyendo noticias, literatura y conversaciones informales. Luego, se les asignó una puntuación de similitud en una escala continua, que va desde 0 (ninguna similitud) hasta 5 (similitud completa), por parte de un grupo de anotadores humanos entrenados. Estas anotaciones se basaron en la percepción de similitud semántica entre las oraciones (*Stsb_multi_mt · Datasets at Hugging Face*, n.d.).

Este dataset es una excelente elección para este proyecto debido a su relevancia y utilidad en la evaluación de la similitud textual en múltiples idiomas. Al tratarse de un dataset multilingüe y anotado con puntuaciones de similitud semántica, proporciona una base sólida y confiable para medir la similitud entre los tweets relacionados con la sostenibilidad y las frases de referencia seleccionadas.

La diversidad de idiomas y géneros presentes en el dataset permite realizar análisis comparativos y precisos en un contexto multilingüe, lo que es especialmente relevante para este proyecto, ya que se trabaja con tweets en español.

Además, el dataset STSB-Multi-MT proporciona una evaluación rigurosa y exhaustiva de la similitud semántica, puesto que las anotaciones fueron realizadas por anotadores humanos entrenados. Esto garantiza la calidad y confiabilidad de las puntuaciones de similitud asignadas a cada par de oraciones.

En conclusión, el campo del NLP busca que las máquinas comprendan y procesen el lenguaje humano de manera similar a los seres humanos. Los modelos tradicionales no capturan la riqueza semántica y contextual del lenguaje, por lo que se utilizan embeddings, representaciones numéricas que capturan información semántica y sintáctica de palabras o frases. Por otro lado, BERT es un modelo de lenguaje que considera ambos contextos simultáneamente, capturando relaciones complejas entre las palabras. BERT se entrena con grandes cantidades de datos no etiquetados y se ajusta para tareas específicas. Su capacidad para capturar contexto y ambigüedad ha mejorado diversas tareas de NLP. Los embeddings y BERT han tenido un impacto en la traducción automática, generación de texto y comparación de frases. En este proyecto, se evaluaron varios modelos de BERT para identificar y analizar tweets sobre sostenibilidad, incluyendo sBERT, BioBERT y DistilBERT, entre otros. Cada uno tiene sus ventajas e inconvenientes en términos de representación semántica, eficiencia y dominio del idioma. No obstante, nos acabamos quedando e implementando con el que consideramos que dio un mejor resultado: "hiiamsid/sentence_similarity_spanish_es". En el siguiente capítulo se evaluarán las diferentes decisiones de diseño que se tomaron para su ejecución.

3 DISEÑO DEL SISTEMA

Para llevar a cabo el proyecto de análisis de tweets sobre sostenibilidad, se deben considerar los siguientes elementos, los cuales serán definidos durante el diseño del sistema:

1. Recopilación de datos: Será necesario recopilar tweets relacionados con la temática de sostenibilidad. Estos tweets pueden obtenerse de cuentas de redes sociales relevantes, como instituciones públicas y oficinas de turismo, que sean representativas en el contexto del proyecto.
2. Definición de frases de referencia y conceptos de sostenibilidad: Será necesario definir un conjunto de frases de referencia y conceptos relacionados con la sostenibilidad. Estas frases y conceptos servirán como criterio para evaluar si los tweets abordan la temática de sostenibilidad y clasificarlos en subtemas específicos.
3. Medida de similitud: Se utilizará una medida de similitud para evaluar la proximidad entre los embeddings de los tweets y las frases de referencia relacionadas con la sostenibilidad. Esta medida permitirá determinar la similitud entre los tweets y su relación con la temática de sostenibilidad.
4. Elección del threshold: Se deberá establecer un umbral adecuado para clasificar los tweets en función de su relevancia con respecto a la sostenibilidad. Este umbral determinará el nivel de similitud requerido entre los embeddings de los tweets y las frases de referencia.
5. Librerías y codificación de los embeddings: Será necesario utilizar librerías específicas para cargar y utilizar el modelo BERT seleccionado, así como para calcular la medida de similitud. Además, los tweets y frases de referencia deberán ser codificados mediante embeddings.

3.1 Recopilación de los tweets

Bajo la supervisión de la doctora Asunción Huertas, del departamento de Comunicación de la URV, se llevó a cabo la selección de destinos y la elaboración de la lista inicial de conceptos de sostenibilidad en el marco de este proyecto. El proceso de selección llevó consigo un exhaustivo proceso de investigación para identificar las cuentas turísticas y los ayuntamientos asociados a cada destino.

El siguiente paso consistió en un análisis minucioso de los tweets publicados por estas cuentas, evaluando aspectos como el número de tweets, el idioma utilizado, la frecuencia de publicación y, lo más importante, la temática relacionada con la sostenibilidad. Para llevar a cabo esta evaluación, se utilizaron criterios establecidos previamente, como la presencia de términos clave relacionados con la sostenibilidad en los tweets y la coherencia en la difusión de mensajes que promovieran prácticas sostenibles. Este proceso se realizó de forma manual. Los resultados obtenidos de este paso se pueden observar en el *Anexo 1*.

Tras este riguroso proceso de análisis y selección, con la ayuda del Dr. Antonio Moreno, se determinó que un conjunto específico de cuentas cumplía con los requisitos establecidos: un volumen considerable de tweets, una actividad constante, así como un enfoque destacado en temáticas relacionadas con la sostenibilidad. A continuación, se procedió a recopilar los tweets correspondientes al periodo comprendido entre 2018 y principios de 2022 de las cuentas

seleccionadas. Es importante destacar que se realizó esta recopilación tanto en español como en catalán. Sin embargo, tras un análisis exhaustivo, se determinó que no se contaba con un modelo óptimo para el procesamiento del catalán, por lo que se decidió continuar el trabajo utilizando únicamente los tweets en español.

En el *Anexo 2* se puede encontrar la *Tabla 10* que muestra todas las cuentas utilizadas en el proyecto, incluyendo información detallada sobre a qué entidad pertenecen y la cantidad de tweets recopilados por año, así como el total estudiado. Además, en la *Tabla 2* se presenta las cuentas seleccionadas para este proyecto, brindando detalles sobre su pertenencia y la cantidad de tweets analizados en cada año.

Destino	**Cuenta**	**2018**	**2019**	**2020**	**2021**	**2022**
Almuñécar	@AlmunecarAyto	2547	2301	2609	2416	420
Chiclana de la Frontera	@ayto_chiclana	1927	1675	1225	1161	263
Isla Cristina	@Ayto_ic	414	159	133	445	22
Marbella	@Ayto_Marbella	1914	1587	1589	1918	418
Estepona	@AytoEstepona	319	612	716	1042	188
Roquetas de Mar	@AytoRoquetas	1168	1065	729	860	277
El Puerto de Santa María	@ElPuerto	2667	2360	2020	2026	467
Fuengirola	@fuengirola	2015	1698	1339	1684	391
Torremolinos	@Torremolinos_On	2503	2001	1403	1089	316
Adeje	@costa_adeje	131	60	428	296	22
	@adeje	1322	739	543	597	190
Arona	@AytoArona	1912	1336	825	464	94
Puerto de la Cruz	@puertodelacruz	350	280	484	464	140
La Oliva	@AytoLaOliva	193	357	453	194	36
Pájara	@aytopajara	191	250	341	304	69
Mogán	@MunicipioMogan	2080	1985	1999	1698	387
San Bartolomé de Tirajana	@AytoSBT	831	494	369	321	119
Baleares	@TurismeBalears	1576	1645	1373	1168	204
Calvià	@_Calvia	239	41	29	260	76
Denia	@DeniaTurismo	214	237	381	638	110
Gandía	@visitgandia	1001	249	484	614	169
Peñíscola	@_peniscola	281	484	353	317	34
	@ajuntpeniscola	482	484	1416	1271	213
Benidorm	@visitbenidorm	787	1133	1380	1478	447
	@BenidormAyto	1342	1100	1039	1017	288
Benicasim	@tmobenicassim	199	605	886	383	66
	@AytoBenicassim	1	0	0	232	90
Cartagena	@AytoCartagenaES	1082	1075	324	1598	476

Tabla 2: Cuentas seleccionas y recuento de tweets analizados en cada año

3.2 Frases y conceptos de referencia

En el contexto de este proyecto, se ha llevado a cabo la creación de un conjunto de frases de referencia destinadas a evaluar y clasificar los tweets recopilados en relación con la temática de la sostenibilidad. Estas frases de referencia, un total de 77, han sido meticulosamente elaboradas con el fin de establecer una conexión inequívoca y relevante entre los conceptos específicos y los aspectos clave de la sostenibilidad, contextualizándolas adecuadamente.

La selección de las frases de referencia ha sido un proceso exhaustivo y riguroso. Para ello, se partió de un conjunto inicial de 282 conceptos relacionados con la sostenibilidad, que fue elaborado en colaboración con el tutor del proyecto. A partir de este conjunto, se realizó una selección cuidadosa de aquellos conceptos más representativos y pertinentes para la temática en cuestión.

Posteriormente, se redactaron las frases de referencia con el objetivo de capturar de manera precisa la esencia de cada concepto y su relación con la sostenibilidad. Se puso especial énfasis en formular las frases de manera clara, concisa y comprensible, de modo que fueran capaces de reflejar de manera acertada la relación entre el concepto y la temática en análisis. Las frases para cada concepto se pueden ver en el *Anexo 3*.

Es importante destacar que las frases de referencia seleccionadas desempeñan un papel fundamental en el análisis y clasificación de los tweets, ya que serán utilizadas como punto de comparación para determinar si un tweet aborda o no la temática de la sostenibilidad. Sin embargo, es importante tener en cuenta que este enfoque también puede presentar limitaciones, ya que las comparaciones se realizarán en función de las frases de referencia seleccionadas, lo que implica que la precisión y la clasificación dependerán en gran medida de la adecuación y representatividad de estas frases en relación con los tweets analizados. Los 77 conceptos seleccionados se pueden observar en la *Tabla 3*.

Variabilidad climática	Retroceso de la costa	Sobreturismo y consumo excesivo de recursos	Compromiso con el cambio climático	Reducción consumo de energía y fuentes de energía renovables
Cambio climático global	Blanqueamiento de corales	Huella ambiental	Campañas climáticas y proyectos con el cambio climático	Economía circular
Cambio climático local	Marejadas ciclónicas y mareas	Clima antropogénico	Greenwashing	Gestión de residuos
Calentamiento global	Agotamiento del ozono	Contaminación por bombeos de aguas residuales	Justicia social y salud mundial en relación al cambio climático	Adaptación al cambio climático
Peligro del calentamiento actual	Temporada esquí más corta y mala calidad nieve	Invasión humana y falta de mitigación	Activismo ambiental y defensa del clima	Servicios sostenibles
Riesgo climático	Escasez de agua	Huella de carbono	Certificaciones ambientales y de sostenibilidad	Certificación de turismo sostenible
Impactos del cambio climático	Especies invasoras y pérdida de habitat	Contaminación por carbono	Esfuerzos de conservación y protocolos de impacto mínimo	Prácticas pro-ambientales
Puntos críticos del cambio climático	Migración de especies y climáticas	Modernización ecológica	Política medioambiental	Medidas de mantenimiento de playas
Peligro del cambio climático	Limitación fabricación de nieve	Credenciales ecológicas	Información medioambiental y comunicación climática en línea	Equidad global
Escépticos del clima	Estrés ecológico	Sociedad post-carbono	Grupos de acción de racionamiento de carbono y mitigación del clima	Reciclaje
Las ONG ambientales	Escala de paradigma ecológico	Acciones contra el cambio climático	Comunicación sobre resiliencia ambiental	Conciencia ambiental
Alargamiento del verano y acortamiento del invierno	Índice Climático del Turismo	Ciencia del cambio climático	Negocios y hoteles sostenibles	Juegos sobre el cambio climático
Nevadas y fuertes lluvias	Confort climático	Comunicación sobre el cambio climático	Ecohoteles y parajes ecológicos	Gases de efecto invernadero
Acidificación de los océanos	Contaminación terrestre y marina	Impuestos al carbono	Alimentos sotenibles	Desperdicio del agua
Fuertes vientos	Derrames de petróleo	Captura, reducción y compensación de carbono	Edificios sostenibles	Senderismo
Reducción de desecho	Prevención de incendios			

Tabla 3: Conceptos sostenibilidad seleccionados

3.3 Similitud coseno

En este proyecto, la medida de similitud seleccionada para evaluar la proximidad semántica entre los tweets y las frases de referencia es la similitud coseno. Esta medida es ampliamente utilizada en el campo del NLP y se basa en el cálculo del ángulo coseno entre dos vectores de características.

La similitud coseno es la indicada para el proyecto debido a sus propiedades y ventajas. En primer lugar, permite medir la similitud entre dos textos o frases, considerando tanto la dirección como la magnitud de los vectores de características. Esto significa que no solo se tiene en cuenta qué palabras están presentes en ambos textos, sino también la importancia relativa de cada palabra en el contexto (Reimers & Gurevych, 2019).

El cálculo de la similitud coseno se realiza mediante el producto escalar entre los vectores de características y la división por el producto de sus magnitudes. Este proceso refleja la similitud de las direcciones y la distancia entre ellos en un espacio vectorial. Cuanto más cercanos sean los vectores, mayor será la similitud coseno, y viceversa. Se expresa siguiendo la fórmula 1.

$$\text{Similitud coseno} = \frac{\vec{u}\,\vec{v}}{|u||v|} \quad (1)$$

La similitud coseno se considera una buena medida para el proyecto debido a su capacidad para capturar las relaciones semánticas y el significado contextual de las oraciones. En lugar de basarse únicamente en la coincidencia exacta de palabras, la similitud coseno tiene en cuenta la similitud de los contextos y la semántica subyacente. Esto resulta especialmente relevante en la clasificación de tweets sobre sostenibilidad, donde el contexto y el significado son cruciales para identificar la relevancia del contenido.

Además, la similitud coseno es computacionalmente eficiente y escalable, lo que la hace adecuada para el procesamiento de grandes volúmenes de datos, como es el caso de los tweets recopilados. Su implementación se encuentra disponible en librerías de NLP, lo que facilita su integración en el flujo de trabajo del sistema de análisis de tweets.

3.4 Elección del threshold

El apartado de elección de threshold se centra en determinar el umbral adecuado para clasificar los tweets en función de la precisión y exhaustividad con respecto a la temática de sostenibilidad. Es importante destacar que el proceso de elección de threshold es fundamental para establecer un equilibrio entre la precisión y la exhaustividad de los resultados. Un threshold muy alto puede resultar en una alta precisión, pero una baja exhaustividad, ya que solo se considerarán aquellos tweets muy similares a las frases de referencia. Por otro lado, un threshold muy bajo puede aumentar la exhaustividad, pero disminuir la precisión al incluir tweets menos relevantes. Para llevar a cabo este proceso, se seleccionaron tres cuentas con un alto volumen de tweets, específicamente las cuentas @fuengirola, @Torremolinos_On y @TurismeBalears. A partir de estas cuentas, se realizó una selección manual de aquellos tweets que se consideraron relevantes para la temática de sostenibilidad.

Una vez obtenida esta selección, se procedió a generar los resultados utilizando cuatro thresholds diferentes: 0.4, 0.42, 0.45 y 0.5. Estos thresholds determinan el nivel de similitud requerido entre los tweets y las frases de referencia para considerarlos como relevantes en términos de sostenibilidad.

El siguiente paso consistió en evaluar cuántos de los tweets previamente seleccionados como relevantes se encontraban entre los resultados obtenidos para cada threshold. Esta evaluación se conoce como exhaustividad, que representa la capacidad del sistema para recuperar correctamente los tweets relevantes. Una mayor exhaustividad indica una mayor capacidad del modelo para identificar los tweets relacionados con la sostenibilidad.

Posteriormente, se llevó a cabo la evaluación de la precisión para determinar si los tweets estaban correctamente clasificados en sus respectivos subtemas relacionados con la sostenibilidad. Para realizar esta evaluación, se compararon los resultados de clasificación obtenidos utilizando los diferentes thresholds con las asignaciones manuales previas de los tweets a los subtemas correspondientes.

La precisión se calculó dividiendo el número de tweets correctamente clasificados en el subtema apropiado (verdaderos positivos) entre la suma de los verdaderos positivos y los falsos positivos (tweets clasificados incorrectamente en un subtema). Una alta precisión indica una mayor confiabilidad en la clasificación de los tweets en relación con sus subtemas de sostenibilidad.

Este análisis de precisión permitió evaluar la exactitud de la clasificación y determinar si los tweets fueron asignados adecuadamente a los subtemas correspondientes. Una precisión más alta indica una mejor coincidencia entre los tweets y los subtemas, lo cual es fundamental para garantizar la calidad y confiabilidad de los resultados del proyecto.

Tras haber evaluado la precisión y la exhaustividad por separado, se procedió al cálculo de una medida que combina ambos valores, conocida como medida F1. La medida F1 es un indicador comúnmente utilizado en la evaluación de modelos de clasificación y representa el equilibrio entre la precisión y la exhaustividad (*A Look at Precision, Recall, and F1-Score | by Teemu Kanstrén | Towards Data Science*, n.d.).

La medida F1 se calcula como la media armónica de la precisión y la exhaustividad, y se expresa mediante la fórmula 2.

$$\text{Medida F1} = 2 * \frac{\text{precisión} * \text{exhaustividad}}{\text{precisión} + \text{exhaustividad}} \quad (2)$$

La importancia de la medida F1 radica en su capacidad para proporcionar una evaluación más completa del desempeño del modelo de clasificación. Al combinar la precisión y la exhaustividad en una sola medida, la medida F1 permite tener una visión equilibrada de la capacidad del modelo para identificar correctamente los tweets relevantes y clasificarlos en los subtemas apropiados relacionados con la sostenibilidad.

En términos más técnicos, la medida F1 tiene en cuenta tanto los verdaderos positivos (tweets correctamente clasificados en el subtema correcto) como los falsos positivos (tweets incorrectamente clasificados en un subtema) y los falsos negativos (tweets relevantes que no

fueron clasificados en el subtema correspondiente). Al considerar estos tres aspectos, la medida F1 proporciona una evaluación más completa y equilibrada del rendimiento del modelo.

La medida F1 se utiliza ampliamente en la evaluación de modelos de clasificación debido a su capacidad para tener en cuenta tanto la precisión como la exhaustividad, evitando así un enfoque sesgado hacia uno u otro. Un valor alto de la medida F1 indica un buen equilibrio entre la capacidad de clasificación precisa y la capacidad de recuperación de los tweets relevantes.

Los resultados de precisión, exhaustividad y medida F1 en tanto por ciento, se encuentran en la *Tabla 4* para la cuenta @fuengirola, en la *Tabla 5* para la cuenta @Torremolinos_On y en la *Tabla 6* para la cuenta @TurismeBalears. Después de calcular la medida F1 para los cuatro thresholds evaluados (0.4, 0.42, 0.45 y 0.5), en las diferentes cuentas, se determinó que el threshold 0.42 ofrecía la medida F1 óptima. La medida F1 obtenida para cada threshold reflejó el equilibrio entre la precisión y la exhaustividad en la clasificación de los tweets relevantes en los subtemas de sostenibilidad.

Al evaluar los resultados, se observó que el threshold 0.42 mostraba un rendimiento destacado en términos de la medida F1. Esto significa que este threshold proporcionaba un equilibrio adecuado entre la capacidad de clasificar correctamente los tweets relevantes (precisión) y la capacidad de recuperar la mayoría de los tweets relevantes (exhaustividad).

Es importante destacar que la elección del threshold óptimo se basó en el objetivo específico del proyecto y en la necesidad de maximizar tanto la precisión como la exhaustividad para garantizar una clasificación precisa y completa de los tweets en los subtemas de sostenibilidad.

	0,4	**0,42**	**0,45**	**0,5**
Precisión	65,02	71,41	75,11	85,87
Exhaustividad	70,19	68,55	61,7	24,56
Medida F1	67,51	69,95	67,75	38,20

Tabla 4: Medida F1 para threshold 0.4, 0.42, 0.45 y 0.5 de la cuenta @fuengirola

	0,4	**0,42**	**0,45**	**0,5**
Precisión	52,92	58,35	62,76	91,35
Exhaustividad	78,05	72,87	65,88	22,87
Medida F1	63,07	64,81	64,28	36,58

Tabla 5: Medida F1 para threshold 0.4, 0.42, 0.45 y 0.5 de la cuenta @Torremolinos_On

	0,4	**0,42**	**0,45**	**0,5**
Precisión	56,6	65,17	71,48	93,33
Exhaustividad	66,11	62,46	60,23	27,98
Medida F1	60,99	63,79	65,37	43,05

Tabla 6: Medida F1 para threshold 0.4, 0.42, 0.45 y 0.5 de la cuenta @TurismeBalears

3.5 Librerías necesarias

En el desarrollo de este proyecto, se han utilizado diferentes librerías que desempeñan un papel fundamental en el NLP y el cálculo de similitud coseno. A continuación, se detallan estas librerías y su función en el sistema de análisis de tweets:

- P**yTorch**: PyTorch es un framework de aprendizaje automático de código abierto que proporciona herramientas y funciones para construir y entrenar modelos de aprendizaje automático. Se utiliza ampliamente en tareas de procesamiento de datos y construcción de modelos. En este proyecto, utilizamos la librería PyTorch para realizar operaciones en tensores y realizar cálculos numéricos en el modelo BERT preentrenado (Paszke et al., 2019).
- **Transformers**: Transformers es una biblioteca de código abierto desarrollada por Hugging Face que proporciona una amplia gama de modelos preentrenados y herramientas para el procesamiento del lenguaje natural (NLP). En este proyecto, utilizamos la librería transformers para cargar y utilizar el modelo BERT preentrenado. También utilizamos las clases AutoTokenizer y AutoModel para la tokenización de texto y la carga del modelo BERT respectivamente (Wolf et al., 2019).
- **scikit-learn**: scikit-learn es una librería de aprendizaje automático de código abierto que proporciona una amplia gama de herramientas y algoritmos para tareas de aprendizaje automático. En este proyecto, utilizamos la función cosine_similarity de la sublibrería metrics.pairwise para calcular la similitud coseno entre vectores de características (Soyusiawaty & Zakaria, 2018).

Además de estas librerías, también se importan otras librerías estándar, como csv y os, para realizar operaciones de lectura y escritura de archivos CSV, y copy para realizar copias de listas sin modificar la original. En la *Tabla 7* se muestra un resumen de las librerías utilizadas en este proyecto, junto con una breve descripción de su funcionalidad en el código.

Librería	**Funcionalidad en el código**
PyTorch	Operaciones en tensores y cálculos numéricos en el modelo BERT
Transformers	Carga del modelo BERT, tokenización de texto y generación de embeddings
scikit-learn	Cálculo de similitud coseno entre vectores de características
os	Manipulación de rutas y archivos en el sistema operativo
csv	Lectura y escritura de archivos CSV
copy	Creación de copias profundas de objetos

Tabla 7: Resumen librerías utilizadas durante el proyecto

3.6 Codificación de los tweets

El código 1 muestra la función get_embedding(tokenizer, model, text, token_length) que se encarga de codificar los embeddings de un texto utilizando el modelo BERT preentrenado seleccionado, es decir, "hiiamsid/sentence_similarity_spanish_es".

```
def get_embedding(tokenizer, model, text, token_length):

    tokens = tokenizer(text, max_length=token_length, padding='max_length',
    truncation=True)

    output       =       model(torch.tensor(tokens.input_ids).unsqueeze(0),
    attention_mask=torch.tensor(tokens.attention_mask).unsqueeze(0)).hidden
    _states[-1]

    return torch.mean(output, axis=1).detach().numpy()
```

Código 1: Codificación de los embeddings

Primero, el parámetro tokenizer en la función get_embedding es una instancia que se encarga de dividir el texto de entrada en unidades más pequeñas llamadas tokens para que se genere un formato entendible por el modelo. El parámetro text es el texto de entrada que debe ser tokenizado. El parámetro token_length especifica la longitud máxima de la secuencia tokenizada. Si la longitud de la secuencia tokenizada supera este valor, se truncará. El padding garantiza que todas las secuencias de tokens tengan la misma longitud, agregando tokens de relleno si es necesario. En este caso, el parámetro padding especifica cómo rellenar la secuencia si su longitud es menor que token_length. Se establece en 'max_length', lo que significa que la secuencia se rellenará hasta la longitud máxima. El parámetro truncation especifica si se debe truncar la secuencia si su longitud supera token_length. En este caso, se establece en True, lo que significa que la secuencia se truncará si su longitud supera token_length.

Una vez que el texto ha sido tokenizado y preparado para el modelo BERT, se procede a obtener la salida del modelo (output) al alimentarlo con los tokens de entrada (tokens.input_ids) y la máscara de atención (tokens.attention_mask). La máscara de atención es una matriz binaria que indica qué tokens son relevantes y cuáles son de relleno. Esta información permite que el modelo se enfoque en los tokens significativos durante el procesamiento.

El uso de unsqueeze(0) se realiza para agregar una dimensión adicional a los datos de entrada, de modo que se ajusten a la estructura de entrada esperada por el modelo BERT. El modelo BERT espera un tensor de forma [batch_size, sequence_length], donde batch_size representa el número de muestras en un lote y sequence_length es la longitud máxima de la secuencia tokenizada.

Una vez que se ha pasado el texto por el modelo BERT, se obtienen las capas ocultas del modelo, que contienen información contextualizada de cada token. Estas capas ocultas se acceden a través del atributo hidden_states. En este caso, se selecciona la última capa oculta (hidden_states[-1]). En general, las capas ocultas más cercanas a la salida suelen contener información más relevante y contextualizada para la tarea en cuestión.

Finalmente, se calcula el promedio de los vectores de características de la última capa oculta utilizando torch.mean(output, axis=1). Esto se realiza para reducir la dimensión de los vectores de características de la última capa oculta a un solo vector representativo. El promedio permite capturar una representación agregada de la información contextualizada contenida en los vectores de características.

Después de calcular el promedio de los vectores de características de la última capa oculta utilizando torch.mean(output, axis=1), se aplica la función detach() para eliminar cualquier

conexión entre el tensor output y su historial de cálculo. Esto se hace para asegurarse de que no se realicen cambios en los gradientes durante el proceso de cálculo del promedio.

A continuación, se utiliza el método numpy() para convertir el tensor resultante en un arreglo NumPy. NumPy es una biblioteca popular en Python utilizada para el procesamiento numérico y computacional, y su uso en este contexto nos permite obtener una representación de los embeddings como un arreglo NumPy, lo cual puede ser más conveniente para su posterior procesamiento o almacenamiento.

El resto del código que se encarga de realizar el procesamiento de los archivos y el cálculo de similitud se encuentra en el *Anexo 4*. En este anexo, se pueden encontrar las funciones y los pasos necesarios para leer los archivos, cargar los modelos preentrenados, generar los embeddings de las frases de referencia y calcular la similitud coseno entre tweets y frases de referencia. Además, se realizan otras operaciones como la manipulación de datos y la escritura de los resultados en archivos CSV.

El objetivo principal de este código es procesar los datos de entrada, obtener los embeddings de los tweets y frases de referencia utilizando el modelo preentrenado y calcular la similitud coseno entre los embeddings. Los resultados se almacenan en archivos CSV para su posterior análisis y evaluación.

Es importante mencionar que este código es una implementación específica para este proyecto y se adapta a las necesidades y objetivos particulares del mismo. Cada función y paso tiene su función específica en el flujo de trabajo y contribuye al logro de los resultados esperados.

En resumen, en este capítulo se ha presentado un sistema de análisis de tweets que utiliza el cálculo de similitud coseno y embeddings generados por el modelo BERT preentrenado. El sistema permite identificar tweets similares a frases de referencia predefinidas, utilizando un umbral específico de 0.42. La implementación se ha apoyado en librerías como PyTorch y Transformers, y se han realizado operaciones de codificación de los tweets, manipulación de archivos y cálculo de similitud coseno. En conclusión, este sistema ofrece una forma eficaz de analizar grandes volúmenes de tweets y extraer información relevante en el ámbito de la sostenibilidad, pudiéndose ver en el siguiente capítulo.

4 RESULTADOS

En esta sección se presentarán los resultados obtenidos del análisis de las cuentas relacionadas con el turismo y la sostenibilidad en las redes sociales. Se evaluará tanto la frecuencia con la que se abordan los temas de sostenibilidad en cada cuenta, como los subtemas más relevantes a nivel de cuenta, comunidad autónoma y en general.

4.1 Frecuencia de los tweets sobre sostenibilidad

En la *Figura 2* se puede observar la frecuencia con la que se abordan los temas de sostenibilidad en las cuentas analizadas. Con frecuencia me refiero a la cantidad de veces que se mencionan temas de sostenibilidad en los tweets emitidos por cada cuenta. Cada cuenta se muestra en el eje horizontal, mientras que la frecuencia de los temas de sostenibilidad se representa en el eje vertical.

En el gráfico, se observa que la mayoría de las cuentas presentan una frecuencia inferior a 0.1, lo que indica que la temática de sostenibilidad no es abordada con gran énfasis en la mayoría de las cuentas analizadas. Estas cuentas pueden tener un enfoque menos pronunciado en la promoción de prácticas sostenibles o pueden estar explorando otros temas en sus publicaciones. Es importante considerar que una frecuencia baja no necesariamente indica una falta de compromiso con la sostenibilidad, ya que las estrategias de comunicación y los enfoques pueden variar entre diferentes cuentas y organizaciones.

Al analizar las cuentas de ayuntamientos (azul) y oficinas de turismo (rojo) por separado, se observa una diferencia significativa en la frecuencia de publicaciones relacionadas con la sostenibilidad. En general, las cuentas de turismo muestran una mayor actividad y frecuencia de tweets en comparación con las cuentas de ayuntamiento en este tema. De hecho, la mayoría de las cuentas de turismo superan el 0.1 de frecuencia, lo que indica una mayor dedicación y compromiso en la difusión de información relacionada con el turismo sostenible y la promoción de destinos.

Sin embargo, es interesante destacar una excepción a esta tendencia. La cuenta @_fuengirola, perteneciente a una oficina turística, muestra una frecuencia ligeramente inferior a 0.1. Esto puede indicar una menor prioridad o enfoque en la difusión de contenidos turísticos específicos en esa cuenta en particular.

Entre las cuentas de ayuntamiento, se destacan dos casos en particular: @Ayto_Marbella y @AytoBenicassim. Estas cuentas muestran una frecuencia más alta, con valores de 0.14 y 0.19 respectivamente. Esto indica un mayor compromiso y esfuerzo por parte de estos ayuntamientos en la difusión de información turística y sostenible, lo que puede tener un impacto positivo en la promoción de sus destinos.

En cuanto a las cuentas de destino que han sido analizadas tanto de oficinas de turismo como de ayuntamientos, como Adeje, Peñíscola, Benidorm y Benicasim, se ha observado una tendencia en la cual las cuentas de oficinas de turismo presentan una mayor frecuencia de tweets en comparación con las cuentas de ayuntamiento. Esta diferencia se destaca especialmente en los casos de Adeje, Benidorm y Benicasim. Sin embargo, es importante mencionar que, en el caso de Peñíscola, ambas cuentas muestran una frecuencia similar, lo que

sugiere un esfuerzo equitativo por parte de ambas entidades en la promoción y difusión de información turística relacionada con la sostenibilidad.

Por otro lado, se identifica a la cuenta *@costa_adeje*. Esta, destaca con una frecuencia de 0.3, lo que sugiere un mayor compromiso y enfoque en la comunicación de aspectos relacionados con la sostenibilidad.

Es importante destacar que la cuenta con la mayor frecuencia puede ser objeto de un estudio más detallado para comprender las estrategias y enfoques utilizados en la difusión de mensajes sobre sostenibilidad.

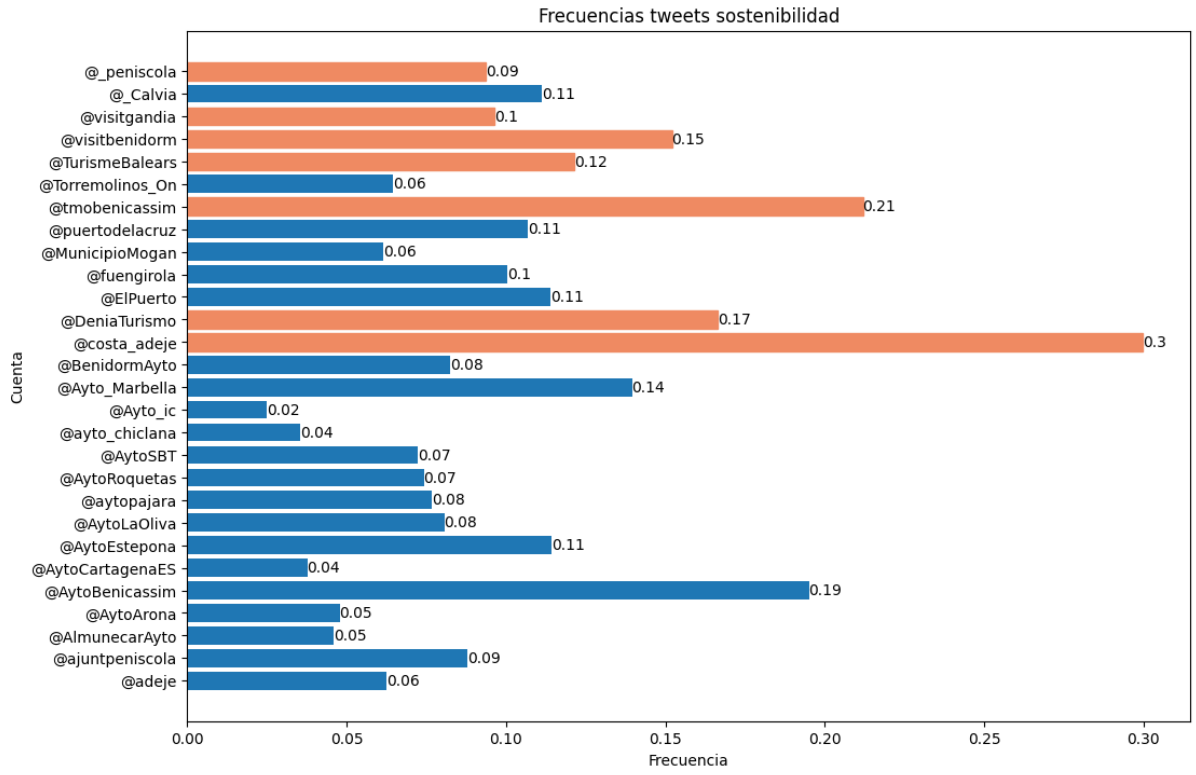

Figura 2: Frecuencia en la que se abordan temas de sostenibilidad en los tweets de las cuentas analizadas

4.2 Análisis de subtemas en los tweets sobre sostenibilidad

A continuación, se llevará a cabo un análisis de los subtemas presentes en los tweets sobre sostenibilidad. Estos subtemas corresponden a los conceptos de referencia previamente definidos y abarcan diversos aspectos relacionados con la sostenibilidad.

El objetivo principal de este análisis es determinar la cantidad de tweets asociados a cada subtema, lo que nos permitirá identificar las temáticas más recurrentes y relevantes en la comunicación sobre sostenibilidad en las redes sociales. Se realizará a nivel de cuenta, comunidad autónoma y en global.

4.2.1 Análisis de subtemas a nivel de cuenta

Debido a la cantidad de cuentas analizadas y los gráficos generados, se ha decidido adjuntar en el *Anexo 5* del informe los gráficos que muestran el recuento de los diferentes subtemas de

sostenibilidad por cuenta. En este anexo, se proporcionará una visualización completa de la distribución de los subtemas en cada cuenta analizada. No obstante, cabe mencionar la cuenta que ha demostrado tener la mayor frecuencia en la comunicación de sostenibilidad: *@costa_adeje*. En la *Figura 3* se puede observar cómo los subtemas de "medidas de mantenimiento de playas" y "ecohoteles y parajes ecológicos", son los más recurrentes. Estos, no solo son los más tratados en la cuenta analizada, sino que también son los más comunes en la totalidad de cuentas evaluadas. Un posible ejemplo de tweet que respalda los temas recurrentes identificados es el siguiente: "Pon rumbo para disfrutar de bosques y paisajes naturales. Descubre el Parque Natural Corona Forestal, el mayor espacio natural protegido del archipiélago. Cuidemos nuestros bosques y protejamos la naturaleza. No arrojes basura en la costa de Adeje".

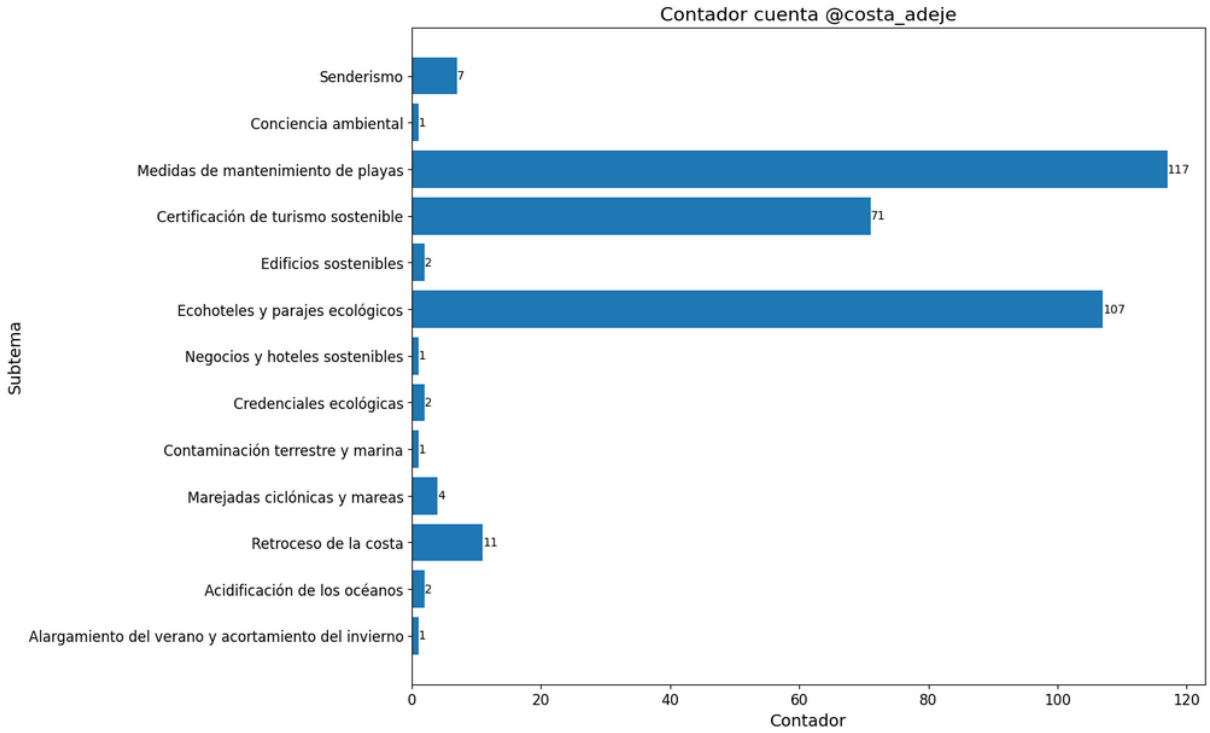

Figura 3: Número de tweets que abarcan los diferentes subtemas en la cuenta @costa_adeje

Se observa que, en general, los temas identificados como más recurrentes se abordan con una frecuencia similar tanto en las cuentas de ayuntamiento como en las cuentas de oficinas de turismo. Esto indica que tanto las entidades municipales como las oficinas de turismo comparten la preocupación por promover la sostenibilidad y abordar los aspectos relacionados con el turismo responsable en sus comunicaciones.

Sin embargo, se destaca una diferencia particular con relación al tema de "medidas de mantenimiento de playas". En este caso, se observa que las cuentas de ayuntamiento abordan este tema con una mayor frecuencia en comparación con las cuentas de oficinas de turismo. Esto sugiere que los ayuntamientos, como autoridades locales, tienen un papel más destacado en la gestión y mantenimiento de las playas, lo cual se refleja en su mayor enfoque en este tema específico.

Es importante tener en cuenta que esta diferencia en la frecuencia de abordaje del tema de "medidas de mantenimiento de playas" no significa que las cuentas de oficinas de turismo no

le den importancia o no se ocupen de este aspecto. Simplemente, indica que los ayuntamientos, como entidades responsables de la administración local, tienden a comunicar más frecuentemente sobre este tema, en particular debido a su rol directo en la gestión de las playas y la implementación de medidas relacionadas.

Por otro lado, es importante destacar que existen algunas cuentas que abordan otros temas en mayor medida. Por ejemplo, la cuenta @AytoArona se centra en la prevención de incendios, publicando regularmente avisos y recomendaciones sobre el riesgo de incendios forestales. Un ejemplo de tweet que respalda este enfoque es el siguiente: "Actualización de la situación de alerta máxima por nivel de riesgo de incendios forestales. El siguiente elemento multimedia incluye contenido delicado. Cambiar configuración para ver. Ayuntamiento de Arona". Otra de las cuentas que también trata con mayor énfasis otros subtemas es la cuenta del ayuntamiento de la localidad de Pájara (@aytopajara). En ella se diferencian tweets como "pájara procede limpieza vertidos barranco canarios ayuntamiento pájara" en los cuáles se intenta tratar de combatir contra la contaminación producida por aguas residuales.

En contraste, la cuenta @ElPuerto trata principalmente sobre el clima, incluyendo información sobre fuertes lluvias y otras condiciones meteorológicas. Aunque estos temas no se corresponden directamente con el subtema específico identificado en el análisis ("nevadas y fuertes lluvias"), los tweets relacionados con el clima demuestran el interés de la cuenta en aspectos ambientales y su impacto en el entorno local. La mayoría de estos tweets son meramente informativos, un ejemplo podría ser "sept comenzamos día cielos nubosos precipitaciones temperaturas van descenso viento sopla sureste índice uv es es alto fuente imagen ayuntamiento puerto santa maría".

Además, se observa que algunas cuentas abordan una gran cantidad de subtemas. Por ejemplo, la cuenta @Ayto_Marbella trata hasta 44 subtemas diferentes en sus tweets. Sin embargo, es importante destacar que en la mayoría de estos subtemas se encuentran pocos tweets. En muchos casos, es común encontrar únicamente un solo tweet relacionado con un subtema específico. Esto sugiere que, aunque la cuenta aborda una amplia variedad de temas, su nivel de profundidad en cada uno de ellos es limitado debido a la falta de volumen de tweets en cada subtema.

Por otro lado, existen cuentas que tratan significativamente menos subtemas. Por ejemplo, la cuenta @Ayto_ic se centra en tan solo 10 temas específicos en sus tweets. Esta diferencia en el número de subtemas tratados se debe principalmente a la cantidad de tweets generados por cada cuenta. Aquellas cuentas, al tener una menor actividad, acaban teniendo menos oportunidades de abordar una amplia variedad de temas

Estos hallazgos nos llevan a comprender que la variedad y profundidad de los subtemas tratados en las cuentas analizadas está fuertemente influenciada por el nivel de actividad y volumen de tweets de cada cuenta.

En los próximos apartados del informe, se analizarán en mayor detalle los resultados obtenidos, destacando los subtemas más recurrentes, posibles patrones y su relevancia en el contexto de la comunicación de la sostenibilidad en las redes sociales de los destinos turísticos.

4.2.2 Análisis de subtemas a nivel de comunidad autónoma

Se agruparon las diferentes cuentas analizadas en función de la comunidad autónoma a la que pertenecían. Las cuentas del proyecto se distribuyeron según se muestra en la *Tabla 8.*

Comunidad autónoma	Cuenta
Andalucía	@AlmunecarAyto
	@ayto_chiclana
	@Ayto_ic
	@Ayto_Marbella
	@AytoEstepona
	@AytoRoquetas
	@ElPuerto
	@fuengirola
	@Torremolinos_On
Islas Canarias	@costa_adeje
	@adeje
	@AytoArona
	@puertodelacruz
	@AytoLaOliva
	@aytopajara
	@MunicipioMogan
	@AytoSBT
Islas Baleares	@TurismeBalears
	@_Calvia
Comunidad valenciana	@DeniaTurismo
	@visitgandia
	@_peniscola
	@ajuntpeniscola
	@visitbenidorm
	@BenidormAyto
	@tmobenicassim
	@AytoBenicassim
Murcia	@AytoCartagenaES

Tabla 8: Distribución cuentas en comunidades autónomas

Se puede observar en la *Figura 4* el análisis de subtemas en los tweets sobre sostenibilidad en Andalucía. En él, se observa que el subtema más tratado es "medidas de mantenimiento de playas". Este resultado puede estar relacionado con la importancia del turismo de sol y playa en la región, así como con la conciencia sobre la necesidad de preservar y mantener adecuadamente las playas como recursos naturales y turísticos.

Este enfoque en las medidas de mantenimiento de playas se debe a la preocupación por la conservación del entorno costero, la gestión adecuada de residuos y la promoción de prácticas sostenibles en relación con el turismo de playa. Los tweets relacionados con este subtema abordan temas como la limpieza de playas, la gestión de residuos, la protección de hábitats marinos y la implementación de prácticas sostenibles en las zonas costeras. Algunos ejemplos pueden ser “alumnos segundo colegio albayyana han han recogido basura playa bajadilla iniciativas gracias info ayto roquetas mar roquetas mar” de la cuenta *@AytoRoquetas* o “hemos

retirado toneladas algas invasoras dispositivo especial limpieza pusimos marcha hace semanas plan cuenta refuerzo personal maquinaria mantendrá temporada verano ayto estepona" de la cuenta *@AytoEstepona*.

Por otro lado, se ha identificado que algunos subtemas presentan resultados menos precisos, como el caso de "nevadas y fuertes lluvias". En este subtema, se han incluido tweets que hacen referencia principalmente a excursiones a Sierra Nevada. La inclusión de estos tweets no es correcta, ya que el subtema se refiere específicamente a fenómenos climáticos de nevadas y fuertes lluvias y su impacto en la sostenibilidad.

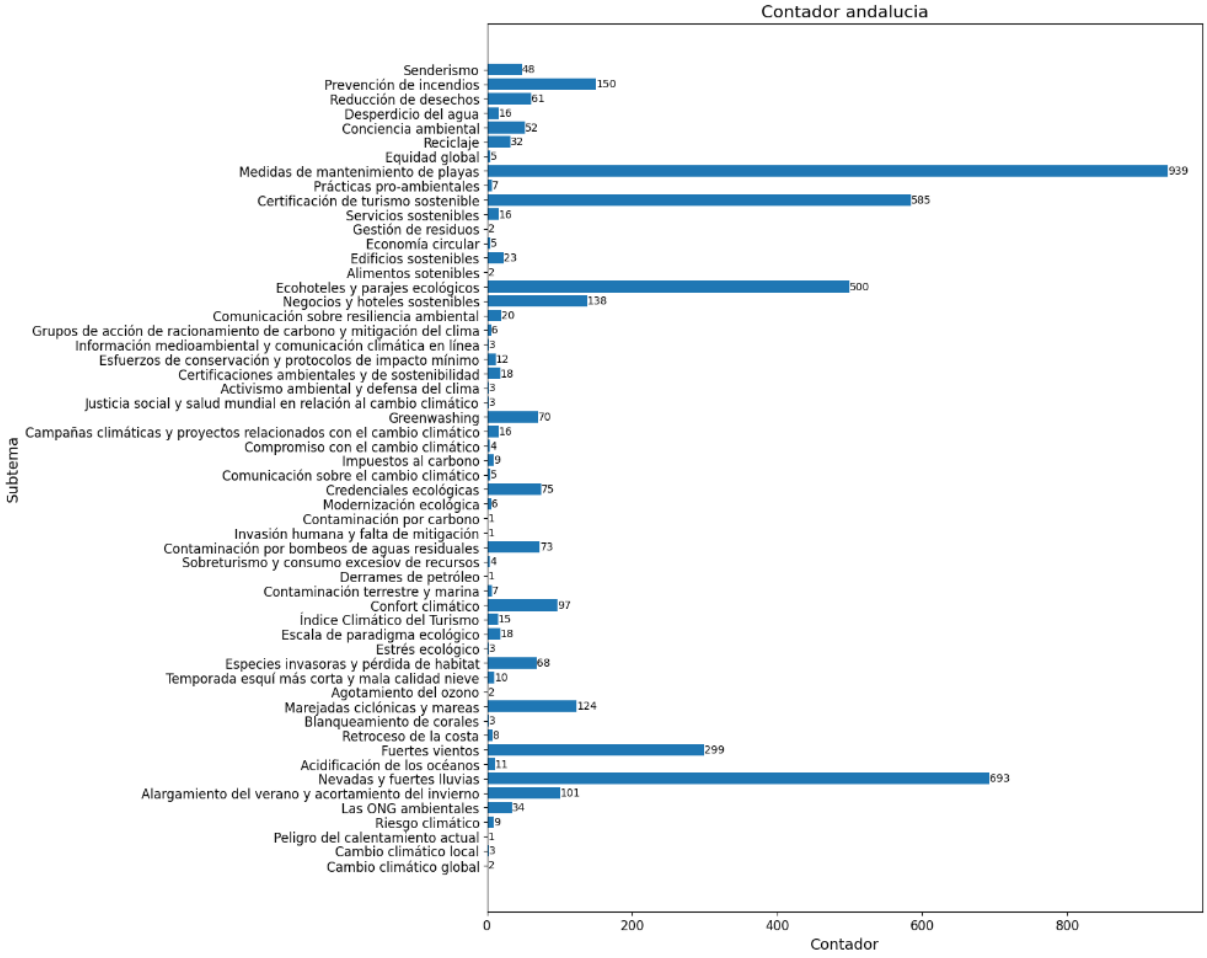

Figura 4: Número de tweets que abarcan los diferentes subtemas en la comunidad autónoma de Andalucía

En el análisis de subtemas en los tweets sobre sostenibilidad en la Comunidad Valenciana, ilustrado en la *Figura 5*, se ha encontrado que el subtema más destacado es "Certificación de turismo sostenible". Esto sugiere que existe un enfoque significativo en la promoción y adopción de prácticas turísticas sostenibles en la región. Se puede ver reflejado en varios tweets como "peñíscola prepara estrategia turística destino enfocada inteligencia turística calidad sostenibilidad noticia completa ajuntament peníscola" de la cuenta *@ajuntpeniscola*.

El subtema "certificado de turismo sostenible" se refiere a la certificación y acreditación de establecimientos turísticos que cumplen con criterios específicos de sostenibilidad. Estos criterios pueden abarcar aspectos como la gestión de residuos, la eficiencia energética, la conservación del entorno natural y la sensibilización sobre prácticas sostenibles.

Sin embargo, se han identificado posibles errores en el análisis de subtemas. En el subtema "ecohoteles y parajes ecológicos", se ha observado que en ocasiones se incluyen tweets que tratan sobre posibles actividades turísticas en la región, pero que no están directamente relacionadas con la sostenibilidad. Estos tweets se refieren a aspectos como promociones de hoteles o atracciones turísticas sin un enfoque específico en prácticas sostenibles. Por ejemplo, “mal tiempo ponemos buena cara to enjoy fabulous massage and spa turismo Benidorm” de la cuenta *@BenidormAyto.*

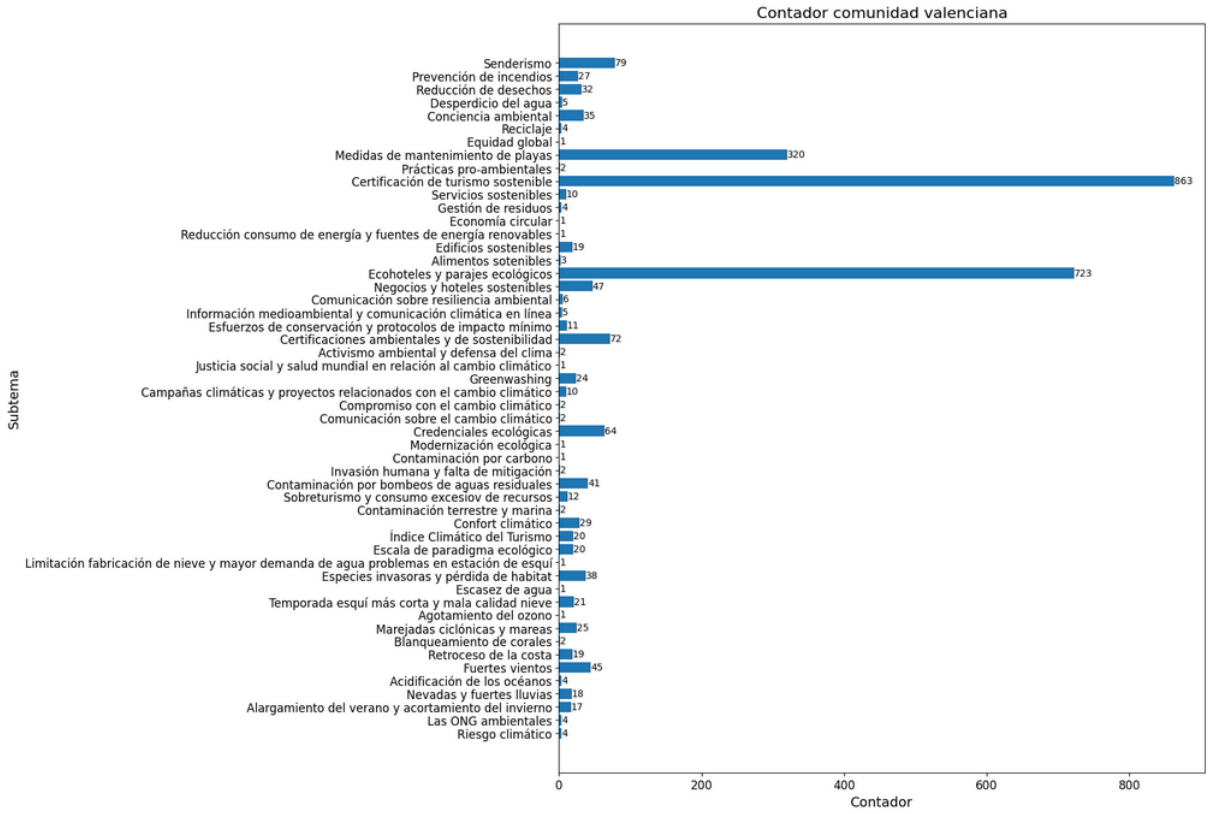

Figura 5: Número de tweets que abarcan los diferentes subtemas en la comunidad valenciana.

La *Figura 6* muestra los subtemas en Canarias. En ella se han identificado dos subtemas destacados: "medidas de mantenimiento de playas" y "ecohoteles y parajes ecológicos". Estos subtemas reflejan la importancia que se le atribuye a la conservación de las playas y al desarrollo de alojamientos turísticos sostenibles y respetuosos con el entorno natural en el archipiélago canario.

El subtema "medidas de mantenimiento de playas" indica que existe una preocupación por garantizar la limpieza y el cuidado de las playas en Canarias. Por otro lado, el subtema "ecohoteles y parajes ecológicos" pone de manifiesto el interés por desarrollar alojamientos turísticos que cumplan con criterios de sostenibilidad y conservación del entorno natural. Un caso podría ser el tweet “conoces proyecto mogán ha sido el único municipio gran canaria

recibir subvención europea proyecto fomentar ecoturismo turismo activo visita página web oficial ayuntamiento mogán" de la cuenta *@MunicipioMogan*.

Sin embargo, al igual que en los análisis anteriores, se han identificado problemas en el subtema "ecohoteles y parajes ecológicos". En algunos casos, se han encontrado tweets que no están directamente relacionados con la sostenibilidad, sino que se centran en experiencias turísticas generales en la región. Estos tweets pueden referirse a actividades de ocio y viajes sin un enfoque específico en la sostenibilidad y la conservación del entorno. Un ejemplo podría ser "visitamos beach club reciente costa adeje fiji es el sitio ideal disfrutar buen coctel camas balinesas vistas gomera servicio impecable ambiente relajado divertido costa Adeje" de la cuenta *@costa_adeje*.

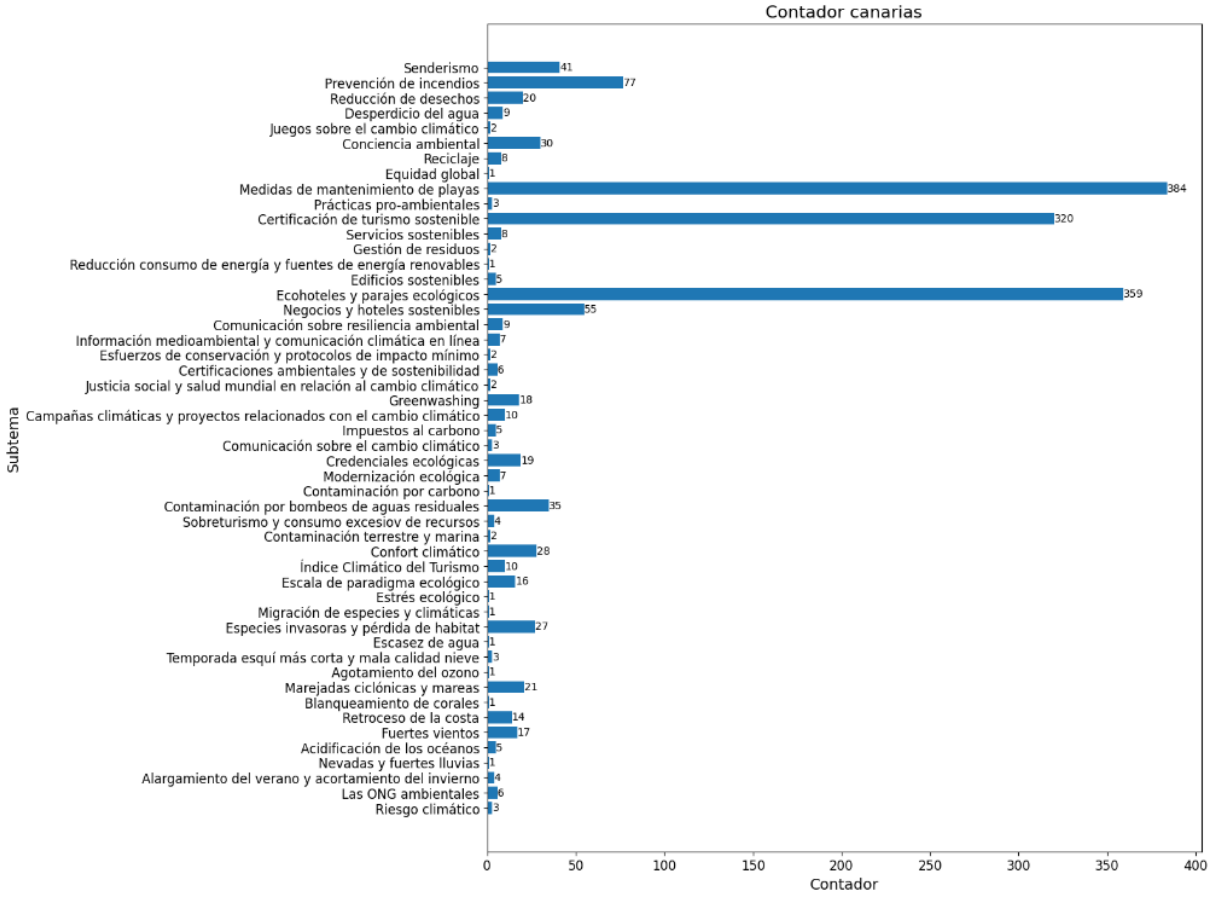

Figura 6: Número de tweets que abarcan los diferentes subtemas en las Islas Canarias

En el análisis de tweets sobre sostenibilidad en Murcia y en las Islas Baleares, *Figura 7 y 8* respectivamente, se observa que, debido al número limitado de cuentas analizadas en estas regiones, existe una menor cantidad de tweets relacionados con la sostenibilidad en comparación con otras comunidades autónomas. Sin embargo, a pesar de esta limitación, se destaca que los subtemas más tratados en ambas regiones son similares a los encontrados en los análisis anteriores.

En particular, se observa que los subtemas de "ecohoteles y parajes ecológicos", "certificación de turismo sostenible" y "medidas de mantenimiento de playas" son abordados con frecuencia en los tweets relacionados con la sostenibilidad en las Islas Baleares y en Murcia. Esto indica

que existe una preocupación por promover alojamientos turísticos respetuosos con el medio ambiente, implementar prácticas sostenibles en el sector turístico y garantizar la conservación y limpieza de las playas en ambas regiones.

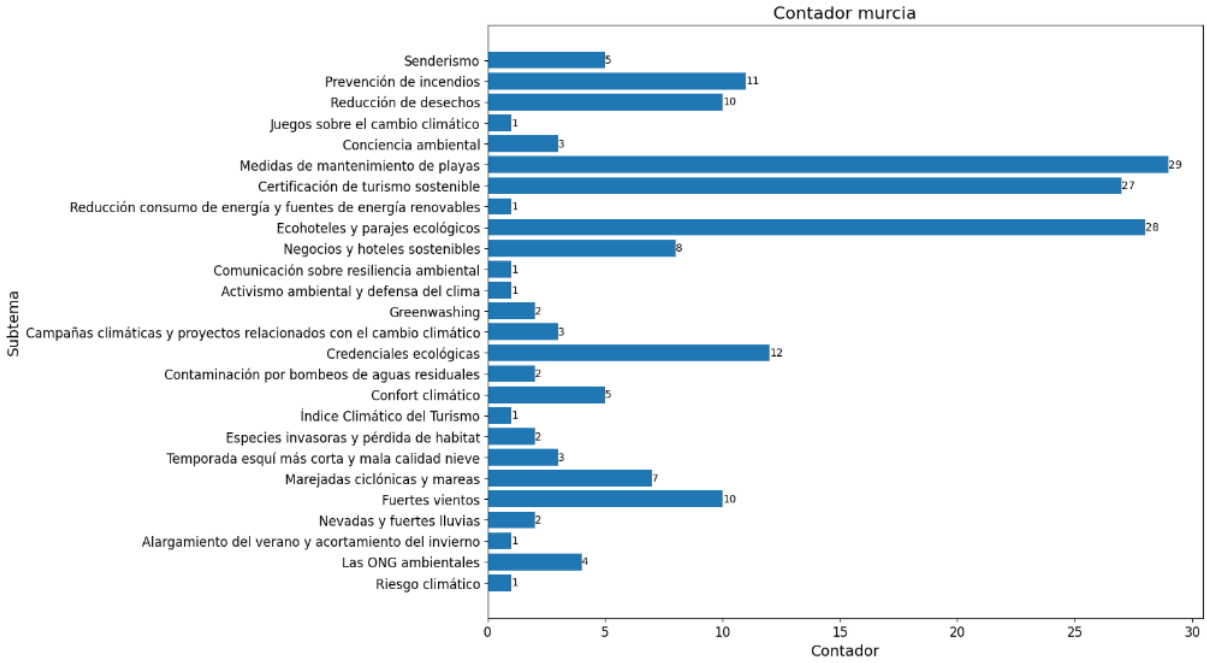

Figura 7: Número de tweets que abarcan los diferentes subtemas en Murcia

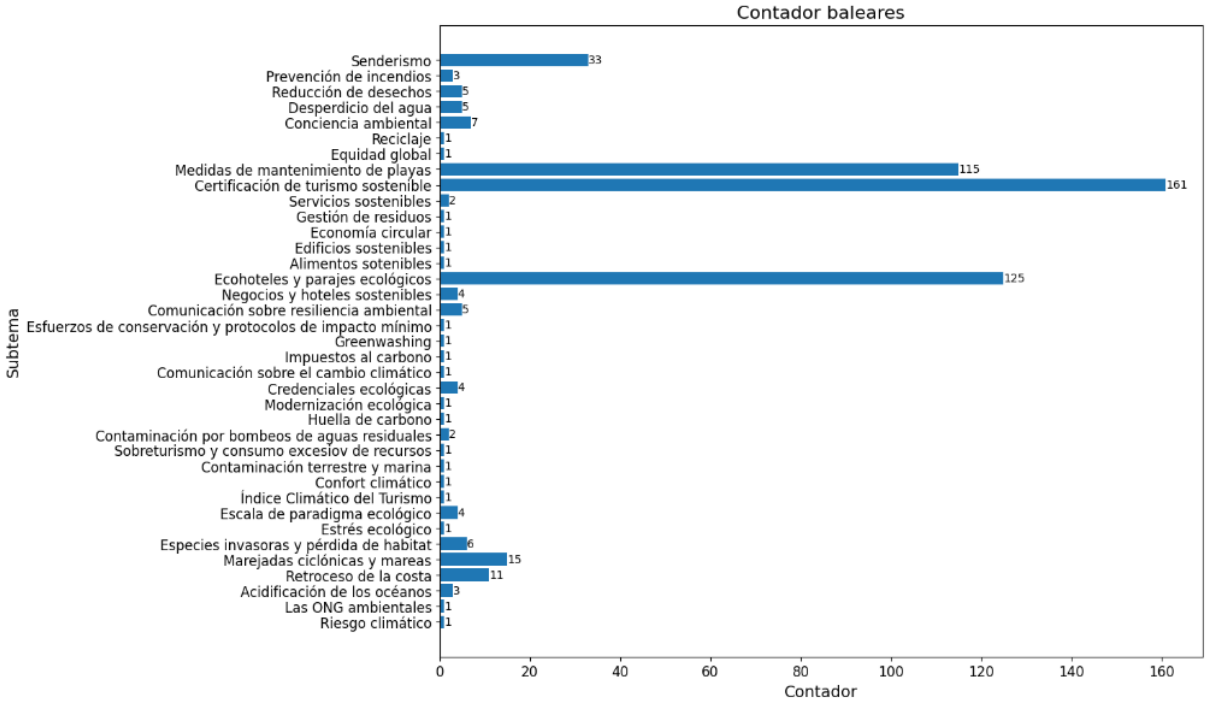

Figura 8: Número de tweets que abarcan los diferentes subtemas en las Islas Baleares

4.2.3 Análisis de subtemas a nivel global

Al analizar todas las cuentas conjuntas, como se refleja en la *Figura 9*, es evidente que se mantiene la misma tendencia observada al analizarlas por comunidad autónoma. Los principales subtemas que se tratan en los tweets relacionados con la sostenibilidad en las cuentas de la costa mediterránea y las Islas Canarias son el "certificado de turismo sostenible", los "ecohoteles y parajes ecológicos" y las "medidas de mantenimiento de playas".

Estos subtemas son relevantes en estas localizaciones debido a la promoción de prácticas sostenibles en el sector turístico, la presencia de entornos naturales destacados y la importancia de mantener la calidad de las playas. Estos temas reflejan los esfuerzos por impulsar un turismo más sostenible y respetuoso con el medio ambiente en estas regiones.

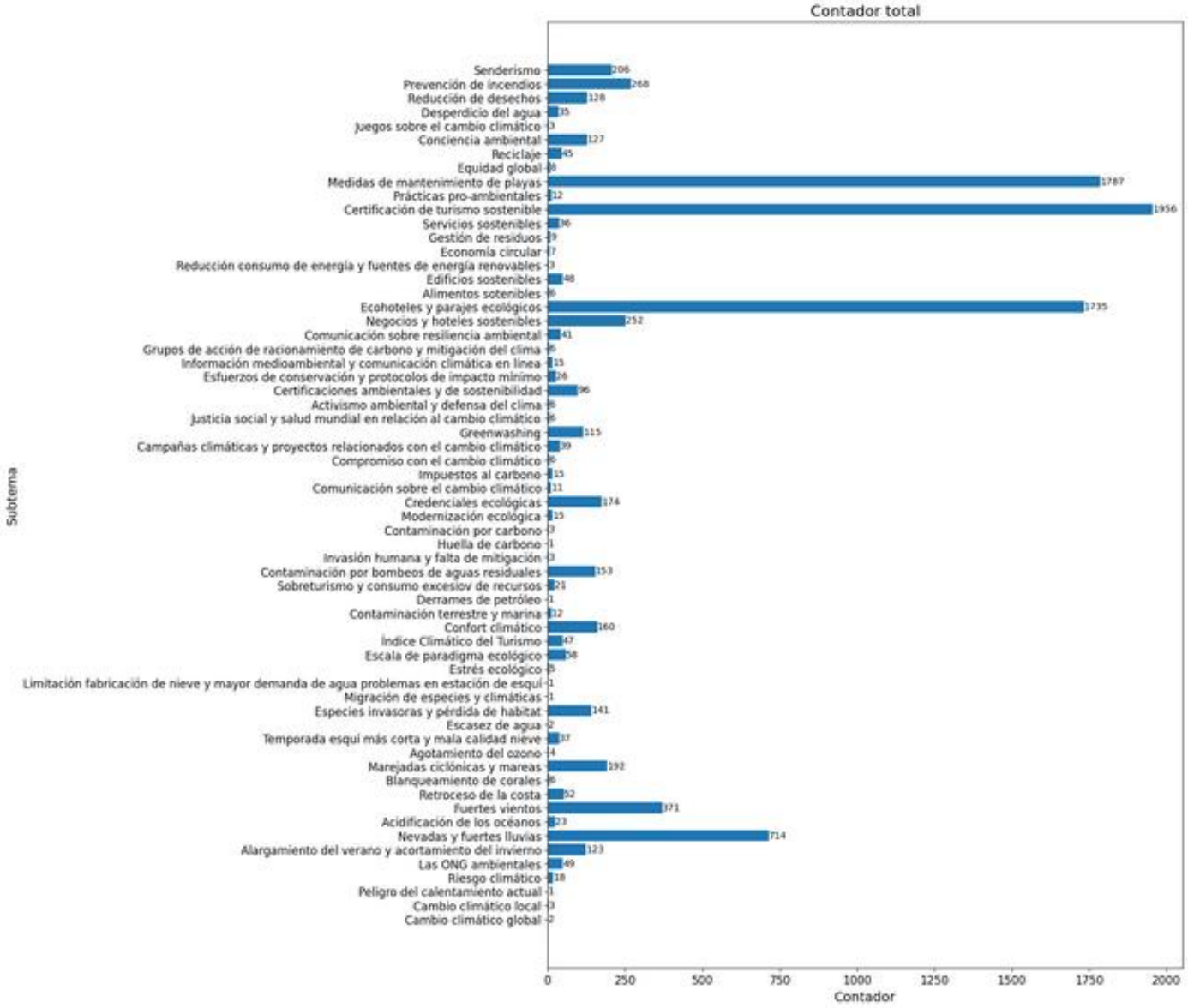

Figura 9: Número de tweets que abarcan los diferentes subtemas en todas las cuentas analizadas

5 CONCLUSIONES Y TRABAJO FUTURO

- El análisis reveló que la mayoría de las cuentas analizadas presentan una baja frecuencia de tweets relacionados con la sostenibilidad, en torno al 10%.
- Se determinó que los subtemas más recurrentes en los tweets analizados sobre sostenibilidad en el turismo son el "certificado de turismo sostenible", "ecohoteles y parajes ecológicos" y "medidas de mantenimiento de playas". Estos subtemas reflejan la importancia de la certificación, el enfoque en la sostenibilidad en la infraestructura hotelera y las acciones para preservar y proteger las playas como atractivos turísticos.
- En el análisis por comunidad autónoma y por cuenta, se observaron patrones consistentes con respecto a los subtemas más tratados.

Es importante destacar que varios factores pueden afectar la precisión y exhaustividad de los resultados obtenidos. En primer lugar, la selección de las frases de referencia desempeña un papel crítico, ya que determina la base para la comparación y clasificación de los tweets analizados. Asimismo, la elección adecuada del threshold es esencial para equilibrar la precisión y la exhaustividad de los resultados, puesto que afecta la inclusión o exclusión de tweets relevantes. Además, el modelo de procesamiento del lenguaje natural utilizado influye en la comprensión semántica y contextual de los textos analizados.

Con el objetivo de mejorar los resultados y la calidad del análisis, se sugieren varias áreas de trabajo futuro. En primer lugar, se recomienda explorar la posibilidad de utilizar frases de referencia más especializadas y adaptadas a los objetivos específicos del proyecto, lo que podría aumentar la precisión y relevancia de los resultados. Además, es fundamental realizar pruebas y ajustes exhaustivos del umbral para encontrar un equilibrio adecuado entre la inclusión de tweets relevantes y la minimización de los falsos positivos. Por último, se plantea la posibilidad de preentrenar modelos con conceptos de sostenibilidad específicos, lo que podría mejorar la capacidad del sistema para identificar y clasificar de manera más precisa los tweets relacionados con la sostenibilidad en el turismo.

En conclusión, el análisis realizado en este proyecto ha demostrado el valor y el potencial de BERT como una herramienta efectiva en el procesamiento del lenguaje natural y la comprensión de textos. Su capacidad para capturar el contexto y el significado de las palabras ha permitido identificar y clasificar con precisión tweets relevantes en el ámbito específico del turismo sostenible a través de las redes sociales. Es evidente que BERT ofrece diversas posibilidades de aplicación en diferentes contextos y sectores, y su relevancia en el futuro está destinada a crecer. Sin embargo, es importante tener en cuenta que su implementación requiere una adaptación adecuada a cada caso en particular.

6 AUTOEVALUACIÓN

Durante el desarrollo de este proyecto, he adquirido valiosos conocimientos y habilidades relacionadas con la investigación, análisis de datos y comunicación de resultados en el ámbito de la sostenibilidad en el turismo a través de las redes sociales. Este proceso ha sido fundamental para alcanzar los objetivos planteados y obtener una sólida base de conocimientos técnicos en esta área específica.

Uno de los aspectos más destacados de mi aprendizaje ha sido el análisis de textos utilizando modelos de procesamiento del lenguaje natural, como BERT. He profundizado en la comprensión de cómo estos modelos avanzados pueden capturar el contexto y el significado de las palabras en un texto, lo que resulta esencial para identificar y extraer información no solo tweets sino también cualquier otro elemento del lenguaje. También comprendí que existen diferentes variantes de BERT preentrenadas, cada una con sus características y especializaciones. Esta habilidad me ha permitido realizar un análisis preciso y detallado de los datos recopilados, y obtener información significativa sobre los subtemas y tendencias en las redes sociales relacionadas con la sostenibilidad turística.

Además, he desarrollado una sólida comprensión de técnicas clave, como la similitud coseno, que me ha permitido comparar los tweets analizados con los conceptos de referencia y evaluar su relevancia. Asimismo, he adquirido experiencia en la selección de umbrales adecuados para determinar qué tweets se consideran más relevantes en términos de sostenibilidad, asegurando así un análisis preciso y significativo de los datos.

En resumen, los conocimientos y habilidades adquiridos a lo largo de este proyecto en el análisis de textos mediante modelos de procesamiento del lenguaje natural y diversas técnicas de tratamiento de datos me proporcionan una base sólida y transferible para futuros proyectos y trabajos en el ámbito de la investigación y análisis de datos. Estas competencias me permitirán realizar análisis rigurosos, obtener información relevante y tomar decisiones fundamentadas en diversos contextos relacionados con mi futuro profesional.

REFERENCIAS

A Look at Precision, Recall, and F1-Score | by Teemu Kanstrén | Towards Data Science. (n.d.). Retrieved May 16, 2023, from https://towardsdatascience.com/a-look-at-precision-recall-and-f1-score-36b5fd0dd3ec

Adewumi, T., Liwicki, F., & Liwicki, M. (2022). Word2Vec: Optimal hyperparameters and their impact on natural language processing downstream tasks. *Open Computer Science*, *12*(1), 134–141. https://doi.org/10.1515/COMP-2022-0236/MACHINEREADABLECITATION/RIS

Almeida, F., & Xexéo, G. (2019). *Word Embeddings: A Survey*. https://arxiv.org/abs/1901.09069v2

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. B. A. (2019). *Publicly Available Clinical BERT Embeddings*. https://arxiv.org/abs/1904.03323v3

BETO: Spanish BERT. Transformer based models are creating… | by elvis | DAIR.AI | Medium. (n.d.). Retrieved May 16, 2023, from https://medium.com/dair-ai/beto-spanish-bert-420e4860d2c6

Cañete, J., Donoso, S., Bravo-Marquez, F., Carvallo, A., & Araujo, V. (2022). ALBETO and DistilBETO: Lightweight Spanish Language Models. *2022 Language Resources and Evaluation Conference, LREC 2022*, 4291–4298. https://arxiv.org/abs/2204.09145v2

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). *What Does BERT Look At? An Analysis of BERT's Attention*. 276–286. https://doi.org/10.18653/v1/w19-4828

Compagnone, M. R., & Fiorentino, G. (2018). Tripadvisor and Tourism: The Linguistic Behaviour of Consumers in the Tourism Industry 2.0. *Strategies of Adaptation in Tourist Communication*, 270–294. https://doi.org/10.1163/9789004359574_015

DeGenaro, D., & Kalita, J. (2022). *CAMeMBERT: Cascading Assistant-Mediated Multilingual BERT*. https://arxiv.org/abs/2212.11456v1

Della Corte, V., Del Gaudio, G., Sepe, F., & Sciarelli, F. (2019). Sustainable Tourism in the Open Innovation Realm: A Bibliometric Analysis. *Sustainability 2019, Vol. 11, Page 6114*, *11*(21), 6114. https://doi.org/10.3390/SU11216114

Etzion, D. (2018). Management for sustainability. *Nature Sustainability 2018 1:12*, *1*(12), 744–749. https://doi.org/10.1038/s41893-018-0184-z

Gai, T., Cao, M., Chiclana, F., Zhang, Z., Dong, Y., Herrera-Viedma, E., & Wu, J. (2023). Consensus-trust Driven Bidirectional Feedback Mechanism for Improving Consensus in Social Network Large-group Decision Making. *Group Decision and Negotiation*, *32*(1), 45–74. https://doi.org/10.1007/S10726-022-09798-7/METRICS

Guerreiro, C., Viegas, M., & Guerreiro, M. (2019). Social Networks and Digital Influencers: Their Role in Customer Decision Journey in Tourism. *Journal of Spatial and Organizational Dynamics*, *7*(3), 240–260. https://www.jsod-cieo.net/journal/index.php/jsod/article/view/198

Guo, S., Zheng, Q., Zhang, L., & Wang, P. (2022). Long-form text matching with word vector clustering and graph convolution. *Proceedings - 2022 International Conference on Machine Learning and Knowledge Engineering, MLKE 2022*, 327–332. https://doi.org/10.1109/MLKE55170.2022.00069

Guo, Y., Jiang, J., & Li, S. (2019). A Sustainable Tourism Policy Research Review. *Sustainability 2019, Vol. 11, Page 3187*, *11*(11), 3187. https://doi.org/10.3390/SU11113187

Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., Silveira-Ocampo, J., Carrino, C. P., Gonzalez-Agirre, A., Armentano-Oller, C., Rodriguez-Penagos, C., & Villegas, M. (2021). *MarIA: Spanish Language Models*. https://doi.org/10.26342/2022-68-3

Hernández-Méndez, J., Muñoz-Leiva, F., Liébana-Cabanillas, F. J., & Marchitto, M. (2016). Análisis de la eficacia publicitaria y usabilidad en herramientas Travel 2.0. Un estudio experimental a través de la técnica de eye-tracking. *Tourism & Management Studies*, *12*(2), 7–17. https://doi.org/10.18089/tms.2016.12202

Higgins-Desbiolles, F. (2018). Sustainable tourism: Sustaining tourism or something more? *Tourism Management Perspectives*, *25*, 157–160. https://doi.org/10.1016/J.TMP.2017.11.017

hiiamsid/sentence_similarity_spanish_es · Hugging Face. (n.d.). Retrieved May 16, 2023, from https://huggingface.co/hiiamsid/sentence_similarity_spanish_es

Huang, W., Cheng, X., Wang, T., & Chu, W. (2019). BERT-Based Multi-head Selection for Joint Entity-Relation Extraction. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11839 LNAI*, 713–723. https://doi.org/10.1007/978-3-030-32236-6_65/COVER

Kanade, A., Maniatis, P., Balakrishnan, G., & Shi, K. (2020). *Learning and Evaluating Contextual Embedding of Source Code* (pp. 5110–5121). PMLR. https://proceedings.mlr.press/v119/kanade20a.html

Kapera, I. (2018). Sustainable tourism development efforts by local governments in Poland. *Sustainable Cities and Society*, *40*, 581–588. https://doi.org/10.1016/J.SCS.2018.05.001

Kim, T., Yoo, K. M., & Lee, S. G. (2021). Self-Guided Contrastive Learning for BERT Sentence Representations. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2528–2540. https://doi.org/10.18653/v1/2021.acl-long.197

Koroteev, M. V. (2021). *BERT: A Review of Applications in Natural Language Processing and Understanding*. https://arxiv.org/abs/2103.11943v1

Li, Z., Lin, H., Shen, C., Zheng, W., Yang, Z., & Wang, J. (2020). Cross2Self-attentive Bidirectional Recurrent Neural Network with BERT for Biomedical Semantic Text Similarity. *Proceedings - 2020 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2020*, 1051–1054. https://doi.org/10.1109/BIBM49941.2020.9313452

Lian, S. L., Sun, X. J., Yang, X. juan, & Zhou, Z. K. (2020). The effect of adolescents' active social networking site use on life satisfaction: The sequential mediating roles of positive feedback and relational certainty. *Current Psychology*, *39*(6), 2087–2095. https://doi.org/10.1007/S12144-018-9882-Y/METRICS

Liu, Z., Lin, Y., & Sun, M. (2020). Representation Learning and NLP. *Representation Learning for Natural Language Processing*, 1–11. https://doi.org/10.1007/978-981-15-5573-2_1

MacKenzie, N., & Gannon, M. J. (2019). Exploring the antecedents of sustainable tourism development. *International Journal of Contemporary Hospitality Management*, *31*(6), 2411–2427. https://doi.org/10.1108/IJCHM-05-2018-0384/FULL/XML

Mariani, M. M., Di Felice, M., & Mura, M. (2016). Facebook as a destination marketing tool: Evidence from Italian regional Destination Management Organizations. *Tourism Management*, *54*, 321–343. https://doi.org/10.1016/J.TOURMAN.2015.12.008

Muller, B., Anastasopoulos, A., Sagot, B., & Seddah, D. (2020). When Being Unseen from mBERT is just the Beginning: Handling New Languages With Multilingual Language Models. *NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, 448–462. https://doi.org/10.18653/v1/2021.naacl-main.38

Nozza, D., Bianchi, F., & Hovy, D. (n.d.). *HONEST: Measuring Hurtful Sentence Completion in Language Models*. https://github.com/MilaNLProc/

Ortiz, L., & González Sánchez, R. (2014). La social media como herramienta de mejora de la experiencia turística: Una aplicación al sector hotelero. *RITUR: Revista Iberoamericana de Turismo, ISSN-e 2236-6040, Vol. 4, Nº. 1, 2014, Págs. 16-34*, *4*(1), 16–34. https://dialnet.unirioja.es/servlet/articulo?codigo=7480349&info=resumen&idioma=SPA

Pan, S. Y., Gao, M., Kim, H., Shah, K. J., Pei, S. L., & Chiang, P. C. (2018). Advances and challenges in sustainable tourism toward a green economy. *Science of The Total Environment*, *635*, 452–469. https://doi.org/10.1016/J.SCITOTENV.2018.04.134

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems*, *32*. https://arxiv.org/abs/1912.01703v1

Plaza-del-Arco, F. M., Molina-González, M. D., Ureña-López, L. A., & Martín-Valdivia, M. T. (2021). Comparing pre-trained language models for Spanish hate speech detection. *Expert Systems with Applications*, *166*, 114120. https://doi.org/10.1016/J.ESWA.2020.114120

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 3982–3992. https://doi.org/10.18653/v1/d19-1410

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, *8*, 842–866. https://doi.org/10.1162/TACL_A_00349/96482/A-PRIMER-IN-BERTOLOGY-WHAT-WE-KNOW-ABOUT-HOW-BERT

Ruhanen, L., Moyle, C. lee, & Moyle, B. (2019). New directions in sustainable tourism research. *Tourism Review*, *74*(2), 245–256. https://doi.org/10.1108/TR-12-2017-0196/FULL/XML

Sánchez Jimenéz, M. Á. (2018). Análisis del retorno de la inversión (ROI) de la actividad en las redes sociales de las provincias andaluzas como destino turístico. *Pasos: Revista de Turismo y Patrimonio Cultural, ISSN-e 1695-7121, Vol. 16, Nº. 4, 2018, Págs. 1067-1088*, *16*(4), 1067–1088. https://dialnet.unirioja.es/servlet/articulo?codigo=6623925&info=resumen&idioma=SPA

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. https://arxiv.org/abs/1910.01108v4

Shimanaka, H., Kajiwara, T., & Komachi, M. (2019). *Machine Translation Evaluation with BERT Regressor*. https://arxiv.org/abs/1907.12679v1

Song, X., Salcianu, A., Song, Y., Dopson, D., & Zhou, D. (2020). Fast WordPiece Tokenization. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, 2089–2103. https://doi.org/10.18653/v1/2021.emnlp-main.160

Soyusiawaty, D., & Zakaria, Y. (2018). Book data content similarity detector with cosine similarity (case study on digilib.uad.ac.id). *Proceeding of 2018 12th International Conference on Telecommunication Systems, Services, and Applications, TSSA 2018*. https://doi.org/10.1109/TSSA.2018.8708758

SPANISH PRE-TRAINED BERT MODEL AND EVALUATION DATA. (n.d.).

Stokłosa, Ł., Marchiori, E., & Cantoni, L. (2019). Understanding the web maturity of Polish DMOs. *Journal of Destination Marketing & Management*, *11*, 192–199. https://doi.org/10.1016/J.JDMM.2018.01.010

stsb_multi_mt · Datasets at Hugging Face. (n.d.). Retrieved May 16, 2023, from https://huggingface.co/datasets/stsb_multi_mt

Tahmid, M., Laskar, R., Huang, J., & Hoque, E. (2020). *Contextualized Embeddings based Transformer Encoder for Sentence Similarity Modeling in Answer Selection Task* (pp. 5505–5514). https://aclanthology.org/2020.lrec-1.676

Teixeira, S., Alías, F., Cardeñoso-Payo, V., Escudero-Mancebo, D., González-Ferreras, C., Fernández-Martínez, F., Luna-Jiménez, C., Kleinlein, R., Griol, D., Callejas, Z., & Montero, J. M. (2022). Fine-Tuning BERT Models for Intent Recognition Using a Frequency Cut-Off Strategy for Domain-Specific Vocabulary Extension. *Applied Sciences 2022, Vol. 12, Page 1610*, *12*(3), 1610. https://doi.org/10.3390/APP12031610

Valeri, M., & Baggio, R. (2021). Social network analysis: organizational implications in tourism management. *International Journal of Organizational Analysis*, *29*(2), 342–353. https://doi.org/10.1108/IJOA-12-2019-1971/FULL/PDF

van Aken, B., Herrmann, S., & Löser, A. (2021). What Do You See in this Patient? Behavioral Testing of Clinical NLP Models. *ClinicalNLP 2022 - 4th Workshop on Clinical Natural Language Processing, Proceedings*, 63–73. https://doi.org/10.18653/v1/2022.clinicalnlp-1.7

Varshney, D., & Vishwakarma, D. K. (2021). A review on rumour prediction and veracity assessment in online social network. *Expert Systems with Applications*, *168*, 114208. https://doi.org/10.1016/J.ESWA.2020.114208

Wang, B., & Kuo, C. C. J. (2020). SBERT-WK: A Sentence Embedding Method by Dissecting BERT-Based Word Models. *IEEE/ACM Transactions on Audio Speech and Language Processing*, *28*, 2146–2157. https://doi.org/10.1109/TASLP.2020.3008390

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. Le, Gugger, S., … Rush, A. M. (2019). *HuggingFace's Transformers: State-of-the-art Natural Language Processing*. https://arxiv.org/abs/1910.03771v5

Xiang, Z. (2018). From digitization to the age of acceleration: On information technology and tourism. *Tourism Management Perspectives*, *25*, 147–150. https://doi.org/10.1016/J.TMP.2017.11.023

Zhan, J., Mao, J., Liu, Y., Zhang, M., & Ma, S. (2020). An Analysis of BERT in Document Ranking. *SIGIR 2020 - Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1941–1944. https://doi.org/10.1145/3397271.3401325

ANEXOS

Anexo 1: Evaluación de los parámetros de interés de las cuentas analizadas

	Destino	Twitter Oficina Turismo	Idioma	Estado	Trata tema	Twitter Ayuntamiento	Idioma	Estado	Trata tema
Andalucía	Mojácar	@MojacarTurismo	castellano	no muy activa	parcialmente				
	Roquetas de Mar	@TurismoRoquetas	castellano	casi inactiva	nada	@AytoRoquetas	castellano	activa	mucho
	Conil de la Frontera	@CadizTurismo	castellano	muy activa	nada				
	Chiclana de la Frontera	@Turismochiclana	castellano	activa	nada	@ayto_chiclana	castellano	muy activa	bastante
	Puerto de Santa María	@TurismoElPuerto	castellano	muy activa	parcialmente	@ElPuerto	castellano	muy activa	bastante
	Tarifa	@CadizTurismo	castellano	muy activa	nada	@puertotarifa	castellano	no muy activa	mucho
	Isla Cristina					@Ayto_ic	castellano	activa	bastante
	Benalmádena	@TuBenalmadena	castellano	muy activa	nada	@BenalmadenaAyto	castellano	muy activa	parcialmente
	Estepona	@EsteponaTurAyto	castellano/inglés	muy activa	parcialmente	@AytoEstepona	castellano	muy activa	mucho
	Fuengirola	@turismofuengi	castellano	casi inactiva	nada	@fuengirola	castellano	muy activa	mucho
	Marbella	@MARBELLATURISMO	inglés	activa	nada	@Ayto_Marbella	castellano	muy activa	bastante
	Torremolinos	@_TTorremolinos	castellano/inglés	no muy activa	parcialmente	@Torremolinos_On	castellano	muy activa	mucho
	Almuñécar	@VisitAlmunecar	castellano	no muy activa	parcialmente	@AlmunecarAyto	castellano	muy activa	mucho
Baleares	Alcudia	@TurismeBalears	castellano	muy activa	bastante	@ajtalcudia	catalán	activa	mucho
	Calviá	@TurismeBalears	castellano	muy activa	bastante	@_Calvia	catalán	muy activa	mucho
	Capdepera	@TurismeBalears	castellano	muy activa	bastante	@ajcapdepera	castellano/catalán	muy activa	bastante
	Llucmajor	@VisitIlucmajor	inglés	no muy activa	nada	@ajllucmajor	castellano/catalán	muy activa	parcialmente
	Manacor	@VisitManacor	castellano/catalán	no muy activa	nada	@ajManacor	catalán	muy activa	bastante
	Muro	@TurismeBalears	castellano	muy activa	bastante	@ajuntament_muro	catalán	muy activa	mucho
	Palma	@passionforpalma	castellano/inglés	activa	parcialmente	@ajuntpalma	catalán	muy activa	mucho
	Santa Margalida	@TurismeBalears	castellano	muy activa	bastante	@AjStaMargalida	catalán	casi inactiva	nada
	Sant Llorenc des Cardassar	@Visitcalamillor	castellano/catalán/inglés	activa	bastante	@Aj_SantLlorenc	castellano/catalán	casi inactiva	bastante
	Santanyi	@TurismeBalears	castellano	muy activa	bastante	@ajsantanyi	castellano	activa	bastante
	Ciutadella	@TurismoMenorca	castellano	activa	nada	@AjCiutadella	catalán	muy activa	mucho
	Eivissa	@EivissaIbiza	castellano/catalán/inglés	activa	parcialmete	@ajeivissa	catalán	activa	bastante
	Sant Antoni de Portmany	@VisitSantAntoni	castellano/catalán/inglés	activa	nada	@sant_antoni	catalán	muy activa	mucho
	Sant Josep de Sa Talaia	@Santjosepibiza	castellano/catalán/inglés	activa	nada	@DeTalaia	catalán	muy activa	parcialmente
	Santa Eulalia des Rius	@visitSE_ibiza	castellano/inglés	activa	nada	@Santa_Eularia	castellano/catalán	muy activa	bastante
Canarias	Mogán	@turismogc	castellano	muy activa	parcialmente	@MunicipioMogan	castellano	muy activa	bastante
	San Bartolomé de Tirajana	@turismogc	castellano	muy activa	parcialmente	@AytoSBT	castellano	muy activa	mucho
	Adeje	@costa_adeje	castellano/inglés	muy activa	mucho	@adeje	castellano	muy activa	mucho
	Arona	@TurismoArona	castellano	muy activa	parcialmente	@AytoArona	castellano	activa	bastante
	Puerto de la Cruz	@VisitPtoCruz	castellano	muy activa	nada	@puertodelacruz	castellano	muy activa	mucho
	Teguise	@TurismoLZT	castellano/inglés	activa	nada	@AyunTeguise	castellano	activa	parcialmente
	Tías	@TurismoLZT	castellano/inglés	activa	nada	@AyunDeTias	castellano	muy activa	parcialmente
	Yaiza	@yaizaturismo	castellano	casi inactiva	nada	@yaizateinforma	castellano	no muy activa	nada
	La Oliva	@iFuerteventura	castellano	activa	nada	@AytoLaOliva	castellano	activa	bastante
	Pájara	@iFuerteventura	castellano	activa	nada	@aytopajara	castellano	activa	bastante

Cataluña	Calella	@calella_bcn	castellano/catalán/inglés	activa	parcialmente	@calellaesmes	catalán	muy activa	bastante
	Santa Susana	@SSusannaturisme	castellano/catalán/inglés	muy activa	parcialmente	@stasusanna_cat	catalán	muy activa	parcialmente
	Sitges	@TurismeDeSitges	castellano/catalán/inglés	muy activa	nada	@AjSitges	catalán	muy activa	bastante
	Castell-Platja D'Aro	@platjaroturisme	catalán	casi inactiva	nada				
	Lloret de Mar	@lloretturisme	catalán	muy activa	parcialmente	@Lloret_de_Mar	catalán	muy activa	bastante
	Roses	@VisitRoses	castellano/catalán/inglés	muy activa	nada	@AjRoses	catalán	muy activa	bastante
	Cambrils	@CambrilsTurisme	castellano/catalán	casi inactiva	nada	@ajcambrils	catalán	muy activa	mucho
	Vila-Seca	@lapinedaplatja	castellano/catalán	activa	parcialmente	@AjVilaseca	catalán	activa	parcialmente
	Salou	@visitsalou	castellano/catalán	muy activa	parcialmente	@AjuntamentSalou	castellano/catalán	muy activa	mucho
Com. Valenciana	Benidorm	@visitbenidorm	castellano/inglés	muy activa	bastante	@BenidormAyto	castellano	muy activa	mucho
	Denia	@DeniaTurismo	castellano	activa	bastante	@DeniaAyun	castellano	casi inactiva	nada
	Benicasim	@tmobenicassim	castellano	muy activa	bastante	@AytoBenicassim	castellano	activa	mucho
	Peñíscola	@_peniscola	castellano	muy activa	parcialmente	@ajuntpeniscola	castellano/catalán	muy activa	bastante
	Gandía	@visitgandia	castellano/catalán	muy activa	bastante	@aj_gandia	catalán	muy activa	mucho
Murcia	Cartagena	@TuriCartagenaES	castellano/inglés	muy activa	nada	@AytoCartagenaES	castellano	muy activa	mucho

Estado
muy activa = un par de tweets diarios
activa = 1 tweet diario
no muy activa = un par de tweets mensuales
casi inactiva = más de un año sin tweets nuevos

Trata tema
mucho = varios tweets sobre el tema por mes
bastante = 1 tweet sobre el tema por mes
parcialmente = un par de tweets sobre el tema por año
nada = ningún tweet encontrado sobre el tema

Tabla 9: Listado cuentas de ayuntamientos y oficinas de turismo junto con la evaluación de los parámetros: idioma, estado y trata sostenibilidad.

Anexo 2: Listado de las cuentas analizadas en este proyecto

Destino	Twitter Oficina Turismo	Tweets 2018	Tweets 2019	Tweets 2020	Tweets 2021	Tweets 2022
Roquetas de Mar	@TurismoRoquetas					
Chiclana de la Frontera	@Turismochiclana					
Puerto de Santa María	@TurismoElPuerto					
Isla Cristina						
Estepona	@EsteponaTurAyto					
Fuengirola	@turismofuengi					
Marbella	@MARBELLATURISMO					
Torremolinos	@_TTorremolinos					
Almuñécar	@VisitAlmunecar					
Alcudia	@TurismeBalears	1576	1645	1373	1168	204
Calviá	@TurismeBalears	1576	1645	1373	1168	204
Capdepera	@TurismeBalears	1576	1645	1373	1168	204
Manacor	@VisitManacor					
Muro	@TurismeBalears	1576	1645	1373	1168	204
Palma	@passionforpalma					
Santanyi	@TurismeBalears	1576	1645	1373	1168	204
Ciutadella	@TurismoMenorca					
Eivissa	@Eivissalbiza					
Sant Antoni de Portmany	@VisitSantAntoni					
Santa Eulalia des Rius	@visitSE_ibiza					
Mogán	@turismogc					
San Bartolomé de Tirajana	@turismogc					

Adeje	@costa_adeje	131	60	428	296	22
Arona	@TurismoArona					
Puerto de la Cruz	@VisitPtoCruz					
La Oliva	@iFuerteventura					
Pájara	@iFuerteventura					
Calella	@calella_bcn					
Sitges	@TurismeDeSitges					
Lloret de Mar	@lloretturisme					
Roses	@VisitRoses					
Cambrils	@CambrilsTurisme					
Salou	@visitsalou					
Benidorm	@visitbenidorm	787	1133	1380	1478	447
Denia	@DeniaTurismo	214	237	381	638	110
Benicasim	@tmobenicassim	199	605	886	383	66
Peñíscola	@_peniscola	281	484	353	317	34
Gandía	@visitgandia	1001	249	484	614	169
Cartagena	@TuriCartagenaES					
	sin repetir @TurismeBalears	4189	4413	5285	4894	1052

Destino	Twitter Ayuntamiento	Tweets 2018	Tweets 2019	Tweets 2020	Tweets 2021	Tweets 2022
Roquetas de Mar	@AytoRoquetas	1168	1065	729	860	277
Chiclana de la Frontera	@ayto_chiclana	1927	1675	1225	1161	263
Puerto de Santa María	@ElPuerto	2667	2360	2020	2026	467
Isla Cristina	@Ayto_ic	414	159	133	445	22

Estepona	@AytoEstepona	319	612	716	1042	188
Fuengirola	@fuengirola	2015	1698	1339	1684	391
Marbella	@Ayto_Marbella	1914	1587	1589	1918	418
Torremolinos	@Torremolinos_On	2503	2001	1403	1089	316
Almuñécar	@AlmunecarAyto	2547	2301	2609	2416	420
Alcudia	@ajtalcudia	400	200	221	156	1
Calviá	@_Calvia	239	41	29	260	76
Capdepera	@ajcapdepera	0	0	1274	1883	710
Manacor	@ajManacor	566	395	565	509	156
Muro	@ajuntament_muro	843	610	111	754	184
Palma	@ajuntpalma	717	1106	1645	1665	412
Santanyi	@ajsantanyi	710	309	29	0	0
Ciutadella	@AjCiutadella	253	64	237	181	64
Eivissa	@ajeivissa	703	389	667	551	79
Sant Antoni de Portmany	@sant_antoni	364	370	445	442	115
Santa Eulalia des Rius	@Santa_Eularia	402	367	813	584	131
Mogán	@MunicipioMogan	2080	1985	1999	1698	387
San Bartolomé de Tirajana	@AytoSBT	831	494	369	321	119
Adeje	@adeje	1322	739	543	597	190
Arona	@AytoArona	1912	1336	825	464	94
Puerto de la Cruz	@puertodelacruz	350	280	484	464	140
La Oliva	@AytoLaOliva	193	357	453	194	36
Pájara	@aytopajara	191	250	341	304	69

Calella	@calellaesmes	611	460	626	619	130
Sitges	@AjSitges	3003	2892	2655	2503	630
Lloret de Mar	@Lloret_de_Mar	304	325	368	449	140
Roses	@AjRoses	1232	1125	1081	1060	271
Cambrils	@ajcambrils	1757	1899	2224	2325	593
Salou	@AjuntamentSalou	2405	1727	1594	1686	427
Benidorm	@BenidormAyto	1342	1100	1039	1017	288
Denia	@DeniaAyun					
Benicasim	@AytoBenicassim	1	0	0	232	90
Peñíscola	@ajuntpeniscola	482	484	1416	1271	213
Gandía	@aj_gandia	1649	1380	1194	1240	224
Cartagena	@AytoCartagenaES	1082	1075	324	1598	476
		41418	35217	35334	37668	9207	178677

Tabla 10: Listado cuentas de ayuntamientos y oficinas de turismo junto con la cantidad de tweets de cada año desde el 2018 hasta principios del 2022.

Anexo 3: Concepto referencia:frase referencia

Variabilidad climática:La variabilidad climática puede tener graves consecuencias para la sostenibilidad de la vida humana y del planeta, ya que puede generar sequías, inundaciones y otros eventos climáticos extremos que afectan la producción de alimentos, la calidad del agua y la biodiversidad.

Cambio climático global:El cambio climático global es un problema urgente que requiere medidas a nivel internacional para reducir las emisiones de gases de efecto invernadero y proteger la sostenibilidad del planeta, ya que afecta a la salud humana, la economía y el medio ambiente.

Cambio climático local:El cambio climático local es un fenómeno que puede tener consecuencias graves para la sostenibilidad de las comunidades, especialmente las más vulnerables, ya que puede alterar los patrones de lluvia, aumentar las temperaturas y aumentar el riesgo de desastres naturales.

Calentamiento global:El calentamiento global es una de las principales causas del cambio climático, y su impacto en la sostenibilidad del planeta se puede ver en la acidificación de los océanos, la pérdida de hielo en los polos y el aumento del nivel del mar.

Peligro del calentamiento actual:El calentamiento actual es un fenómeno preocupante que puede tener graves consecuencias para la sostenibilidad del planeta, ya que puede provocar sequías, olas de calor y otros eventos climáticos extremos que afectan la producción de alimentos, la biodiversidad y la calidad del aire.

Riesgo climático:El riesgo climático es una realidad que debemos afrontar para garantizar la sostenibilidad del planeta, y para ello es necesario adaptarnos a los cambios climáticos y reducir nuestra huella de carbono.

Impactos del cambio climático:Los impactos del cambio climático pueden ser devastadores para la sostenibilidad del planeta, ya que pueden provocar sequías, inundaciones, la extinción de especies y la pérdida de biodiversidad.

Puntos críticos del cambio climático:Los puntos críticos del cambio climático son aquellos lugares donde el impacto del cambio climático es más grave y puede tener consecuencias irreversibles para la sostenibilidad del planeta, como la degradación de los arrecifes de coral y la pérdida de glaciares.

Peligro del cambio climático:El peligro del cambio climático es real y puede tener consecuencias graves para la sostenibilidad de la vida humana y del planeta, como el aumento del nivel del mar, la extinción de especies y la disminución de la producción de alimentos.

Escépticos del clima:Los escépticos del clima son aquellos que dudan de la existencia del cambio climático y su impacto en la sostenibilidad del planeta, pero la evidencia científica es clara y debemos tomar medidas urgentes para reducir nuestra huella de carbono y proteger el futuro del planeta.

Las ONG ambientales:Las ONG ambientales trabajan incansablemente para combatir la fatiga por el cambio climático y reducir el efecto invernadero, así como proteger el medio ambiente mundial de los problemas ambientales más apremiantes.

Alargamiento del verano y acortamiento del invierno:El alargamiento del verano y el acortamiento del invierno son signos evidentes del cambio climático, que ha provocado escasez

de agua, fenómenos meteorológicos extremos y olas de calor que aumentan la temperatura y reducen la precipitación.

Nevadas y fuertes lluvias:Las nevadas son cada vez menos frecuentes debido al aumento de la temperatura y la disminución de la capa de nieve, lo que ha agravado las sequías y los riesgos de incendios forestales e inundaciones en algunas regiones.

Acidificación de los océanos:La acidificación de los océanos y la disminución de las precipitaciones están afectando el ambiente marino, lo que está provocando ciclones y tormentas tropicales más intensas y tejidos de tormenta más severos, así como un mayor surf y un aumento de los incendios forestales en las zonas costeras.

Fuertes vientos:Los huracanes son un ejemplo dramático de los fenómenos meteorológicos extremos que están aumentando en intensidad y frecuencia debido al cambio climático global, lo que está exacerbando la pobreza y la desigualdad en todo el mundo.

Retroceso de la costa:El retroceso de la costa y la erosión de la playa son consecuencias evidentes del aumento del nivel del mar, lo que pone de manifiesto la vulnerabilidad climática de las regiones costeras.

Blanqueamiento de corales:El blanqueamiento de corales y los cambios morfológicos que se están produciendo en los ecosistemas marinos son una muestra clara de los impactos del cambio climático en el medio ambiente mundial.

Marejadas ciclónicas y mareas:Las marejadas ciclónicas y las mareas reales cada vez más frecuentes son eventos extremos que están afectando gravemente a las comunidades costeras y aumentando su vulnerabilidad climática.

Agotamiento del ozono:El agotamiento del ozono está teniendo consecuencias biológicas graves, como la pérdida de biodiversidad y la vulnerabilidad de las especies a las radiaciones ultravioleta.

Temporada esquí más corta y mala calidad nieve:La temporada de esquí más corta y la mala calidad de la nieve son síntomas evidentes del impacto del cambio climático en los ecosistemas de montaña, lo que puede tener graves consecuencias para la economía de las regiones que dependen del turismo invernal.

Escasez de agua:La escasez de agua y la mayor demanda de energía son problemas cada vez más acuciantes en muchas regiones del mundo, lo que hace necesario un uso más eficiente de los recursos naturales y una transición hacia fuentes de energía más sostenibles.

Especies invasoras y pérdida de habitat:La presencia de especies invasoras y la pérdida de hábitat son problemas que están afectando gravemente a la biodiversidad y a la interacción entre las especies en muchos ecosistemas terrestres y acuáticos.

Migración de especies y climáticas:La migración de especies y las migraciones climáticas son fenómenos cada vez más frecuentes, que están afectando a los ecosistemas y a la disponibilidad de recursos naturales en muchas regiones del mundo.

Limitación fabricación de nieve y mayor demanda de agua problemas en estación de esquí:La limitación de la fabricación de nieve y la mayor demanda de agua para la producción de nieve artificial son problemas que están poniendo en peligro la sostenibilidad de muchas estaciones de esquí.

Estrés ecológico:El estrés ecológico es un problema que está afectando gravemente a muchas especies y ecosistemas en todo el mundo, lo que hace necesario un enfoque más integrado y sostenible de la gestión de los recursos naturales.

Escala de paradigma ecológico:La escala de paradigma ecológico de las comunidades locales es un factor importante a considerar en los esfuerzos de mitigación del cambio climático.

Índice Climático del Turismo:El Índice Climático del Turismo puede ser una herramienta útil para evaluar la vulnerabilidad del sector turístico frente al cambio climático y guiar la toma de decisiones.

Confort climático:El confort climático de las personas puede verse seriamente comprometido si se produce un aumento en el malestar climático, lo que puede tener implicaciones económicas y sociales.

Contaminación terrestre y marina:La contaminación terrestre y de origen marino es una amenaza para la salud humana y los ecosistemas, y se necesita una gestión adecuada de los residuos para reducir su impacto.

Derrames de petróleo:Los derrames de petróleo son uno de los desastres ambientales más devastadores que pueden ocurrir en los océanos, y pueden tener consecuencias a largo plazo para los ecosistemas marinos.

Sobreturismo y consumo excesiov de recursos:El sobreturismo y el consumo excesivo de recursos son problemas que afectan la sostenibilidad del turismo y la calidad de vida de las comunidades locales.

Huella ambiental:La huella ambiental es una medida útil para evaluar el impacto de nuestras acciones en el medio ambiente y puede ayudar a identificar áreas para reducir nuestra huella ecológica.

Clima antropogénico:El clima antropogénico se refiere a los cambios en el clima causados por las actividades humanas, como la quema de combustibles fósiles y la deforestación.

Contaminación por bombeos de aguas residuales:La contaminación por bombeos de aguas residuales y la acumulación de flotsam y echazón son problemas que afectan la calidad del agua y la salud de los ecosistemas acuáticos.

Invasión humana y falta de mitigación:La invasión humana en áreas protegidas y la falta de mitigación del cambio climático son factores que contribuyen a la pérdida de biodiversidad y el estrés ecológico de los ecosistemas.

Huella de carbono:La huella de carbono es una medida importante que indica la cantidad de gases de efecto invernadero, como la emisión de CO2, que un individuo, empresa o país emite en su producción y consumo.

Contaminación por carbono:Para reducir la contaminación por carbono, es necesario fomentar el almacenamiento de carbono, como la reforestación, y el uso de alternativas de combustible de bajo carbono, como la energía renovable.

Modernización ecológica:La modernización ecológica se refiere a la adopción de tecnologías y prácticas más sostenibles en la producción y el consumo, lo que permite un crecimiento sostenible y reduce las emisiones de carbono.

Credenciales ecológicas:Las credenciales ecológicas, como la certificación de sostenibilidad, pueden ayudar a las empresas a demostrar su compromiso con la sostenibilidad y mejorar su reputación.

Sociedad post-carbono:La sociedad post-carbono es aquella en la que la economía y la tecnología no dependen de la quema de combustibles fósiles y se ha logrado una neutralidad de carbono.

Acciones contra el cambio climático:Las acciones contra el cambio climático, incluyendo la mitigación del cambio climático y la adopción de tecnologías bajas en carbono, son esenciales para enfrentar el problema del cambio climático.

Ciencia del cambio climático:La ciencia del cambio climático es fundamental para comprender los efectos del cambio climático y cómo mitigarlos.

Comunicación sobre el cambio climático:La comunicación sobre el cambio climático es importante para educar a la población y crear conciencia sobre los riesgos del cambio climático y las posibles soluciones.

Impuestos al carbono:Los impuestos al carbono son una herramienta efectiva para reducir las emisiones de carbono y fomentar el uso de alternativas de bajo carbono.

Captura, reducción y compensación de carbono:La captura, reducción y compensación de carbono son medidas que se pueden tomar para lograr la neutralidad de carbono y reducir la huella de carbono en la producción y consumo.

Compromiso con el cambio climático:El compromiso con el cambio climático implica la adopción de estrategias de mitigación que aborden la huella de carbono y promuevan la energía alternativa, así como la justicia social y la equidad global en la transición hacia una economía baja en carbono.

Campañas climáticas y proyectos relacionados con el cambio climático:Las campañas climáticas y proyectos relacionados con el cambio climático son esenciales para crear conciencia y fomentar la acción contra el cambio climático, mediante debates, juegos y estrategias que involucren a la sociedad y promuevan la comunicación climática en línea.

Greenwashing:El lavado verde, también conocido como greenwashing, es una práctica engañosa que se utiliza para dar una imagen de sostenibilidad y cuidado ambiental, cuando en realidad se está haciendo poco o nada para reducir la huella de carbono y promover prácticas sostenibles.

Justicia social y salud mundial en relación al cambio climático:La justicia social y la salud mundial están estrechamente relacionadas con la sostenibilidad y el cambio climático, ya que las comunidades más vulnerables son las que más sufren las consecuencias del cambio climático y las prácticas insostenibles.

Activismo ambiental y defensa del clima:El activismo ambiental y la defensa del clima son esenciales para promover la acción contra el cambio climático y fomentar políticas de mitigación, así como para monitorear el consumo y promover prácticas sostenibles.

Certificaciones ambientales y de sostenibilidad:Las certificaciones ambientales y de sostenibilidad, como la certificación de turismo sostenible, los requisitos de sostenibilidad y los indicadores de sostenibilidad, son herramientas importantes para promover prácticas pro-ambientales y el ahorro de energía y agua.

Esfuerzos de conservación y protocolos de impacto mínimo:Los esfuerzos de conservación y los protocolos de impacto mínimo son prácticas que buscan minimizar el impacto humano en los ecosistemas y promover la biodiversidad, mientras que los servicios sostenibles y las prácticas pro-ambientales buscan reducir el consumo y fomentar la sostenibilidad.

Política medioambiental:La acción relacionada con el clima y la política medioambiental son esenciales para promover la mitigación del cambio climático y fomentar la transición hacia una economía baja en carbono, así como para promover la conciencia ambiental y el comportamiento relacionado con el clima.

Información medioambiental y comunicación climática en línea:La información medioambiental y la comunicación climática en línea son herramientas importantes para crear conciencia y fomentar la acción contra el cambio climático, así como para promover interacciones verdes y la adopción de prácticas sostenibles.

Grupos de acción de racionamiento de carbono y mitigación del clima:Los grupos de acción de racionamiento de carbono y las iniciativas de mitigación del clima son esenciales para promover la reducción de emisiones y el control de emisiones, así como para promover la compensación de carbono y la neutralidad de carbono.

Comunicación sobre resiliencia ambiental:La comunicación sobre resiliencia ambiental es crucial para informar a la población sobre los desafíos que enfrentamos y cómo podemos adaptarnos al cambio climático de manera sostenible.

Negocios y hoteles sostenibles:Los negocios sostenibles deben implementar prácticas sostenibles y ofrecer productos y servicios ecológicos para garantizar un futuro sostenible.

Ecohoteles y parajes ecológicos:Los ecohoteles ofrecen experiencias ecológicas únicas que promueven prácticas sostenibles y respetuosas con el medio ambiente.

Alimentos sotenibles:Los alimentos sostenibles se producen utilizando prácticas agrícolas sostenibles y promueven la biodiversidad y el uso responsable de los recursos naturales.

Edificios sostenibles:Los edificios sostenibles son construcciones diseñadas para maximizar la eficiencia energética y minimizar el impacto ambiental durante su ciclo de vida.

Reducción consumo de energía y fuentes de energía renovables:La reducción del consumo de energía y la implementación de fuentes de energía renovable son clave para reducir las emisiones de gases de efecto invernadero.

Economía circular:La economía circular busca reducir el desperdicio y optimizar el uso de recursos naturales a través del reciclaje, la reutilización y la reparación.

Gestión de residuos:La gestión de residuos eficiente es fundamental para minimizar el impacto ambiental y promover prácticas sostenibles.

Adaptación al cambio climático:La adaptación al cambio climático implica tomar medidas para mitigar los efectos del cambio climático y reducir la vulnerabilidad de las comunidades y los ecosistemas.

Servicios sostenibles:Los servicios sostenibles son aquellos que se prestan de manera respetuosa con el medio ambiente y promueven prácticas sostenibles.

Certificación de turismo sostenible:La certificación de turismo sostenible es un requisito clave para garantizar que los negocios sostenibles, como los ecohoteles, ofrezcan experiencias ecológicas y productos ecológicos.

Prácticas pro-ambientales:Las prácticas pro-ambientales, como la reducción del consumo de agua y el reciclaje, son medidas clave de adaptación al clima que deben ser implementadas por empresas con políticas medioambientales sólidas.

Medidas de mantenimiento de playas:La construcción del malecón, el bombeo de arena y la restauración de playas son medidas de reposición de playas necesarias para contrarrestar los efectos del cambio climático, como el aumento del nivel del mar.

Equidad global:La equidad global es un principio clave de la sostenibilidad y debe ser considerada en todas las políticas de mitigación y adaptación al clima.

Reciclaje:El reciclaje es la clave para reducir la cantidad de basura que generamos al utilizar contenedores adecuados, reusar lo que podamos y tirar correctamente lo demás es fundamental para cuidar el medio ambiente

Conciencia ambiental:El desarrollo de la conciencia ambiental es crucial para fomentar la toma de decisiones sostenibles en todos los ámbitos de la sociedad y reducir el impacto negativo en el medio ambiente.

Juegos sobre el cambio climático:Los juegos sobre el cambio climático pueden ser una herramienta efectiva para educar y concienciar a las personas sobre la importancia de abordar este problema global y fomentar la toma de decisiones sostenibles.

Gases de efecto invernadero:La emisión excesiva de gases de efecto invernadero, como el dióxido de carbono, el metano y el óxido nitroso, es una de las principales causas del calentamiento global y del cambio climático.

Desperdicio del agua:La reducción del desperdicio del agua es fundamental para garantizar la sostenibilidad de los recursos hídricos y promover prácticas sostenibles en todos los ámbitos de la sociedad.

Reducción de desechos:Implementar sistemas de reciclaje y compostaje, fomentar la reutilización de productos y reducir el uso de materiales de un solo uso son algunas de las estrategias clave para lograr una reducción efectiva de desechos y avanzar hacia una economía más circular y sostenible.

Prevención de incendios:Las charlas de prevención de incendios son una herramienta valiosa para educar a la comunidad sobre las medidas que se deben tomar para prevenir incendios y cómo actuar en caso de emergencia.

Senderismo:El senderismo es una actividad física y recreativa que consiste en caminar por senderos, rutas o montañas, generalmente en áreas naturales y es una excelente forma de ejercitarse, estar en contacto con la naturaleza y disfrutar de paisajes y vistas impresionantes.

Anexo 4: Código para análisis de los tweets

```
import torch

from transformers import AutoTokenizer, AutoModel

from sklearn.metrics.pairwise import cosine_similarity

import csv

import os

import copy


# Cargar el tokenizador y modelo pre-entrenado para el idioma español

tokenizer_esp =
AutoTokenizer.from_pretrained("hiiamsid/sentence_similarity_spanish_es", )

model_esp =
AutoModel.from_pretrained("hiiamsid/sentence_similarity_spanish_es",
output_hidden_states=True)




def get_embeddings(tokenizer, model, references, token_length):

    """

        Obtener los embeddings de las frases de referencia

    """

    embeddings = []

    for text in references:

        tokens = tokenizer(text, max_length=token_length,
padding='max_length', truncation=True)

        output = model(torch.tensor(tokens.input_ids).unsqueeze(0),


attention_mask=torch.tensor(tokens.attention_mask).unsqueeze(0)).hidden_sta
tes[-1]

        embedding = torch.mean(output, axis=1).detach().numpy()

        embeddings.append(embedding)

    return embeddings




def get_embedding(tokenizer, model, text, token_length):

    """

        Obtener el embedding del tweet a analizar

    """
```

```
    tokens = tokenizer(text, max_length=token_length, padding='max_length',
truncation=True)
    output = model(torch.tensor(tokens.input_ids).unsqueeze(0),

attention_mask=torch.tensor(tokens.attention_mask).unsqueeze(0)).hidden_sta
tes[-1]
    return torch.mean(output, axis=1).detach().numpy()



def calculate_similarity(tokenizer_sim, model_sim, text1, embeddings,
keywords, token_length=20):
    """
       Calcula la similitud entre el embedding del tweet y los embeddings
de referencia. Nos quedaremos con el subtema con mayor similitud, siempre y
cuando supere el threshold de 0.42.
    """
    sim_max = None
    iter_max = None
    i = 0
    out = get_embedding(tokenizer_sim, model_sim, text1,
token_length=token_length)
    for embed in embeddings:
        sim = cosine_similarity(embed, out)[0][0]
        if sim > 0.42 and (sim_max is None or sim > sim_max):
            sim_max = sim
            iter_max = i
        i += 1
    if sim_max is not None and iter_max is not None:
        keywords[iter_max][1] += 1
        keywords[iter_max][2] += f"{text1} {sim_max}"
        keywords[iter_max][2] += f"\n---\n "



def split_text(lang):
    """
        Obtiene de la lista de referencia los conceptos y las descripciones
    """
    keywords = []
```

```
    descriptions = []
    if lang == 'esp':
        archivo = os.path.join(os.path.dirname(os.path.abspath(__file__)),
'referencias_esp.txt')
    else:
        print('Lang no es válido')
        archivo = None
    with open(archivo, encoding='utf-8') as f:
        for line in f:
            parts = line.split(':')
            if len(parts) == 2:
                row = [parts[0].strip(), 0, ""]
                keywords.append(row)
                descriptions.append(parts[1].strip())
    return descriptions, keywords



def process_files(folder_path, lang):
    """
       Procesa los archivos con los tweets y calcula la similitud con las
referencias.
    """
    variables = {}
    descriptions, keywords = split_text(lang)
    tokenizer_path = f"tokenizer_{lang}"
    model_path = f"model_{lang}"
    tokenizer = globals()[tokenizer_path]
    model = globals()[model_path]
    keywords_original = copy.deepcopy(keywords)
    length = 20
    descriptions_embeddings = get_embeddings(tokenizer, model,
descriptions, token_length=length)
    for filename in os.listdir(folder_path):
        file_path = os.path.join(folder_path, filename)
        name = filename.split('.')[0]
        keywords = copy.deepcopy(keywords_original)
```

```
        with open(file_path, newline='', encoding='utf-8') as csvfile:
            reader = csv.reader(csvfile)
            for row in reader:
                sentence = ' '.join(row)
                calculate_similarity(tokenizer, model, sentence,
descriptions_embeddings, keywords)
            variables[name] = keywords
    return variables



def save_matrix_to_csv(variable_name, matrix):
    """
       Guarda los resultados en un archivo CSV para cada cuenta.
    """
    results_path = os.path.join(os.path.dirname(os.path.abspath(__file__)),
'results')
    if not os.path.exists(results_path):
        os.makedirs(results_path)
    print("Hi "+results_path+" "+variable_name)
    file_path = os.path.join(results_path, f"{variable_name}.csv")
    with open(file_path, mode="w", newline="", encoding='utf-8') as file:
        writer = csv.writer(file)
        for row in matrix:
            writer.writerow(row)



# Procesar archivos en la carpeta "esp"
results = {}
path_esp = os.path.join(os.path.dirname(os.path.abspath(__file__)), 'esp')
results_esp = process_files(path_esp, 'esp')
results.update(results_esp)
# Guardar los resultados en archivos CSV
for var_name, var_value in results.items():
    save_matrix_to_csv(var_name, var_value)
```

Anexo 5: Análisis de subtemas a nivel de cuenta

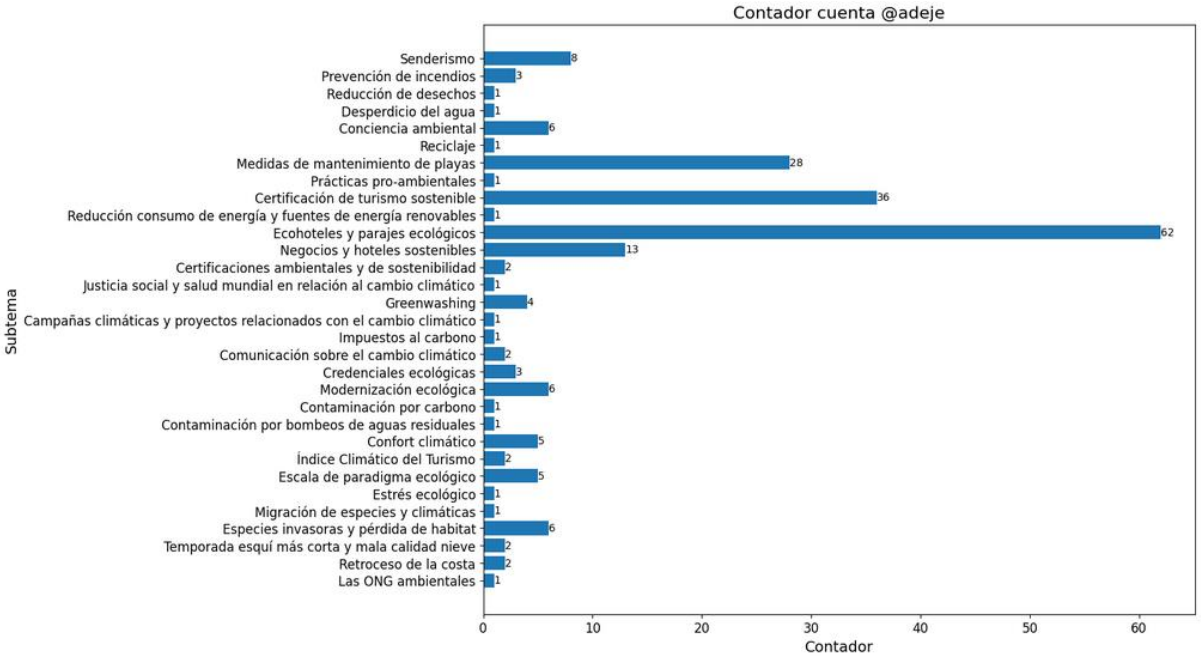

Figura 10: Número de tweets que abarcan los diferentes subtemas en la cuenta @adeje

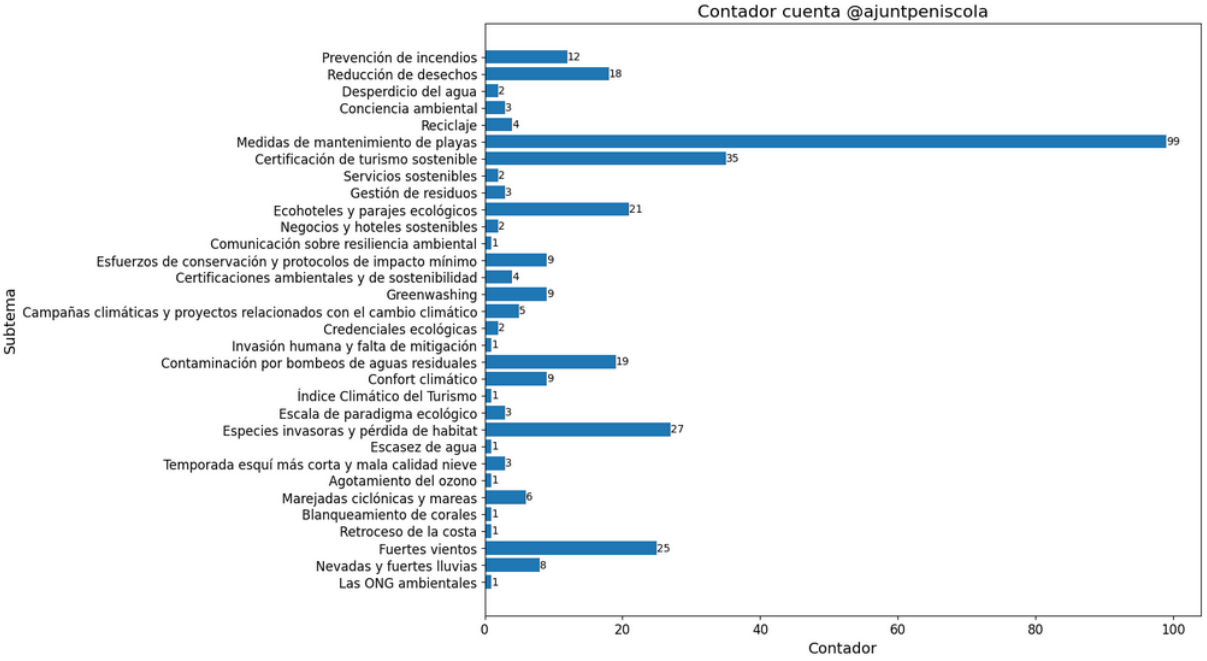

Figura 11: Número de tweets que abarcan los diferentes subtemas en la cuenta @ajuntpeniscola

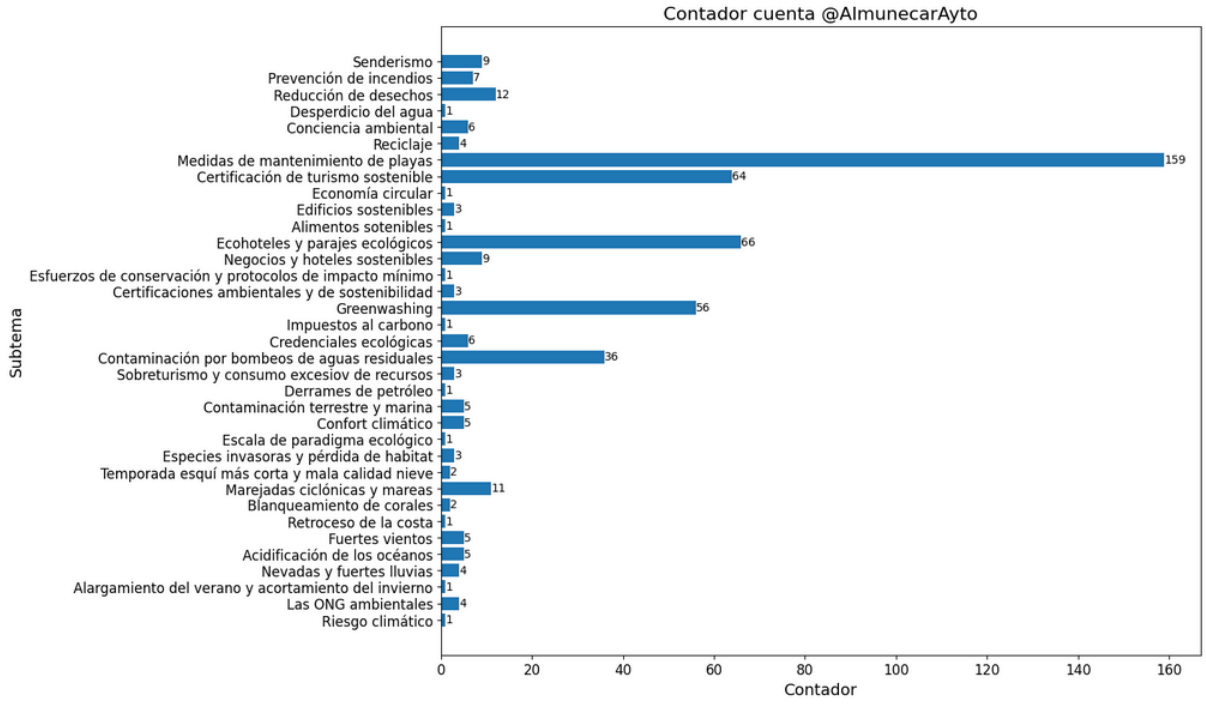

Figura 12: Número de tweets que abarcan los diferentes subtemas en la cuenta @AlmunecasAyto

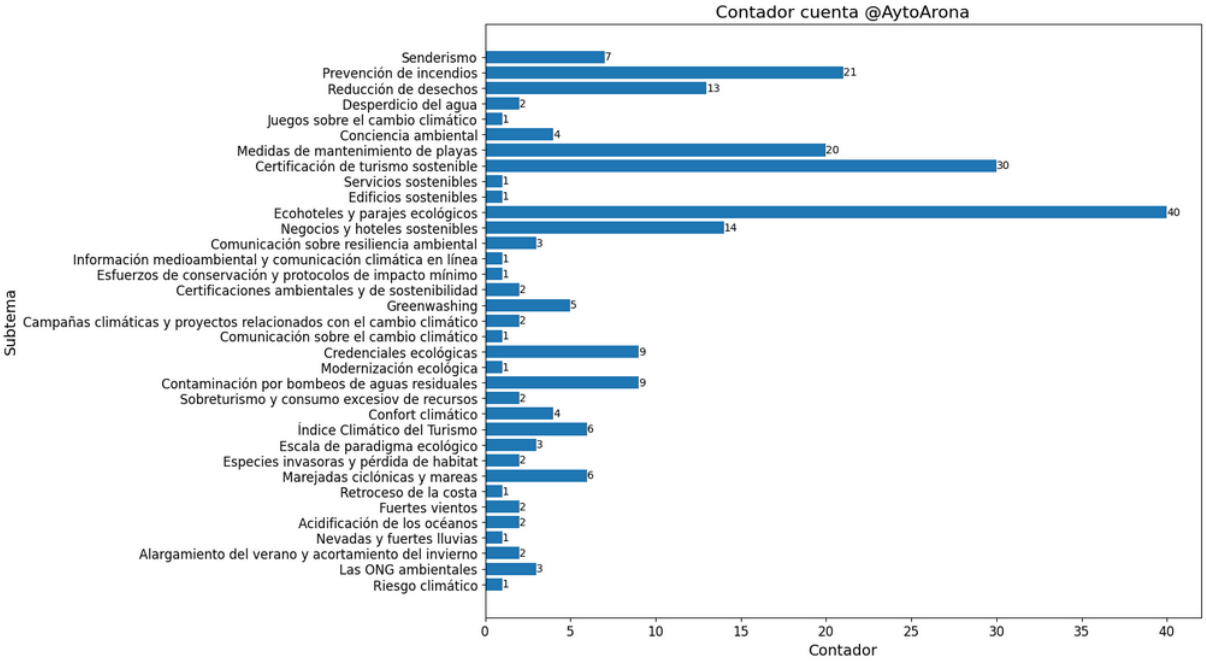

Figura 13: Número de tweets que abarcan los diferentes subtemas en la cuenta @AytoArona

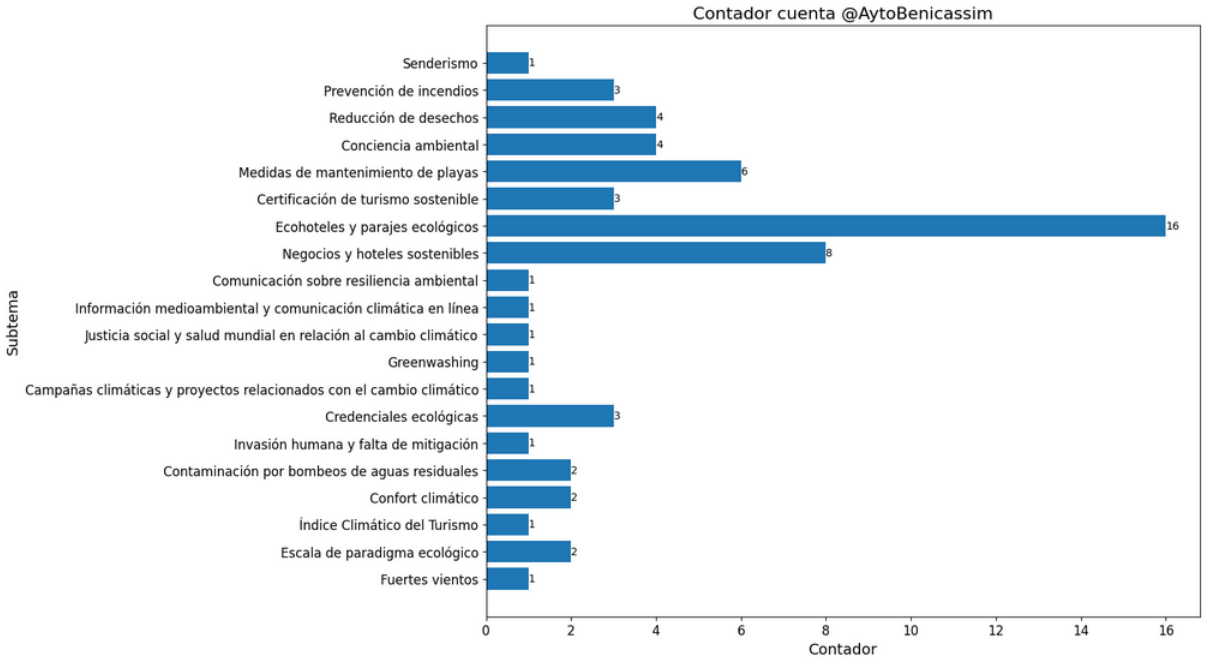

Figura 14: Número de tweets que abarcan los diferentes subtemas en la cuenta @AytoBenicassim

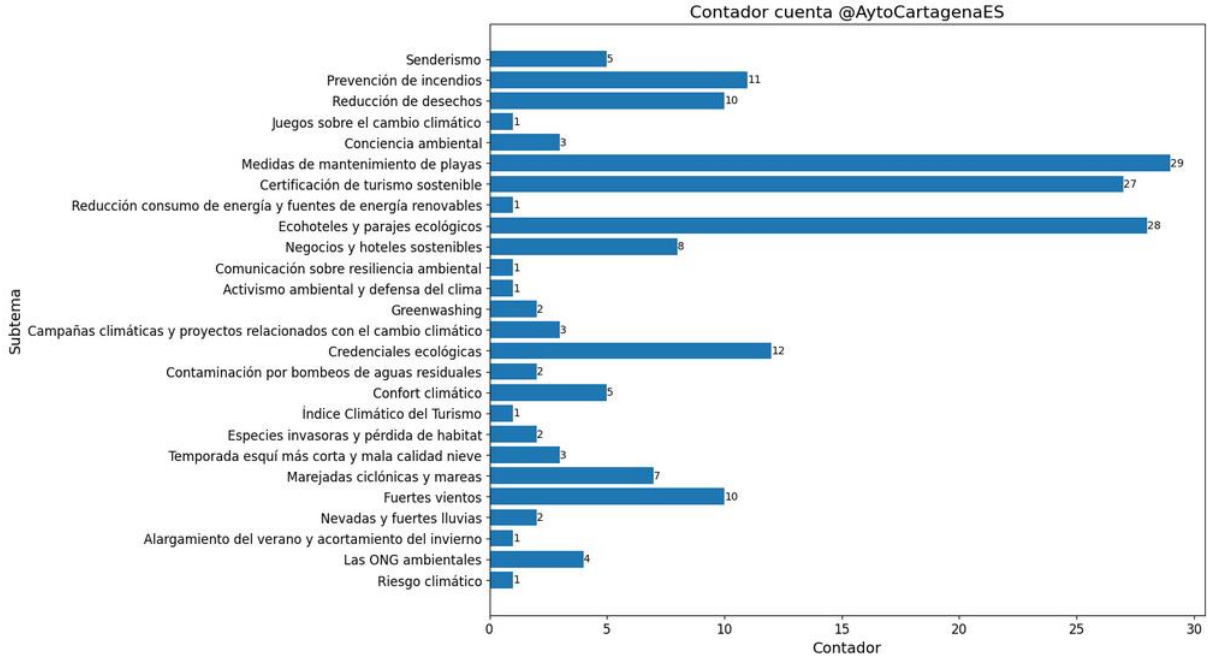

Figura 15: Número de tweets que abarcan los diferentes subtemas en la cuenta @AytoCartagenaES

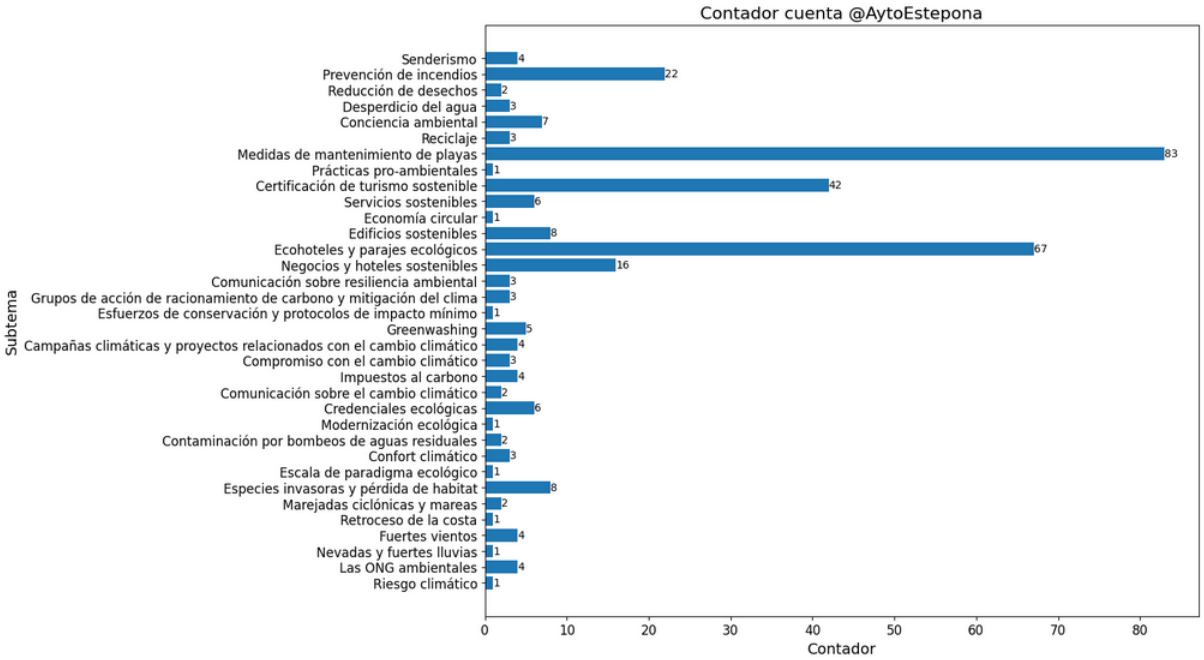

Figura 16: Número de tweets que abarcan los diferentes subtemas en la cuenta @AytoEstepona

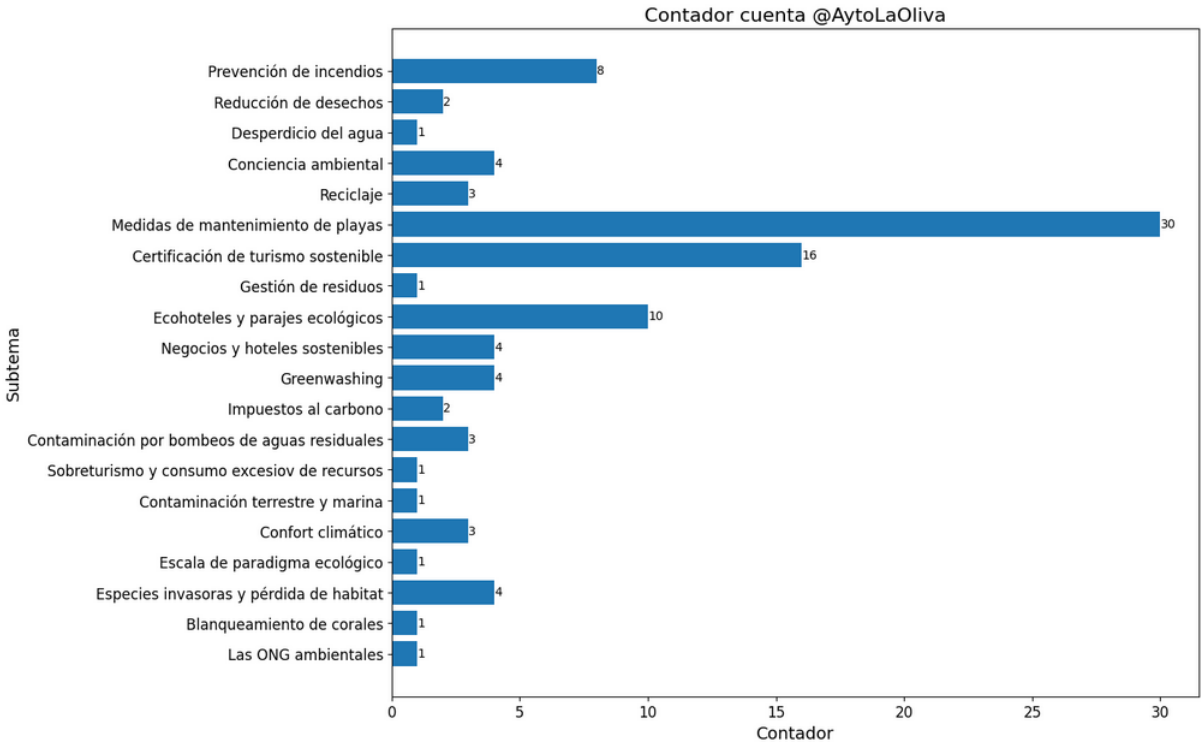

Figura 17: Número de tweets que abarcan los diferentes subtemas en la cuenta @AytoLaOliva

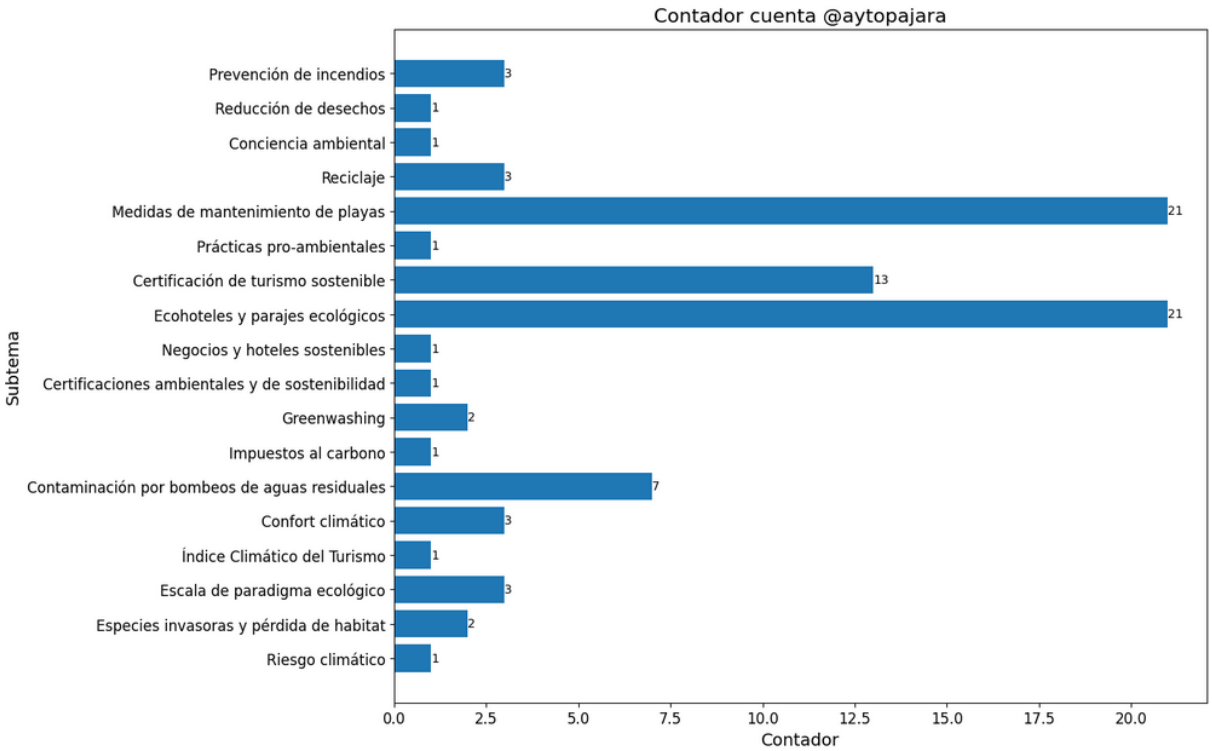

Figura 18: Número de tweets que abarcan los diferentes subtemas en la cuenta @aytopajara

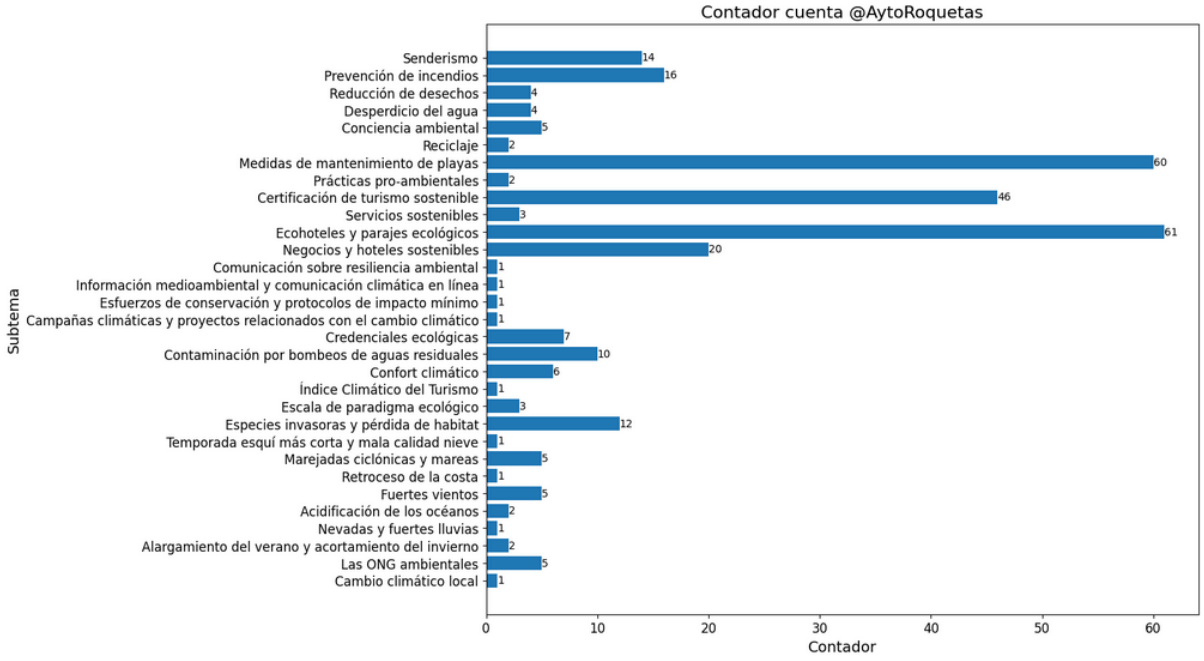

Figura 19: Número de tweets que abarcan los diferentes subtemas en la cuenta @AytoRoquetas

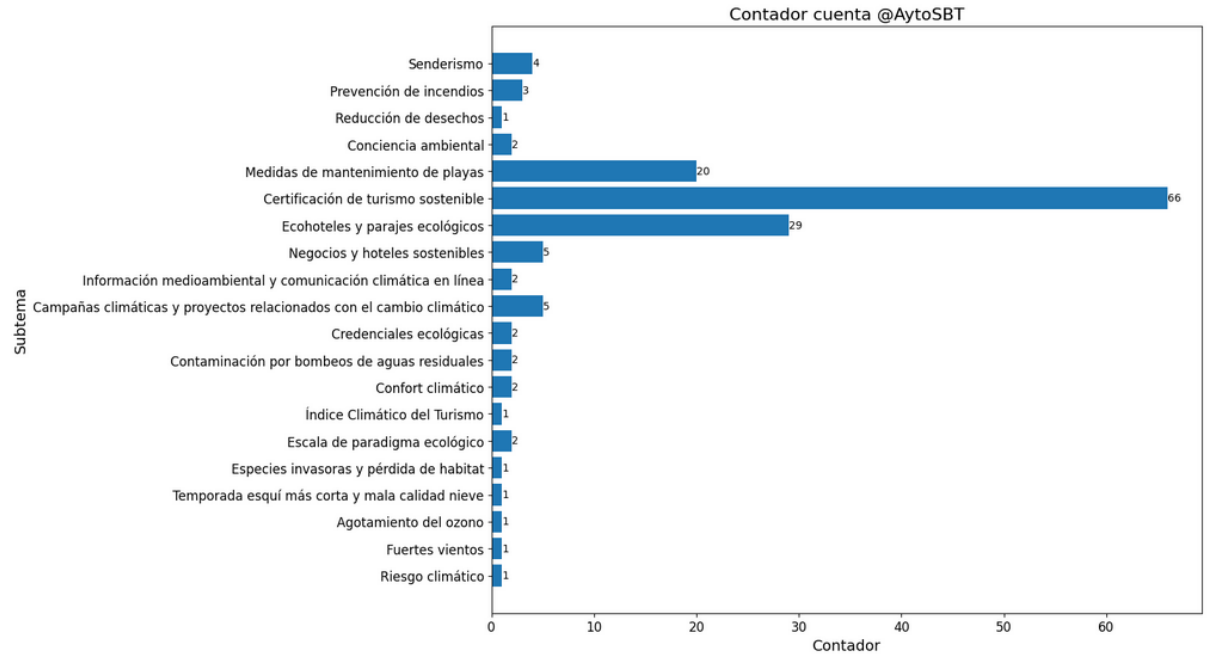

Figura 20: Número de tweets que abarcan los diferentes subtemas en la cuenta @AytoSBT

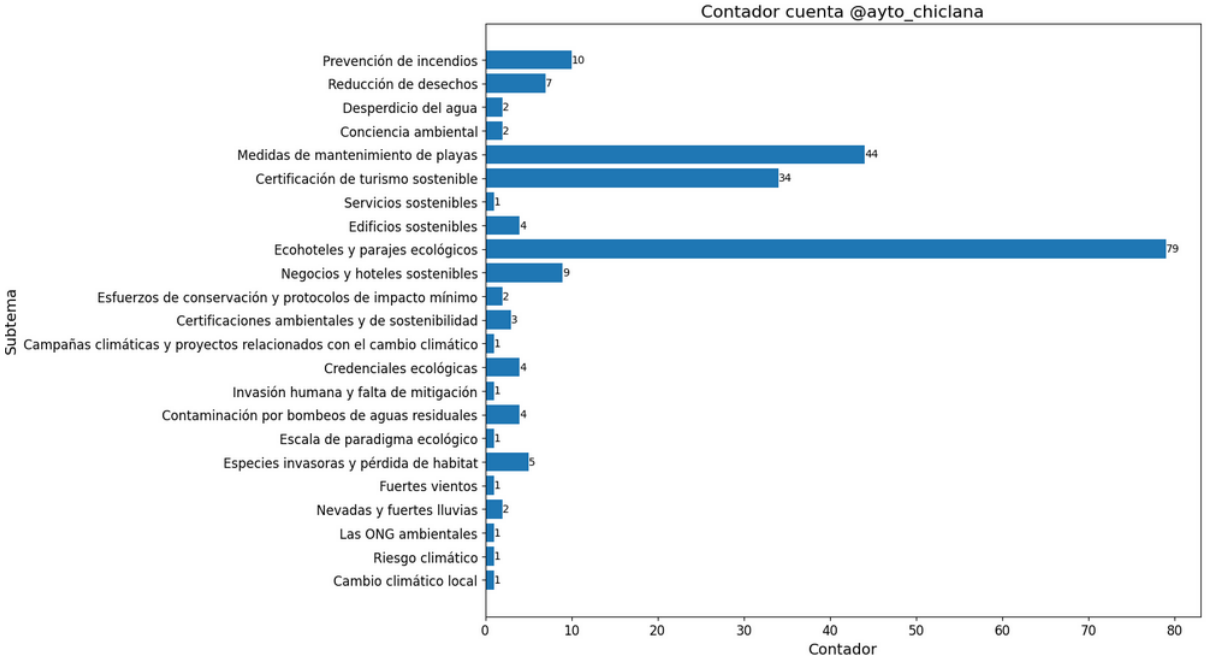

Figura 21: Número de tweets que abarcan los diferentes subtemas en la cuenta @ayto_chiclana

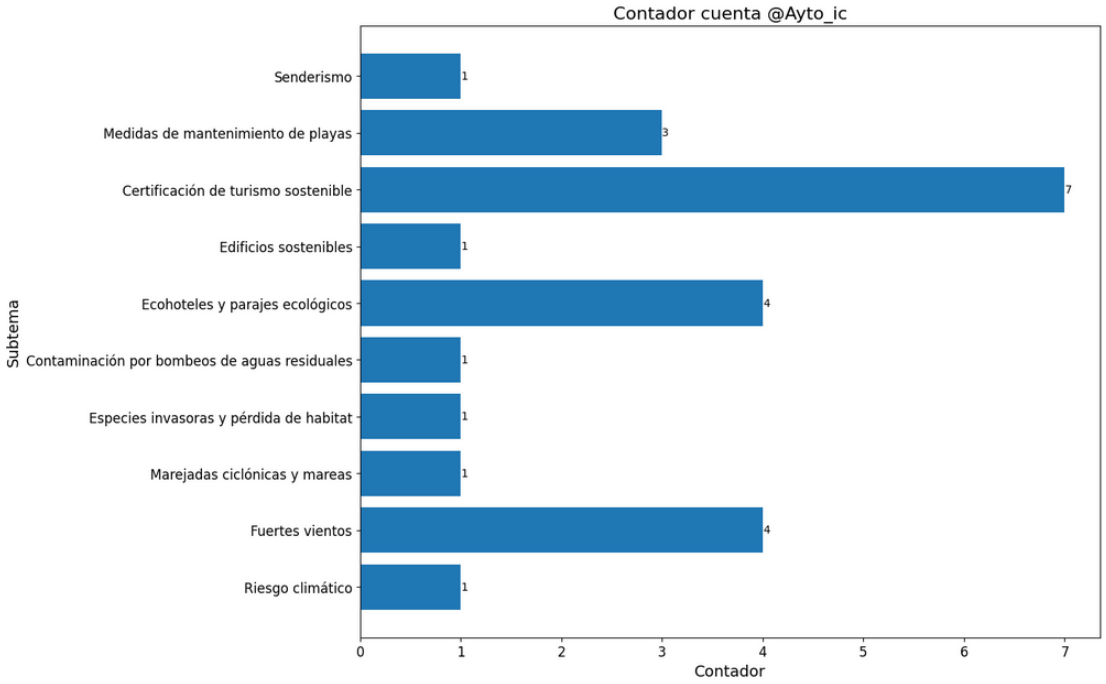

Figura 22: Número de tweets que abarcan los diferentes subtemas en la cuenta @Ayto_ic

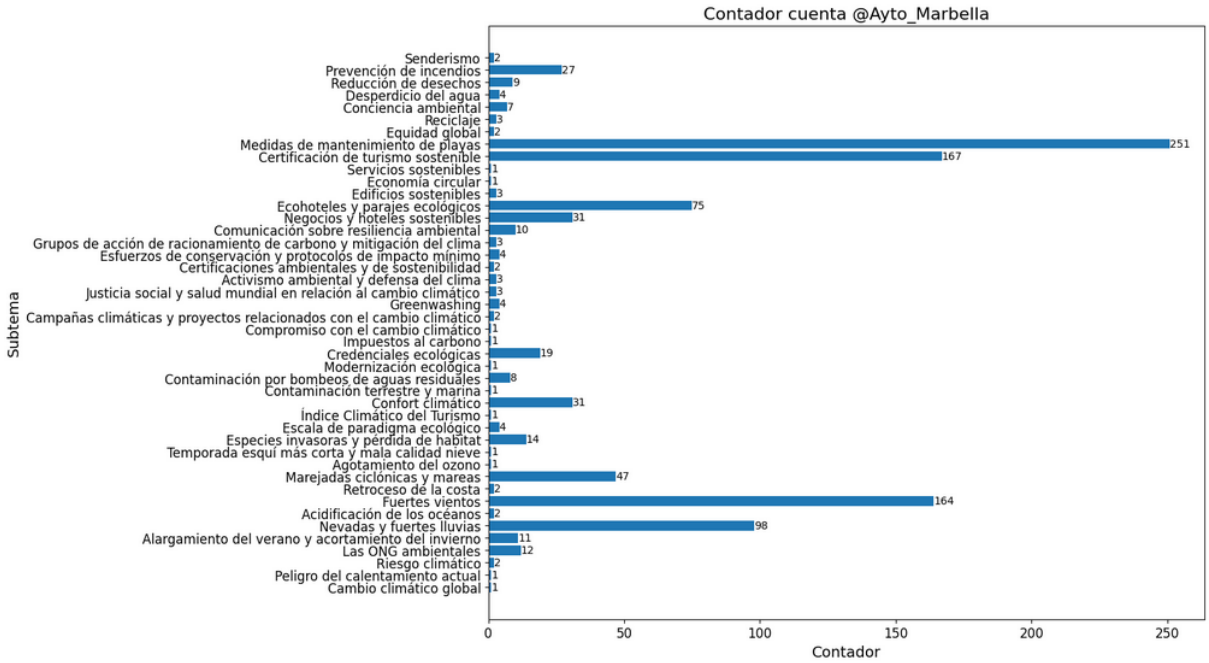

Figura 23: Número de tweets que abarcan los diferentes subtemas en la cuenta @Ayto_Marbella

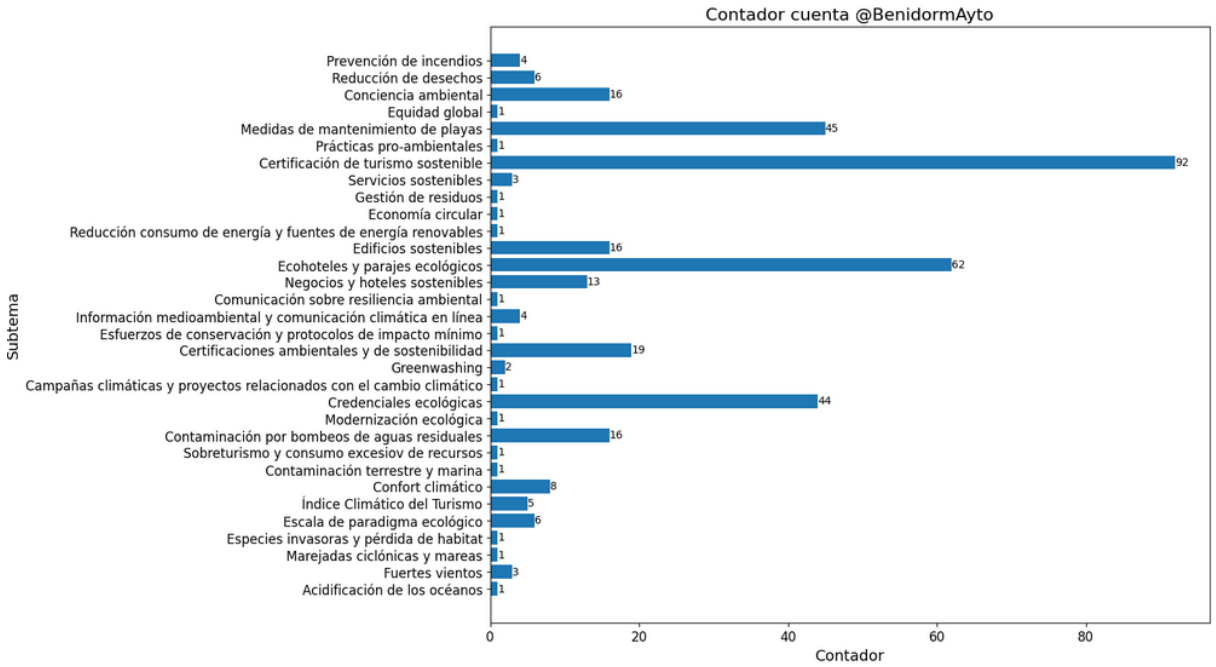

Figura 24: Número de tweets que abarcan los diferentes subtemas en la cuenta @BenidormAyto

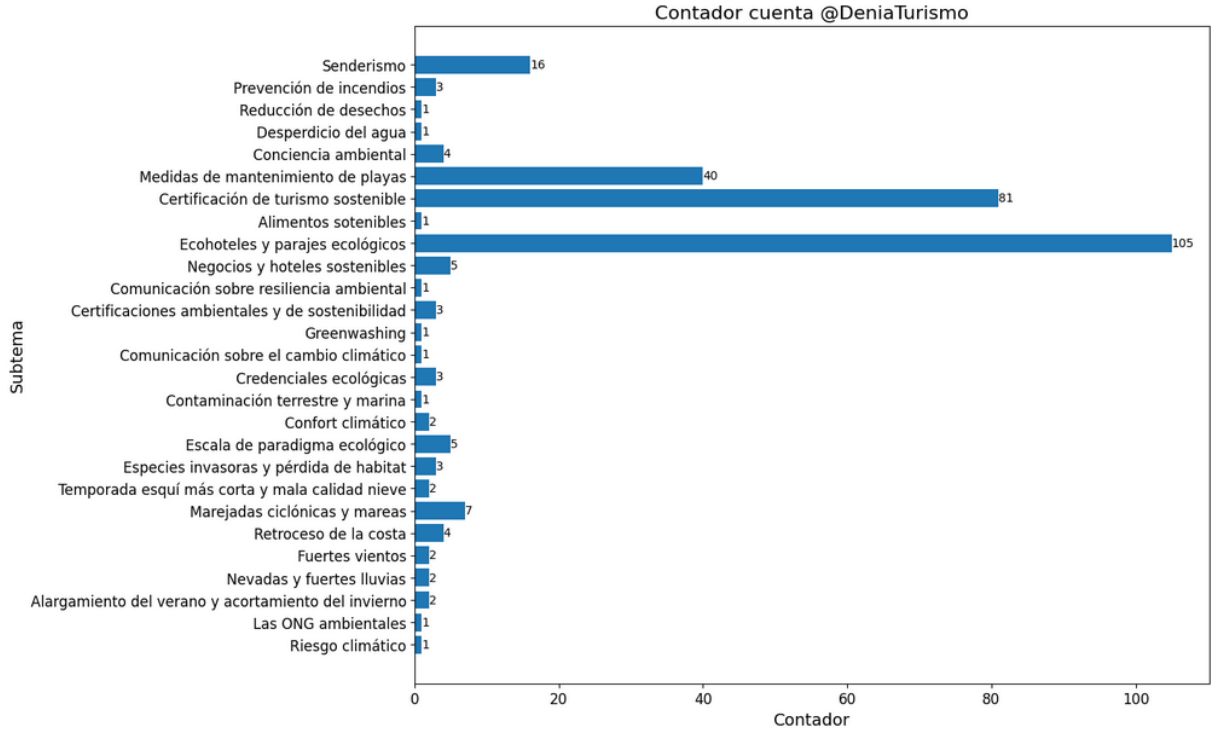

Figura 25: Número de tweets que abarcan los diferentes subtemas en la cuenta @DeniaTurismo

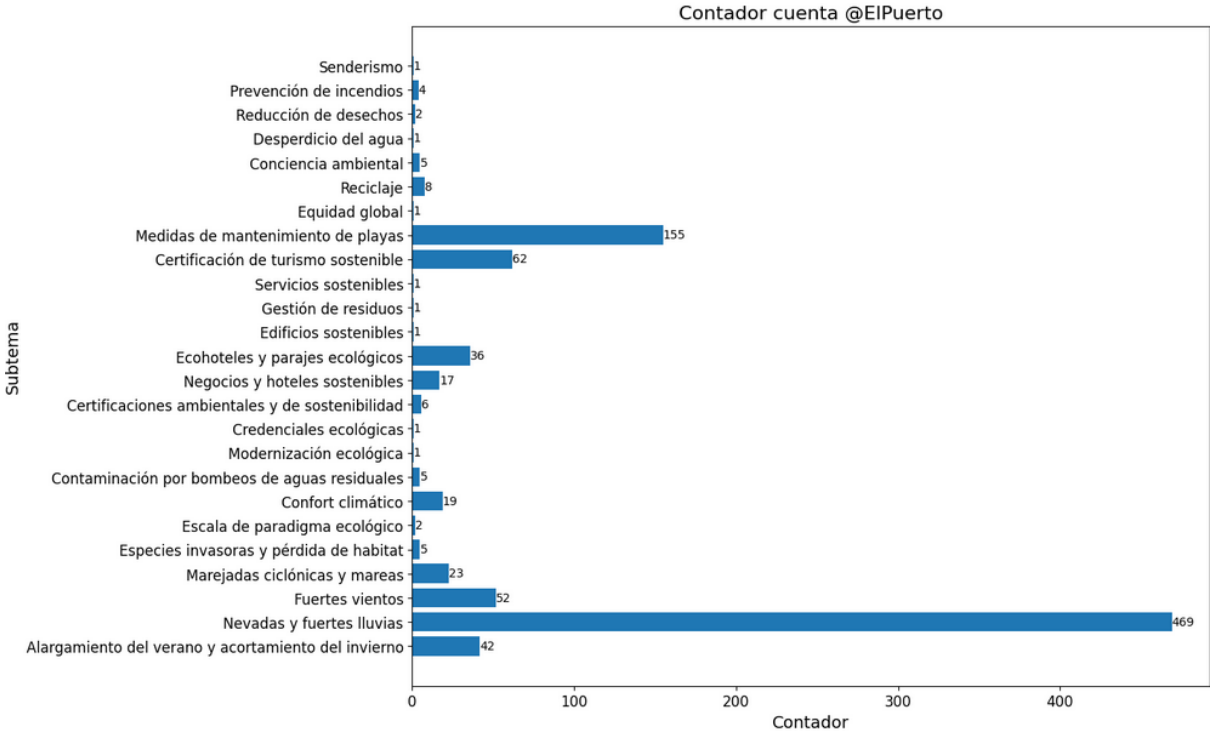

Figura 26: Número de tweets que abarcan los diferentes subtemas en la cuenta @ElPuerto

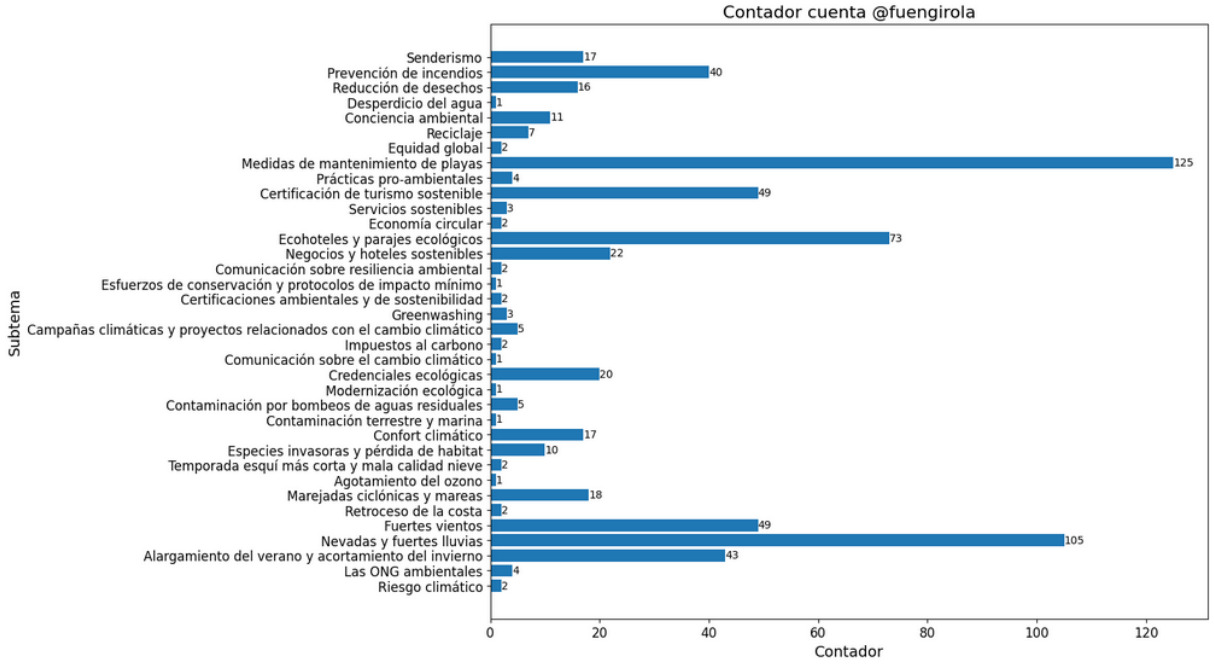

Figura 27: Número de tweets que abarcan los diferentes subtemas en la cuenta @fuengirola

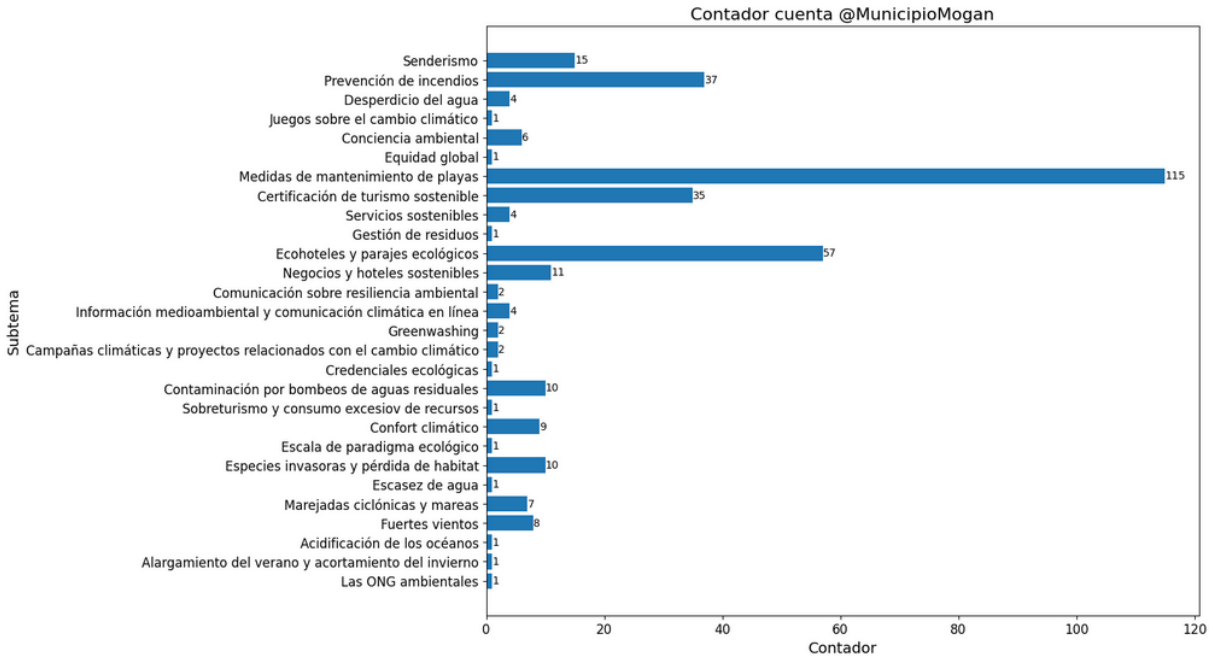

Figura 28: Número de tweets que abarcan los diferentes subtemas en la cuenta @MunicipioMogan

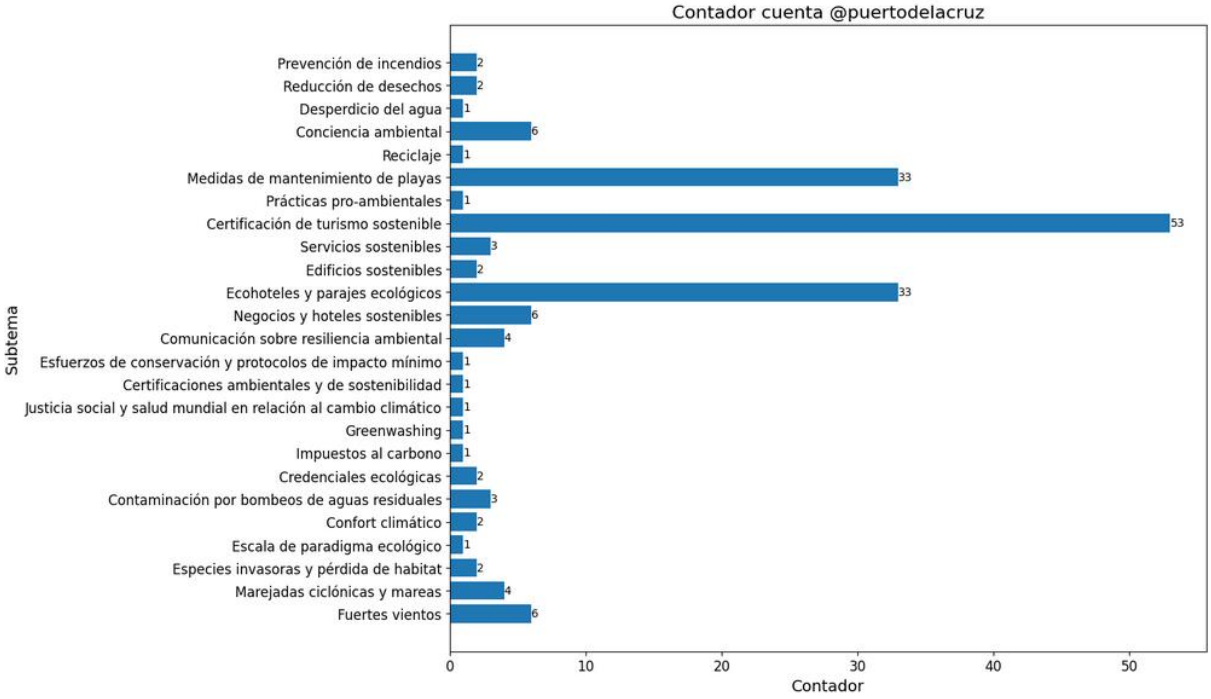

Figura 29: Número de tweets que abarcan los diferentes subtemas en la cuenta @puertodelacruz

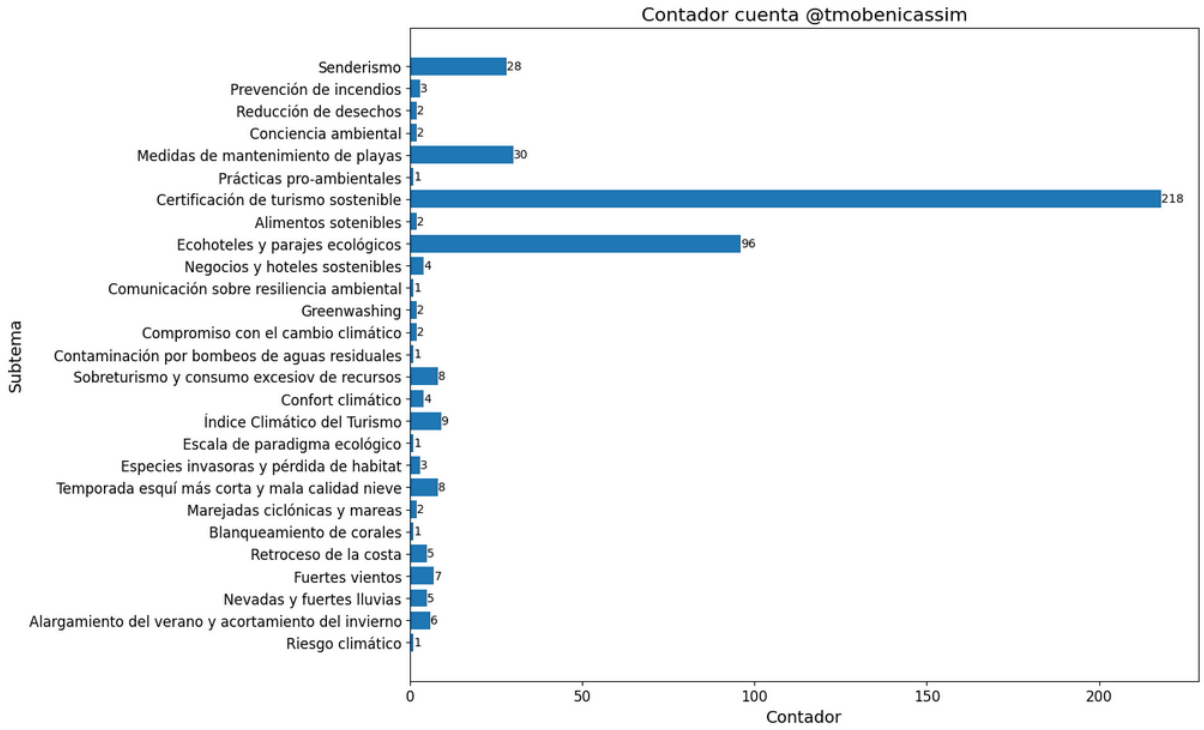

Figura 30: Número de tweets que abarcan los diferentes subtemas en la cuenta @tmobenicassim

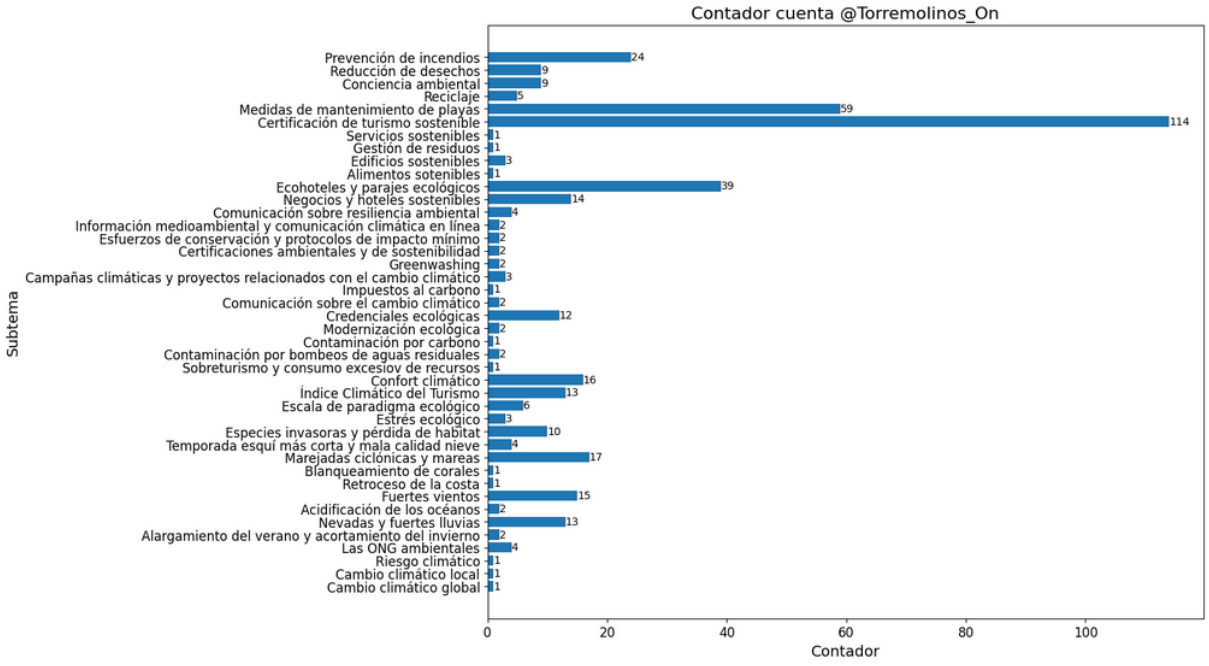

Figura 31: Número de tweets que abarcan los diferentes subtemas en la cuenta @Torremolinos_On

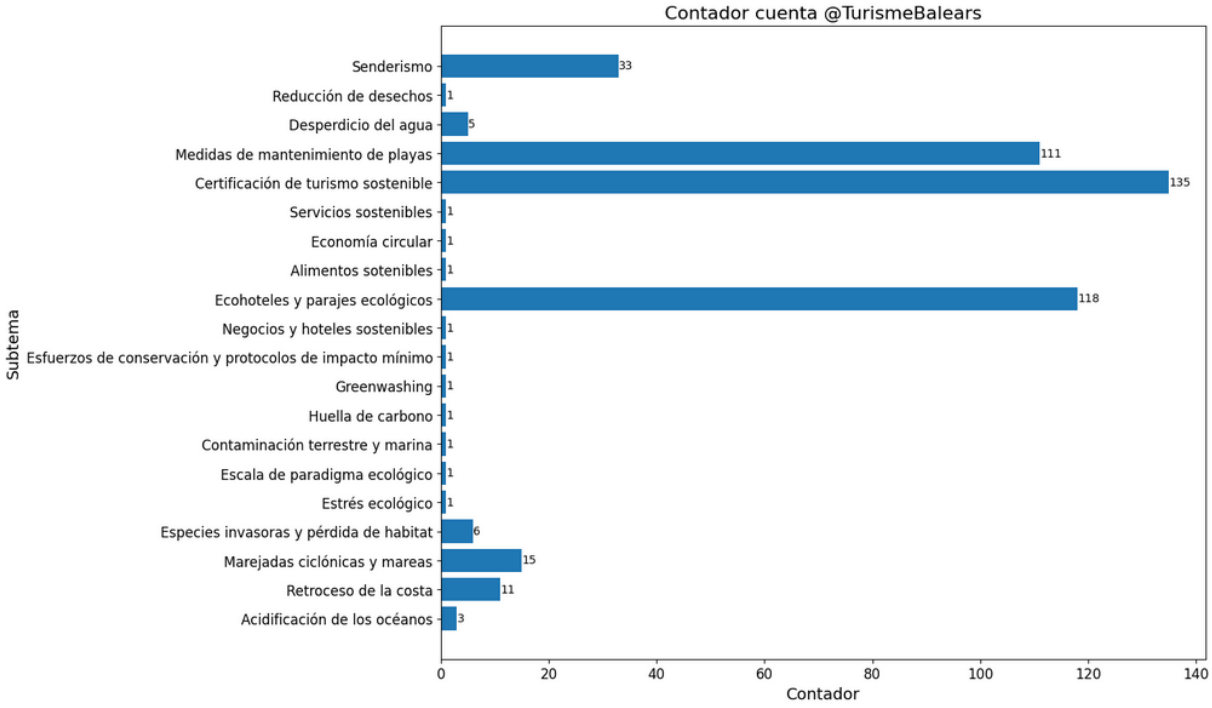

Figura 32: Número de tweets que abarcan los diferentes subtemas en la cuenta @TurismeBalears

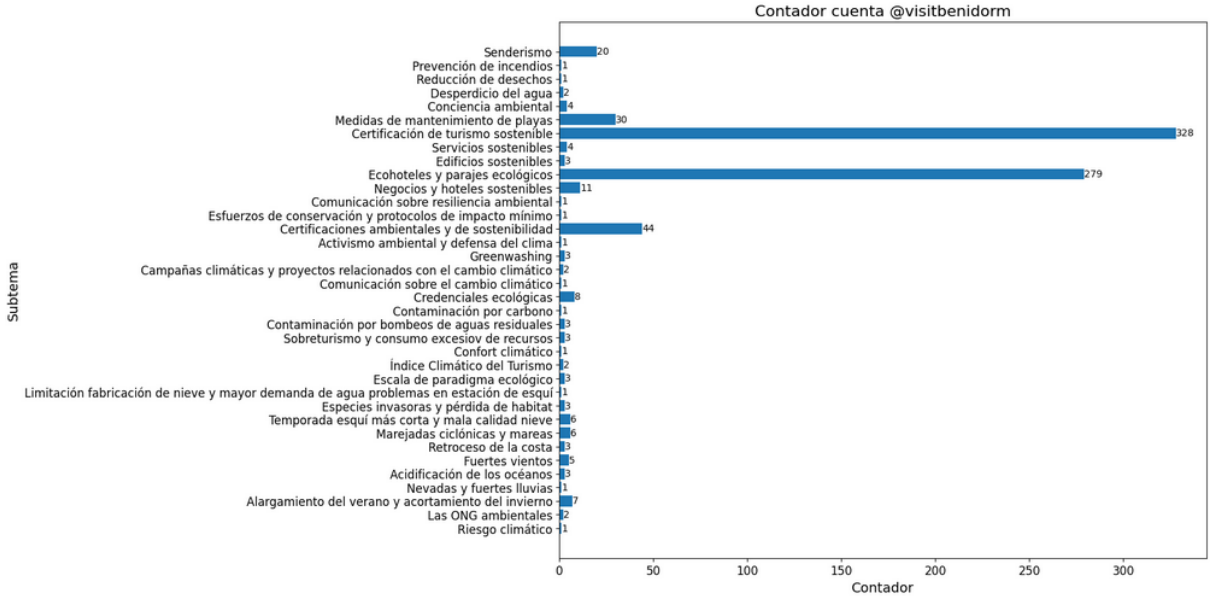

Figura 33: Número de tweets que abarcan los diferentes subtemas en la cuenta @visitbenidorm

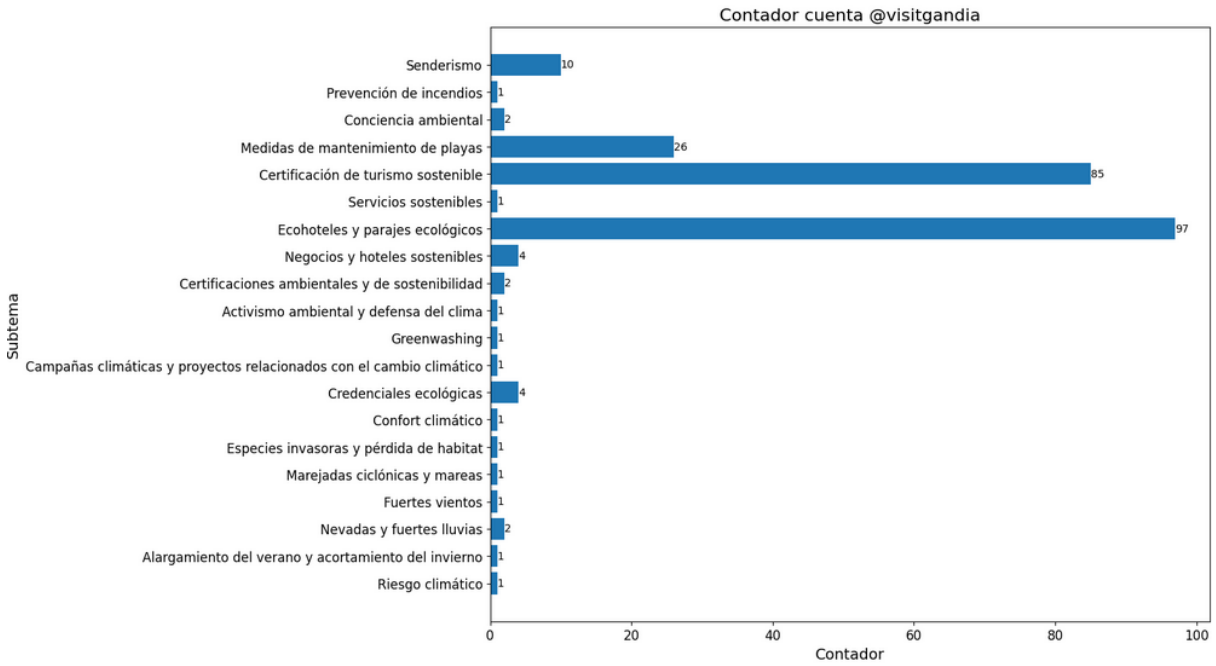

Figura 34: Número de tweets que abarcan los diferentes subtemas en la cuenta @visitgandia

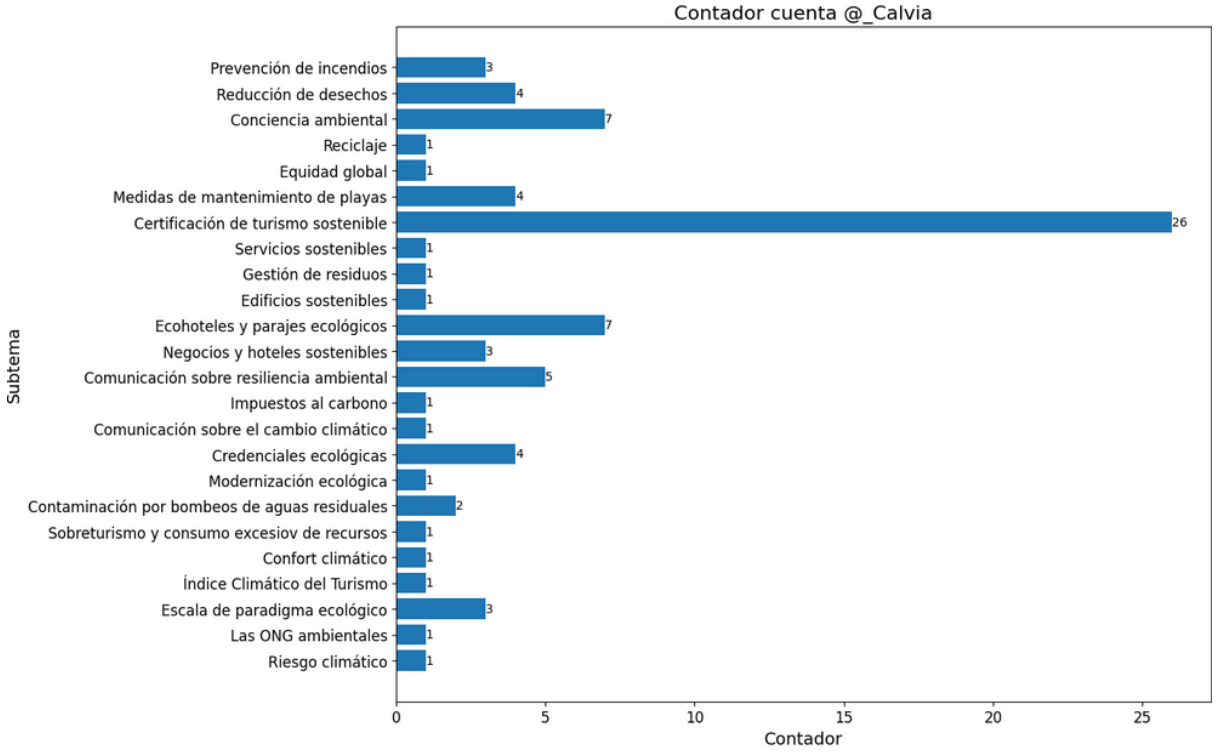

Figura 35: Número de tweets que abarcan los diferentes subtemas en la cuenta @_Calvia

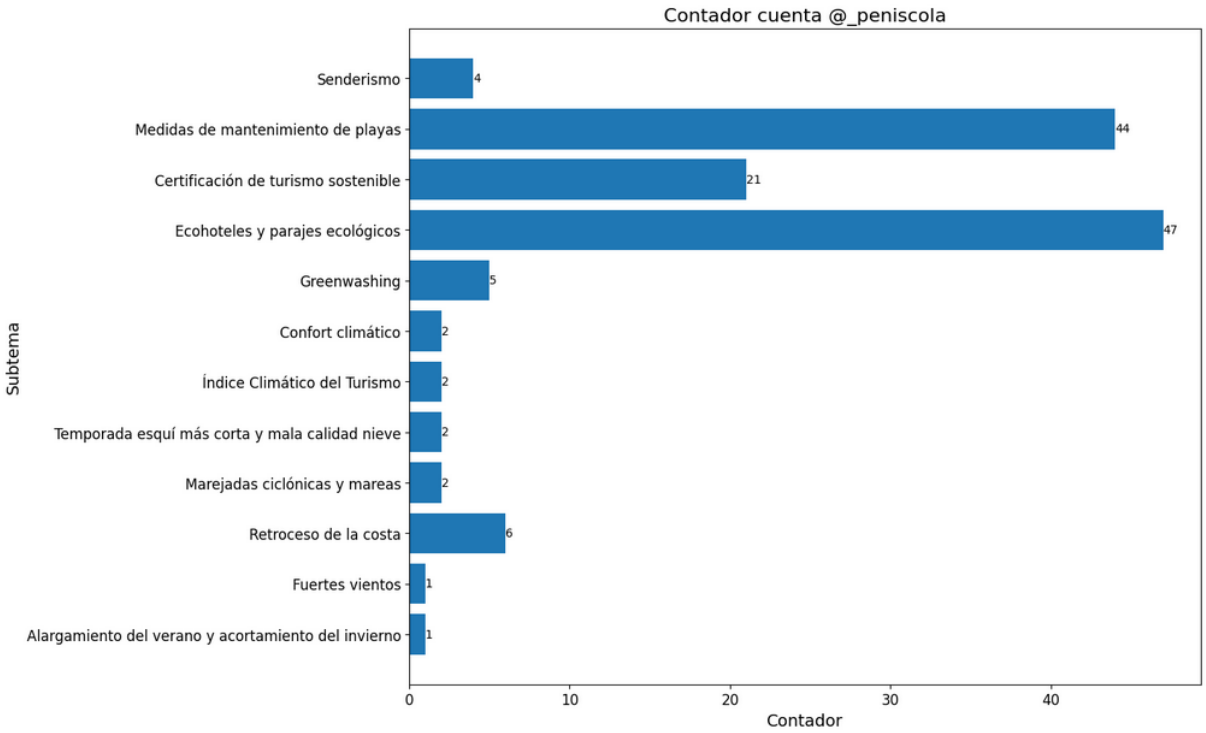

Figura 36: Número de tweets que abarcan los diferentes subtemas en la cuenta @_peniscola