COMILLAS
UNIVERSIDAD PONTIFICIA

Facultad de Ciencias Económicas y Empresariales

EXTRACCIÓN DE LAS PERCEPCIONES DE MARCA EN LA RED SOCIAL DE TWITTER: FACTORES DE ÉXITO DE SHEIN EN ESPAÑA

Nombre: Maria Moreno Bastante
Director: Jenny Alexandra Cifuentes Quintero

MADRID| Abril 2023

Resumen

El presente trabajo de grado se enfoca en realizar un análisis exhaustivo de la percepción de la marca Shein en España a través de redes sociales. La marca ha experimentado un crecimiento significativo en los últimos años y en especial durante la pandemia, lo que ha generado un gran interés en su análisis. Además, se ha recibido críticas por su modelo de negocio y su impacto social y ambiental, lo que ha generado un mayor interés en su estudio. El objetivo principal del estudio es obtener una visión clara de cómo se percibe la marca en el mercado español y qué aspectos son valorados positiva o negativamente. En este contexto, durante la primera etapa de desarrollo, se ha realizado un exhaustivo análisis de la literatura relacionada, identificando la técnica LDA como la más adecuada para el modelado de categorías de discusión y el diccionario VADER como estrategia para el análisis de sentimientos. De esta manera, a través de la API de Twitter, se recopilaron publicaciones asociados con la marca desde el comienzo de su período de expansión en 2017 hasta la actualidad, volumen de datos que ha sido analizado utilizando diferentes técnicas de minería de datos, incluyendo análisis de N-gramas, modelado de tópicos y análisis de sentimiento.

Con el análisis de N-gramas, la técnica TF-IDF permitió identificar que los conceptos más recurrentes en el texto estaban relacionadas principalmente con la recepción y entrega de los pedidos. El modelado de tópicos ha permitido identificar las principales categorías asociadas con la marca, tales como descuentos, consumo, looks y prendas, canal de compra y bikinis, y recepción de pedidos. De entre todos los tópicos, el tópico Recepción de pedidos, mostró un mayor volumen de reseñas en comparación con el resto, siendo el tópico Consumo el que menos reseñas presenta. El análisis de sentimiento ha demostrado que la mayoría de las publicaciones asociadas con Shein en cada uno de los tópicos y a nivel general, presentan un sentimiento neutral, aunque se observó un cambio significativo en el tópico de consumo y descuentos con respecto al periodo anterior y posterior de la pandemia, siendo más positivo en el período anterior a la crisis sanitaria.

En general, este trabajo ha permitido obtener una visión más clara de cómo la marca es percibida por los consumidores en España mediante redes sociales. Los resultados, pueden ser de gran utilidad para Shein a la hora de definir y establecer las estrategias futuras de la marca en el país así como mejorar la calidad de sus productos y servicios.

Abstract

The present work focuses on conducting an exhaustive analysis of the perception of the Shein brand in Spain through social networks. The brand has experienced significant growth in recent years and especially during the pandemic, which has generated great interest in its analysis. In addition, it has received criticism for its business model and its social and environmental impact, which has generated further interest in its study. The main objective of the study is to obtain a clear view of how the brand is perceived in the Spanish market and which aspects are valued positively or negatively. In this context, during the first stage of development, an exhaustive analysis of the related literature has been carried out, identifying the LDA technique as the most appropriate for the modeling of discussion categories and the VADER dictionary as a strategy for sentiment analysis. In this way, through the Twitter API, posts associated with the brand were collected from the beginning of its expansion period in 2017 to the present, a volume of data that has been analyzed using different data mining techniques, including N-gram analysis, topic modeling and sentiment analysis.

With N-gram analysis, the TF-IDF technique allowed us to identify that the most recurrent concepts in the text were mainly related to the receipt and delivery of orders. Topic modeling identified the main categories associated with the brand, such as discounts, consumption, looks and garments, shopping channel and bikinis, and receiving orders. Among all the topics, the topic Receiving orders, showed a higher volume of reviews compared to the rest, being the topic Consumption the one with the least number of reviews. The sentiment analysis has shown that most of the publications associated with Shein in each of the topics and at a general level, present a neutral sentiment, although a significant change was observed in the topic of consumption and discounts with respect to the period before and after the pandemic, being more positive in the period prior to the health crisis.

Overall, this work has provided a clearer picture of how the brand is perceived by consumers in Spain through social networks. The findings may be of great use to Shein in defining and establishing future brand strategies in the country as well as improving the quality of its products and services.

Agradecimientos

Quiero agradecer este trabajo principalmente a mi tutora, Jenny Alexandra Cifuentes Quintero, por haberme guiado a lo largo de todo el proceso y por brindarme un gran apoyo y ayuda para la elaboración de este trabajo.

Índice general

Índice de figuras

Índice de tablas

Acrónimos

CRM	Customer Relationship Management
LDA	Análisis Latente de Dirilecht
NBR	Net Brand Reputation
PIB	Producto interior bruto
PFC	Proyecto Fin de Carrera
SVM	Support Vector Machine
URL	Localizador uniforme de recursos
VADER	Valence Aware Dictionary and Sentiment Reasoner

Capítulo 1

Introducción

1.1. Motivación

La industria de la moda ha experimentado una gran evolución durante los últimos años, contando actualmente con una de las mayores cadenas de distribución y consumo del mercado. La dinámica cambiante de esta industria ha involucrado principalmente la reducción de la producción en masa, el aumento del número de temporadas de moda y la modificación de las características estructurales de la cadena de suministro (Bhardwaj y Fairhurst, 2010). Además de estos factores, el marketing digital ha sido uno de los grandes factores diferenciadores e impulsores de la competitividad en este mercado. Estos cambios han obligado a los minoristas a desear un bajo costo y flexibilidad en el diseño, así como la mejora en la calidad, la entrega y la rapidez de su respectiva comercialización. Como consecuencia, el concepto de *fast fashion* ha surgido para dar respuesta a las nuevas necesidades del mercado, ocasionando que, en muchos casos, las empresas hayan preferido mover sus centros productivos a países donde los costos de producción son más baratos. Así, productos de moda con precios más bajos pueden ser adquiridos por la mayoría de la sociedad, ampliando el público objetivo y convirtiendo a estas empresas en un agente económico de gran importancia (Mazaira, Gonzalez, y Avendaño, 2003). De hecho, para el año 2021, la industria de la moda aproximadamente generaba una actividad valorada en 1.5 trillones de dólares a nivel global, de los cuales 91.21 billones pertenecían al sector *fast fashion* (Smith, 2022). En España, actualmente, esta industria genera alrededor del 2.8 % del PIB (Modaes.es, 2022), y aunque esta cifra ha ido disminuyendo levemente desde el año 2000, la moda sigue siendo el producto más demandado en compras online con un 54 % para el año 2021, superando significativamente a productos tecnológicos y viajes, y creciendo más de dos puntos respecto al año anterior (Adevinta, 2021).

Dentro de los principales representantes de esta industria, encontramos marcas tan conocidas como H&M, Forever21 y el grupo Inditex. Sin embargo, aunque hasta ahora Inditex con marcas como Zara, ha sido la empresa líder del modelo *fast fashion* (McKinsey & Com-

pany, 2019), esta posición se está viendo amenazada por la aparición de un nuevo modelo de negocio de moda ultra rápida o moda a tiempo real representado por Shein, una marca de origen Chino. Shein destaca en primer lugar por el gran volumen de artículos disponibles simultáneamente para la venta con hasta 600.000 productos, agregando 2.800 nuevos estilos cada semana en su sitio web (Oxford Analytica, 2021). Asimismo, esta marca se enfoca en prendas low-cost, con precios inferiores a los de sus principales competidores y con diseños versátiles para un público objetivo considerablemente amplio. Además de estos factores Shein tiene numerossas estrategias de marketing y de publicidad como KOL, streaming en directo, programas de afiliación, publicidad en redes sociales, re-mercadeo, etc, convirtiendose, por ejemplo, en la marca más popular entre los adolescentes de TikTok (Shen, 2022).

Sin embargo, aunque los indicadores de este mercado a nivel económico son prometedores, el mercado fast fashion ha recibido abundantes críticas por su escasa consideración en cuestiones medioambientales y sociales, ubicando el tema asociado a los costos no financieros de la moda en la agenda pública mundial (Niinimäki et al., 2020). Asimismo, las numerosas denuncias recibidas, por parte de grandes y pequeñas empresas contra Shein, con motivo del plagio en los diseños de la ropa han supuesto grandes problemas para la reputación de esta marca. De hecho, como fue publicado en el año 2021, por la organización mundial de noticias The Guardian, la cuenta de Instagram del grupo de moda Diet Prada destacó las acusaciones de la diseñadora Bailey Prado sobre la copia de Shein de más de 45 de sus diseños (The Guardian, 2021).

Por otro lado, Shein como marca *fast fashion*, se ha visto sumergida en varios escándalos mediáticos debido a denuncias sobre las condiciones laborales de sus trabajadores. En ese sentido, la agencia de noticias Reuters reportó en el año 2021 que Shein no había informado los detalles asociados a su cadena de suministro, información exigida por la legislación británica de acuerdo con la Ley de Esclavitud Moderna del 2015. Asimismo, Reuters no pudo evaluar de forma independiente las condiciones de trabajo o los salarios de las fábricas, y aseguró que hasta el año pasado, el sitio web de la empresa había afirmado falsamente que sus condiciones de trabajo estaban certificadas por organismos internacionales de normas laborales (Reuters, 2021). A este respecto, un estudio publicado a finales del año 2021 por la ONG suiza de derechos humanos Public Eye denunció que los trabajadores de la marca trabajan alrededor de 75 horas semanales y su salario se calcula en función de prendas producidas en lugar de horas trabajadas o calidad del trabajo realizado (Public Eye, 2021).

Desde el punto de vista económico, un artículo publicado en la agencia de noticias de moda *Business of Fashion* estimó la facturación de Shein en el año 2020 en alrededor de 9700 millones de euros, una cifra significativamente mayor a los 6600 millones de Inditex y los 5100 millones de H&M. Estos resultados económicos evidencian el aumento en ventas de un 60 % frente al año anterior, debido en gran parte al periodo de pandemia del COVID 19.

En resumen, el éxito reportado a nivel mundial, y particularmente en España, y el impacto mediático de la marca con diversos tipos de noticias, positivas y negativas, motivan

el objetivo de este trabajo. De esta manera, este estudio pretende brindar un análisis de los factores asociados al modelo de mercado de Shein dentro de la industria fast fashion, así como explorar los principales factores de discusión sobre la percepción de marca en España, teniendo como proxy la opinión ciudadana reportada en redes sociales como Twitter. Teniendo en cuenta los diversos factores que motivan el desarrollo de este análisis, y que han sido abordados con anterioridad, en la siguiente sección se describen los objetivos generales y específicos de este trabajo.

1.2. Objetivos

1.2.1. Objetivo General

El objetivo general de este trabajo consiste en estudiar la percepción ciudadana de la marca *fast fashion* "Shein" en España, identificando las estrategias y factores que han impactado en su reputación y en el éxito en el mercado de la moda. Para ello, se realizará la recopilación de datos a través de la red social Twitter, y se identificarán y analizarán las principales categorías de discusión de la marca, mediante técnicas avanzadas de minería de texto como el modelado de tópicos y el análisis de sentimientos.

1.2.2. Objetivo Específicos

A continuación se desglosa el objetivo general en objetivos más concretos, lo que permitirá analizar y extraer las conclusiones necesarias durante el desarrollo del presente trabajo.

- ***Descripción de la relevancia de la industria fast fashion a nivel global y en España.***

  Se llevará a cabo una descripción de la industria fast fashion a modo de contextualización con el objetivo de entender su modelo de mercado, relevancia e impacto ambiental y económico no solo en el mercado de estudio (España) sino a nivel mundial.

- ***Contextualización del impacto de "Shein" como marca líder en la industria fast fashion en los últimos años en China y otros países.***

  Se considera relevante de cara al estudio propuesto introducir las principales características de la marca Shein en el contexto de la moda fast fashion. Asimismo, se pretende detallar factores importantes de su estrategia de mercado, y de su impacto en el mercado de la moda.

- ***Análisis de los resultados obtenidos con el uso de técnicas de minería de textos, aplicadas al estudio de la percepción ciudadana de la marca en áreas relacionadas con el marketing digital.***

El análisis propuesto se llevará a cabo mediante la implementación de técnicas avanzadas de minería de datos, en concreto *modelado de tópicos* y *Análisis de sentimientos*. Por tanto, se contextualizarán los trabajos realizados en el área y se detallarán los principales resultados de estas técnicas en la identificación de la percepción pública sobre la marca y como herramienta de marketing.

- ***Análisis de la percepción ciudadana en España en redes sociales sobre la marca "Shein", a través del modelado automático de tópicos y el análisis de sentimientos.***

  Finalmente, se recolectarán los datos de la red social de Twitter y se llevará a cabo el análisis automático sobre los datos recopilados. Para ello, se identificarán las categorías de discusión abordadas en twitter y partiendo de estos resultados, se realizará un análisis temporal del volumen de tweets, así como un seguimiento de los sentimientos involucrados en cada categoría. Estos resultados permitirán evidenciar los factores positivos y negativos más relevantes de la marca Shein en España.

1.3. Organización de la Memoria

Este trabajo se estructura en cinco capítulos principales. En el capítulo 1, se proporciona una breve introducción que incluye la motivación del estudio y los objetivos principales de análisis. El capítulo 2 contextualiza la industria de la moda rápida o *fast-fashion*, presenta la marca de estudio Shein y revisa la bibliografía relevante de estudios similares que emplean técnicas de minería de datos. Por su parte, el capítulo 3 describe los procesos de minería de datos y la metodología concreta empleada en este estudio, incluyendo diferentes subsecciones como la obtención y preprocesamiento de datos, análisis de N-gramas, modelado de tópicos y análisis de sentimiento. En el capítulo 4, se presentan los resultados obtenidos en cada una de las partes mencionadas anteriormente. Finalmente, el quinto y último capítulo resume los puntos principales del trabajo y se presentan las conclusiones obtenidas.

Capítulo 2

El fenómeno Shein: Extracción de percepciones de marca a partir de la red social de Twitter

2.1. Relevancia de la industria fast fashion

La industria de la moda ha experimentado un crecimiento significativo durante los últimos 10 años alcanzando un tamaño de mercado para el año 2021 de más de 1500 millones de dolares y con una previsión de incremento de más del 25 % para el año 2026 (Ver Figura 2.1). Este incremento no se ha alcanzado sólamente en términos de ingresos, si no que además, la producción textil requerida en el mercado ha aumentado, a nivel per cápita en el mundo, de 5,9 kg a 13 kg por año desde 1975 hasta el año 2018 (Peters, Sandin, y Spak, 2019). De forma similar, el consumo mundial de ropa ha aumentado a aproximadamente 62 millones de toneladas al año, y se estima que siga incrementándose hasta los 102 millones de toneladas para el año 2030 (Niinimäki et al., 2020). Como consecuencia de este continuo crecimiento, las marcas de moda han doblado la producción de prendas en el periodo 2000-2014, y se espera que la tendencia continue (Remy, Speelman, y Swartz, 2016).

Este enfoque ha generado resultados exitosos, evidenciados en su tasa continua de crecimiento y en las cifras superiores de rendimiento, en comparación con los modelos tradicionales de negocio minorista. Con el fin de responder a estas nuevas necesidades del mercado, han surgido nuevos actores como los minoristas online los cuales ofrecen mayor agilidad en el diseño y menores tiempos de entrega de los nuevos productos. Bajo esta nueva tendencia, las predicciones de demanda requeridas para la planificación en el proceso de producción se hacen en tiempo real, por lo que el tiempo de respuesta es casi inmediato. Esta cualidad, junto a los efectos de la pandemia del Covid-19, han provocado que la tendencia hacia el uso del comercio online haya ido en aumento durante los últimos años. Con los consumidores confinados en sus casas, el mundo de la moda se trasladó al comercio por Internet desde

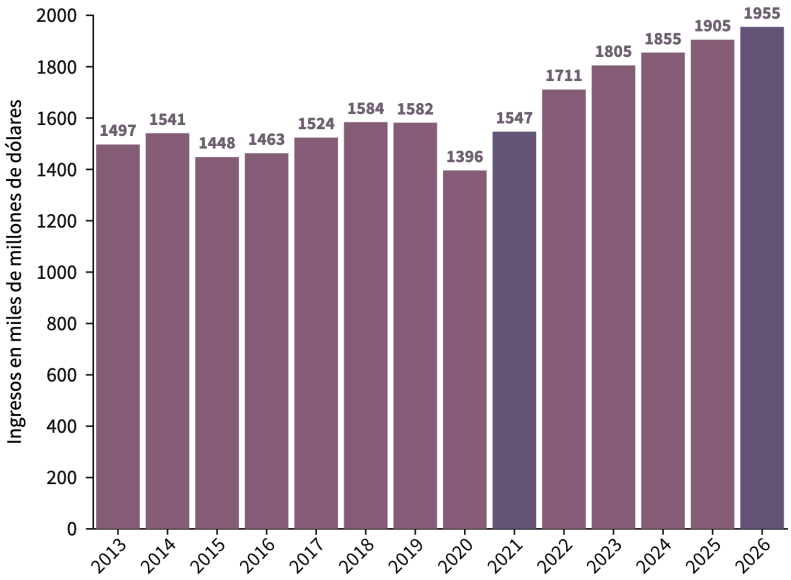
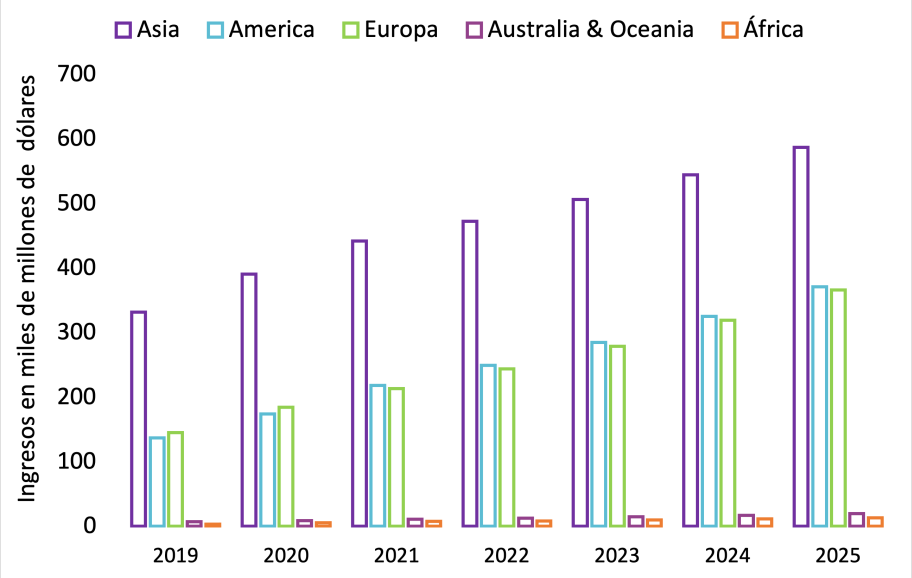


Figura 2.1: Ingresos en la industria de la moda a nivel global
Fuente de Datos: Statista (Apparel Market Statista, 2022) . Elaboración Propia

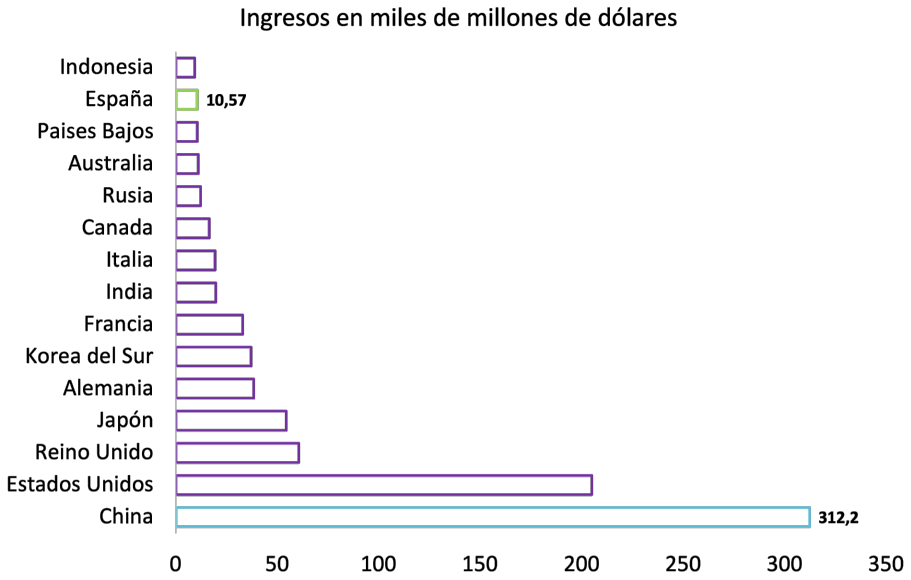
el año 2019, y es así como en el año 2020, se registró un incremento en las ventas online del 16 % al 29 % a escala mundial (McKinsey & Company, 2021). Particularmente, en el año 2022 se espera que este sector del mercado crezca un 10,6 % en el sector de la moda, contando con China como el país con mayores ingresos del comercio electrónico del mundo (Ver Figura 2.2a), y con previsiones que mantienen al mercado asiático como el agente más importante del sector en los próximos años (Ver Figura 2.2b).

En este contexto de la industria *fast fashion*, los fabricantes de moda se han visto obligados a crear cadenas de producción más flexibles y con costos más bajos, con el fin de adaptarse a los altos niveles de demanda y competitividad del mercado. En términos económicos, el mercado mundial de las empresas pertenecientes al sector *fast fashion* alcanzó un valor de casi 68634,9 millones de dólares en 2020 y se espera un crecimiento hasta los 163468,5 millones en 2025 a una tasa del 19,0 %. El resultado de estas previsiones se basa en el aumento de las inversiones extranjeras directas, el crecimiento de los mercados emergentes, y del desarrollo de los medios de comunicación y los avances tecnológicos reportados en los últimos años (Research & Markets, 2021).

En la actualidad, las dos grandes marcas que lideran la industria europea *fast fashion* son H&M y el grupo minorista español Inditex, que posee una familia de marcas que incluye a Zara, Berksha y Massimo Dutti, entre otras. Los resultados financieros de Inditex en 2019 situaron al minorista en la primera posición de una clasificación de marcas de ropa europeas basada en las ventas mundiales, con H&M en segundo lugar (Ver Figura 2.3). En el mercado español, la industria *fast fashion* tiene gran relevancia pues genera alrededor del 2.8 % del PIB (Modaes.es, 2022), siendo Inditex la primera empresa española en adoptar este modelo



(a) Evolución de ingresos en e-commerce a nivel global



(b) Países con mayores ingresos del e-commerce

Figura 2.2: Ingresos del e-commerce a nivel mundial en la industria de la moda
Fuente de Datos: Statista (Ecommerce Countries Statista, 2022; Ecommerce Worldwide Statista, 2022) .
Elaboración Propia

de mercado en el país. En la Figura 2.4 puede observarse un resumen de las 5 principales firmas españolas *fast fashion*, clasificadas por sus ingresos para el año 2019. Como se muestra, Inditex lidera en facturación, tiendas, y empleados por una amplia diferencia con las compañías que le siguen, como Mango y Tendam. Particularmente, Inditex, para el año 2019, superaba los 176.000 empleados, contaba con tiendas en más de 100 países, vendía online en más de 200, y su valor en bolsa se situaba por encima de los 100.000 millones de euros (Miranda y Roldán, 2021).

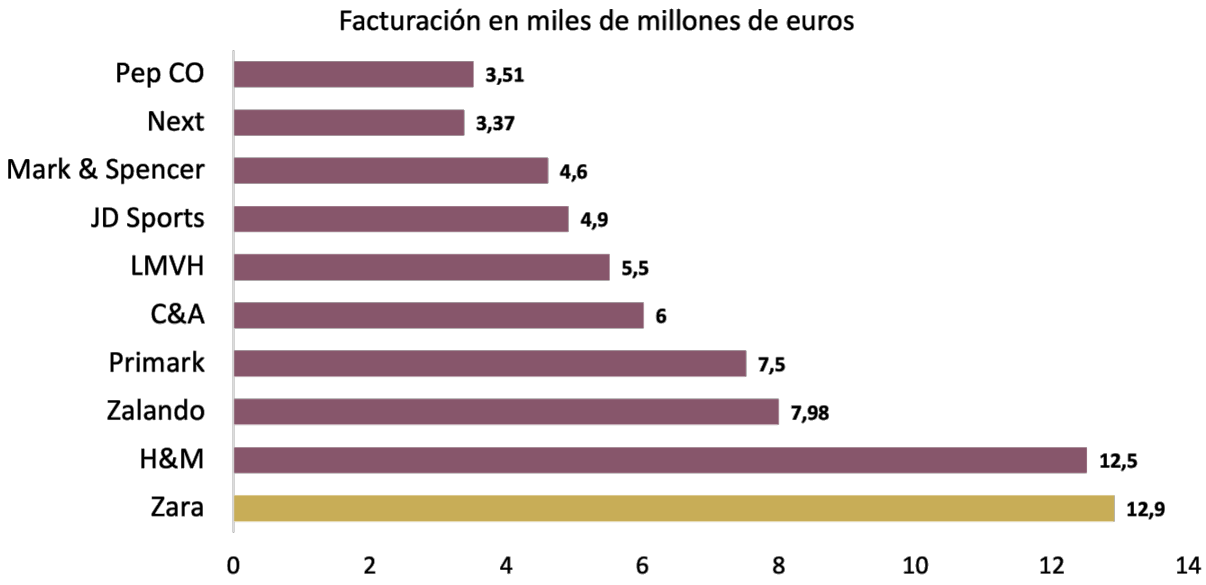


Figura 2.3: Marcas europeas líderes en la industria fast-fashion según su facturación (2019)
Fuente de Datos: Statista (Leading Fashion Statista, 2022) . Elaboración Propia

Empresa	Cadenas	Fundador	Empleados	Tiendas	Facturación (millones de euros)
Inditex	Zara, Massimo Dutti, Bershka, Pull&Bear, Stradivarius, Oysho, Zara Home, Uterqüe, Lefties	Amancio Ortega	176.611	7.469	28.286
Mango	Mango, Mango Man, Mango Kids, Violeta by Mang	Isak Andic	14.892	2.188	2.374
Tendam	Springfield, Women'secret, Cortefiel, Fifty, Pedro del Hierro, Hoss Intropia	Felipe García Quirós	10.735	1.993	1.187
Pepe Jeans	Pepe Jeans, Hackett, Façonnable	Nain, Aurun and Milan Shah	2.701	440	530
Mayoral	Mayoral, Newborn, Nakutavake, Mayoral Shoes, Abel&Lula	Francisco Domínguez	1505	260	365

Figura 2.4: Mayores empresas de moda española en el mercado internacional 2019.
Tabla adaptada de: (Miranda y Roldán, 2021)

2.2. Factores de riesgo: sostenibilidad y medio ambiente

La industria *fast fashion* puede llegar a considerarse como un mercado altamente atractivo, ya que genera ingresos significativos como consecuencia de la creciente demanda y el gran volumen de ventas. Sin embargo, dada la proliferación mundial del mercado *fast fashion* y el volumen de artículos producidos y desechados, esta industria representa una amenaza medioambiental (Castro, 2022). De hecho, las diversas consideraciones ambientales en términos de la reducción de los agentes contaminantes y de un adecuado manejo de los residuos no han sido la principal preocupación de los productores y minoristas del mercado *fast fashion*. En un sentido opuesto, la industria se ha centrado en reducir los costos y los tiempos de entrega de los productos a sus respectivos consumidores (Perry, 2017).

La industria textil es así de las más contaminantes del mundo, ubicandose en segundo lugar tras la industria petrolera, debido principalmente al excesivo uso de recursos naturales. Esto se debe, por un lado, a que se emiten gran cantidad de gases perjudiciales a la atmósfera y por otro, al gran consumo de agua que se genera, pues esta es necesaria en muchos de los diversos procesos de la cadena de producción. Así, la industria de la confección textil y el calzado generó entre el 5 % y el 10 % de los impactos de la contaminación mundial global para el año 2016, donde solo la confección representó el 6.7 % de los impactos climáticos mundiales (3290 millones de toneladas métricas de CO_2eq) (Quantis, 2018). En la figura 2.5 se evidencia el impacto medioambiental de las diferentes etapas involucradas en el proceso textil con respecto a factores como el agua, los recursos naturales, la calidad del ecosistema, la salud y el cambio climático. Como se puede ver, dentro de las diferentes etapas, la asociada a teñido y acabado de las prendas tiene un alto impacto con respecto a todos los indicadores estudiados. Por su parte, la producción de fibra tiene el mayor impacto sobre la extracción de agua dulce y en la calidad del ecosistema, debido principalmente al cultivo de algodón. Finalmente, el alto impacto asociado a las etapas de teñido y acabado y al hilado se debe principalmente al procesamiento intensivo de energía y a la alta dependencia de la energía fósil (Quantis, 2018).

El consumo de agua en esta industria, por su parte, involucra alrededor de 79 billones de litros al año, siendo responsable de un 20 % de la contaminación de las aguas industriales, por el tratamiento y el tinte de los textiles, y de un 35 % de la contaminación primaria por microplásticos en los océanos, material incluido en las fibras usadas para el proceso de fabricación (Kant, 2011; UNFCC, 2018). Además, como consecuencia del corto ciclo de vida que se le da a las prendas, se producen grandes cantidades de desecho que son depositadas en el ecosistema. Según un estudio de Bloomberg, solo en Estados Unidos se desechan más de 11.3 millones de toneladas de residuos cada año, gran parte de los cuales acaban en los vertederos o se queman(Dottle y Gu, 2022).

Por otra parte, es importante mencionar también el impacto social que genera esta industria. Para poder cumplir con los niveles de producción deseados y establecer bajos precios

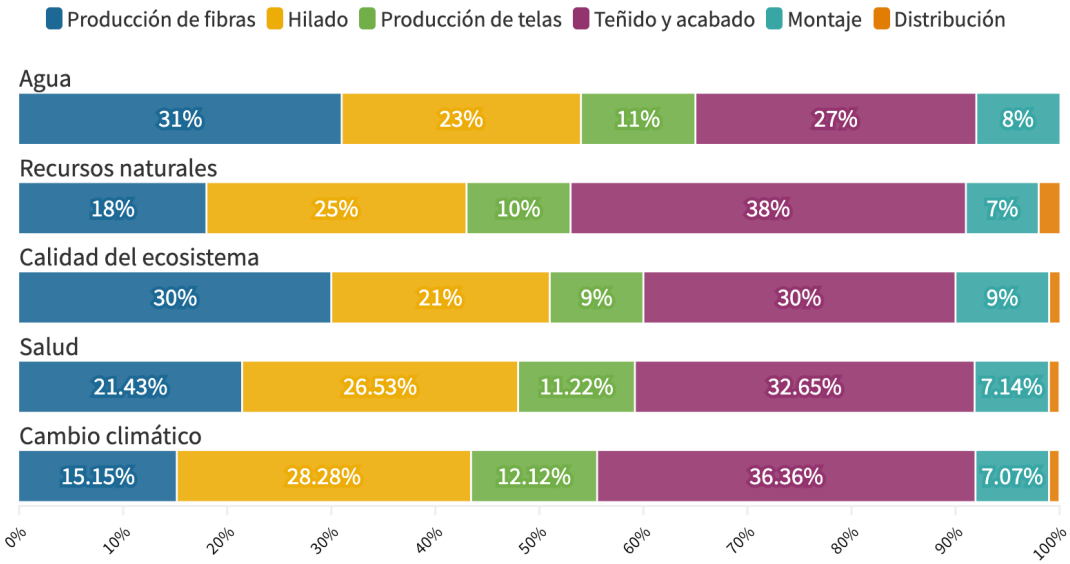


Figura 2.5: Impacto medio ambiental por proceso textil
Fuente de Datos: Quantis (Quantis, 2018) . Elaboración Propia

en las prendas, las empresas ubican sus centros productivos en países en desarrollo, donde la mano de obra es más barata. En muchos de estos países, las condiciones laborales son precarias contando con entre 14 y 16 horas laborales diarias y con un salario medio que no es suficiente para cubrir el costo de vida (Tamayo, 2018). De hecho, un informe del Centro Stern para la Empresa y los Derechos Humanos de la Universidad de Nueva York analizó la situación y encontró que países africanos como Etiopía, en su afán por atraer la inversión extranjera, han promovido los salarios más bajos de todos los países productores de ropa: sólo 26 dólares al mes para el año 2018 (Barret y Baumann, 2019). Muchas de las marcas *fast fashion* más conocidas del mundo, como H&M y PVH, emplean a miles de trabajadores en Etiopía con un salario con el cual son incapaces de mantener a sus familias. Como puede verse en la Figura 2.6, los salarios mensuales son más altos en algunos países en vía de desarrollo, pero siguen siendo extremadamente bajos en general, destacando Bangladesh y Ethiopia, con 95 y 26 dólares mensuales respectivamente. Este fenómeno, conocido como *esclavitud moderna*, se ha asentuado en estos países debido principalmente a la falta de regulación laboral por parte de los respectivos gobiernos y a su alta accesibilidad comercial, ya que los productos cuentan con limitadas restricciones en los procesos de exportación con el resto del mundo.

Con el fin de combatir el grave impacto medioambiental de esta industria, se están llevando a cabo varias iniciativas que ofrecen alternativas de consumo más sostenibles como el modelo llamado *slow fashion*, el cual se opone a la idea base de consumo acelerado del modelo *fast fashion*. Este modelo defiende los derechos de los trabajadores, lucha por la transparencia y propone utilizar tejidos naturales y libres de tóxicos, creando prendas de mayor calidad y durabilidad y a su vez estableciendo un consumo a medio-largo plazo para reducir

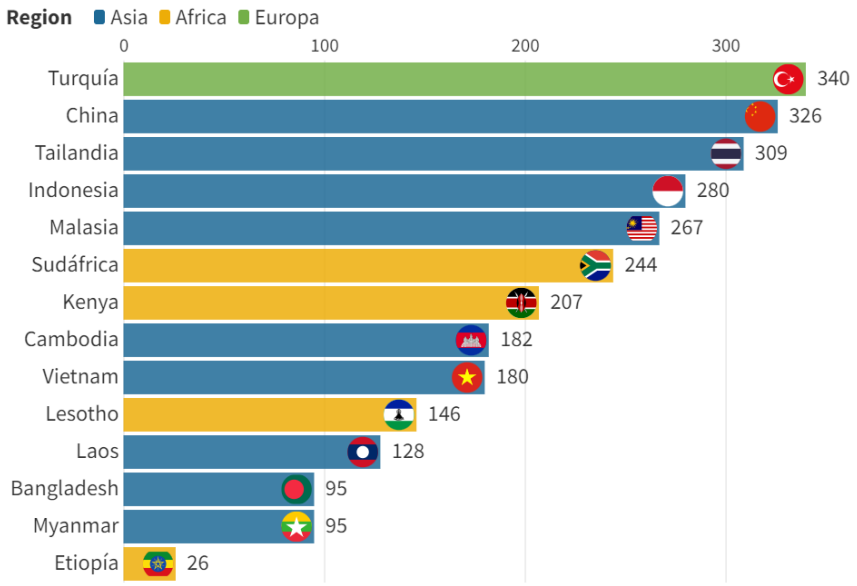


Figura 2.6: Salario mínimo mensual en la industria textil (en dólares) para el año 2018
Fuente de Datos: Statista (Clothing Production Statista, 2018) . Elaboración Propia

la producción y el consumo mediante un incremento de la calidad en las prendas (del Arrabal Fernández Matilla, 2017). Este tipo de enfoques emergentes suponen una amenaza para las firmas de fast fashion, más aún en el caso de que surjan nuevas regulaciones y normativas que fomenten un consumo mas consciente y sostenible para cuidar el medioambiente.

Finalmente, cabe destacar la polémica en la que se ha visto envuelta esta marca continuamente como consecuencia de las denuncias realizadas por marcas competidoras sobre la realización de copias e imitaciones de sus diseños. En varias ocasiones Shein no sólo ha sido acusada de copiar diseños, si no de además utilizar las propias imágenes de producto de otras marcas para promocionar sus propios producto en la web (Eco Club, 2022). Por este motivo, Shein ha recibido numerosas denuncias y se ha visto envuelta en varios escándalos donde los usuarios han denunciado estos hechos a través de redes sociales como Twitter o Instagram. Sin embargo, el problema recae en que es muy dificil patentar un diseño, lo que hace que este mercado sea tan competitivo y que por tanto gigantes como Shein, que cuentan con más recursos y costes mas reducidos, puedan recrear los diseños de otras marcas a precios mas bajos, afectando negativamente sobretodo a pequeñas empresas, que se ven gravemente afectadas por este gigante. Tras recibir dichas denuncias, en algunas ocasiones la marca ha decidido hacer declaraciones públicas pidiendo disculpas para finalmente retirar el producto de su catálogo, pero en otras muchas ocasiones esto no ha sido así (Godoy, 2021). Esta problemática podría afectar de forma muy negativa a la marca en un futuro en caso de que se produzca una regulación de la ley respecto al plagio de los diseños, además de que sus clientes potenciales lleguen a percibir a la marca de forma negativa como consecuencia de dichos actos en perjuicio contra otras marcas del mercado de la moda.

2.3. Shein: Factores de éxito y modelo de negocio

Debido a la creciente demanda de la industria *fast fashion*, este modelo ha evolucionado con el fin de mejorar la velocidad en los procesos de producción y de entrega de productos, lo que ha resultado en conceptos emergentes como el *ultra fast fashion* y el *real time fashion*, cuyo mayor exponente es la firma China Shein.

Shein es una marca de moda ultra rápida fundada en 2008 por Chris Xu en China. Esta marca comenzó con el nombre de Sheinside y en sus inicios, se dedicaba a vender vestidos de novia a precios muy asequibles fuera de China(Moi Global, 2022). Sin embargo, la gran capacidad de producción de este país y la facilidad de acceso a nuevas tecnologías, permitió la expansión de su catálogo más allá de los vestidos de novia, y rápidamente se estableció como una marca reconocida en los mercados de ropa online de todo el mundo. Durante este proceso de expansión, la empresa produjo artículos de moda femenina, de todo tipo, a precios muy bajos. Además, realizó una integración vertical de todos sus procesos con el fin de controlar su cadena de valor y así, acelerar la creación de sus propios diseños y la producción y distribución de todos sus productos(Guptal, 2022). De hecho, con el fin de disminuir sus tiempos de entrega, su centro productivo y de distribución se ubicó en Guangzhou, una localización estratégica de gran relevacia en la industria textil que le ha permitido distribuir a gran velocidad.

Esta reestructuración incluyó también al desarrollo de un *software* propio, donde pudo recibir actualizaciones de los pedidos al instante, atender al amiento de su público y enviar datos en tiempo real. Es bajo estos cambios que en 2017, la versión actual de Shein empezó a tomar forma. La marca además incluyó anuncios en programas televisión diurnos en Estados Unidos, y los *influencers* de la moda se encargaron de mostrar a Shein y a sus botines junto a otros minoristas, como Fashion Nova y Zaful. Sin embargo, fue el uso temprano de TikTok y la capacidad de comercializar productos virales lo que disparó la popularidad de Shein.

A partir de estas iniciativas el interés en Shein ha ido aumentando, de hecho su índice de importancia en búsquedas en Google se ha incrementado gradualmente desde el año 2017 y se disparó en marzo de 2020, periodo en el que empezaron las medidas de confinamiento por la pandemia del Covid 19 en muchos países del mundo (Ver figura 2.7). A diferencia de otros minoristas en años anteriores, Shein no vio un rebote durante la temporada de vacaciones de 2020, pero se recuperó rápidamente en la primavera de 2021, alcanzando un nuevo máximo de interés para junio de ese año, alcanzando y sobrepasando desde esa fecha los índices de búsqueda de la firma Inditex, con la marca Zara y de la firma H&M. Esta gran popularidad ha logrado que Shein irrumpiese en el mercado estadounidense y europeo, y que incrementase sus ingresos exponencialmente en casi un 400 % del año 2017 al 2021 (Ver figura 2.8). Además, su valoración de 5.000 millones de dólares en 2019 alcanzó los 47.000 millones en el año 2021, y se espera que salga a bolsa en el presente año (Business of Apps - Shein, 2022).

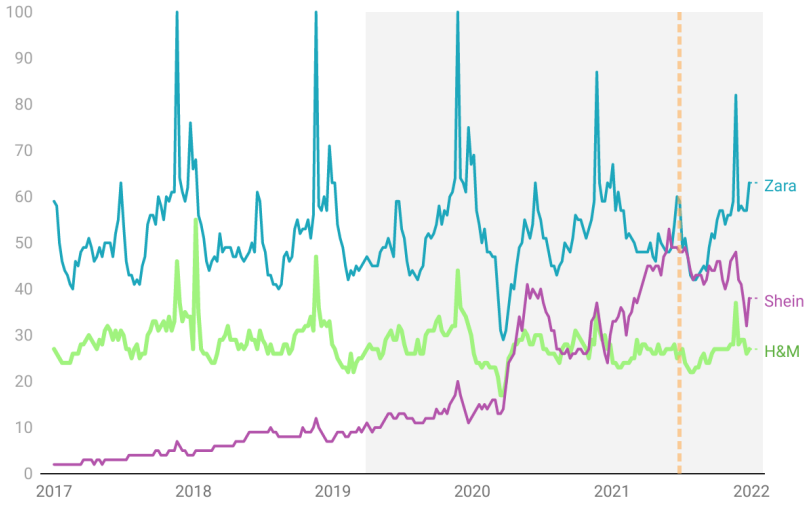


Figura 2.7: Evolución de búsquedas de Google de Shein frente a sus principales competidores
Fuente de Datos: Google trends. Elaboración Propia

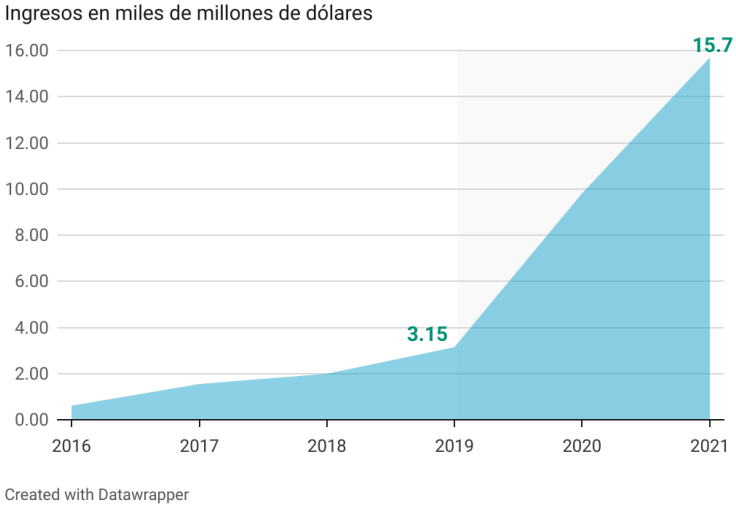


Figura 2.8: Ingresos de Shein en los últimos 5 años
Fuente de Datos: Daxue Consulting (Daxue Consulting, 2022). Elaboración Propia

Actualmente, Shein cuenta con una cantidad inmensa de productos de gran variedad (alrededor de 600.000) en su página web, llegando a introducir hasta 2800 nuevos estilos cada semana (Oxford Analytica, 2021). En España, el éxito de la marca ha sido tal que ha llegado a ocupar el octavo puesto en ventas de todas las empresas de venta online, siendo la segunda dentro de las empresas dedicadas solamente a la venta de productos de moda, con ingresos que suman los 410 millones de dólares, acercándose al grupo Inditex que con su marca Zara registra ingresos en el país de alrededor de 500 millones de dólares (Statista, 2021). En nuestro país, la oferta de diseños vanguardistas, poco accesibles principalmente para los jóvenes, pero con precios muy bajos han hecho de esta marca la más popular en mujeres de 24 años, sobrepasando considerablemente a Zara y al Corte Inglés (Ver Figura 2.9). De hecho,

la media de sus precios es significativamente más baja que las marcas tradicionales de ropa en España. Por ejemplo, en la categoría de vestidos de mujer, como puede verse en la Figura 2.10, Shein ofrece un precio de venta al público medio muy bajo (15 euros), el cual es un 50 % inferior al de Zara o Mango y sólo un 10 % superior al de Primark, marca que cuenta con la media de precios más baja. Asimismo, la moda de su franja de precios se ubica en el intervalo de 10-15 euros, correspondiente al intervalo de precios más bajos respecto a sus competidores (MHE Consumer, 2021).

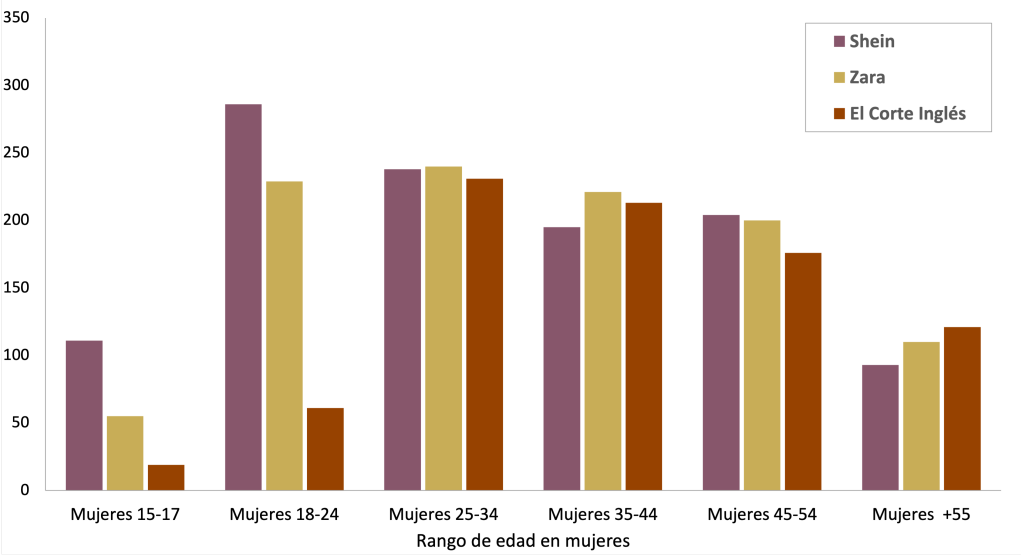


Figura 2.9: Grado de interés en mujeres por rango de edad en las principales marcas de moda
Fuente de Datos: Comscore MMX (MMX, 2022) . Elaboración Propia

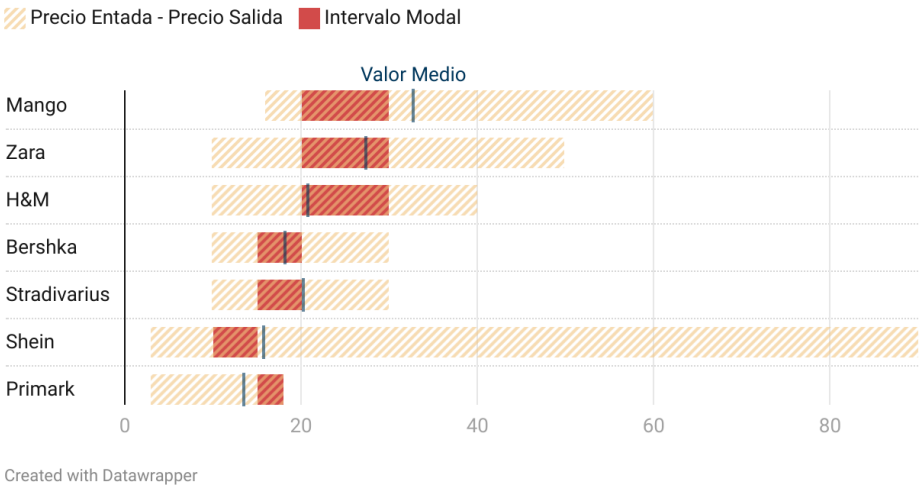


Figura 2.10: Rango de precios en vestidos para mujer por marca
Fuente de Datos: Market Mapping MHE (MHE Consumer, 2021) . Elaboración Propia

Finalmente, es importante destacar que esta marca ha conseguido crear toda una experiencia interactiva para el usuario a través de imágenes, vídeos, reviews, descuentos y premios

y una actualización de productos constante en su canal de venta. El público objetivo de la marca, como se ha visto anteriormente, son principalmente los jóvenes pertenecientes a la generación Z y por tanto, la marca ha hecho gran uso de las tecnologías y redes sociales. De hecho, a día de hoy la marca no invierte en publicidad tradicional, si no que hace un uso inteligente de redes sociales como TikTok e *influencers* (Daxue Consulting, 2022). A través de esta plataforma, la empresa genera contenido que es de especial interés para los jóvenes hasta tal punto que han creado su propia comunidad de microinfluencers. Es así como en términos de aplicaciones móviles, Shein se ha convertido en la número uno del *ecommerce*, siendo actualmente la app comercial más descargada incluso por encima de Amazon, solamente siendo superada por aplicaciones asociadas a redes sociales y de mensajería instantánea (Ver Figura 2.11). En España la situación no es diferente, particularmente en términos de aplicaciones móviles de moda, la Figura 2.12 muestra que para finales del año pasado, Shein era líder indiscutible con una cuota de mercado del 50.3 %, siguiendole Vinted, a gran diferencia, con una cuota de mercado del 37,9 %.

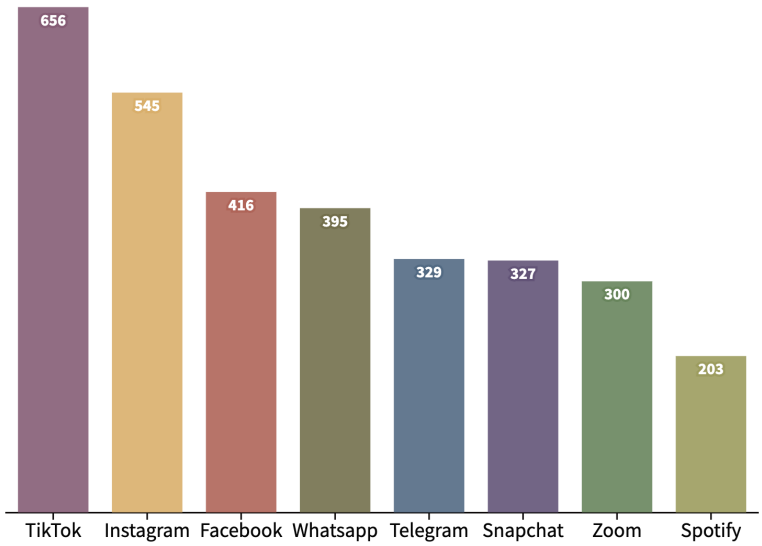


Figura 2.11: Top 10 apps mas descargadas a nivel global en 2021 (en millones de unidades)
Fuente de Datos: Apptopia (Apptopia, 2022) . Elaboración Propia

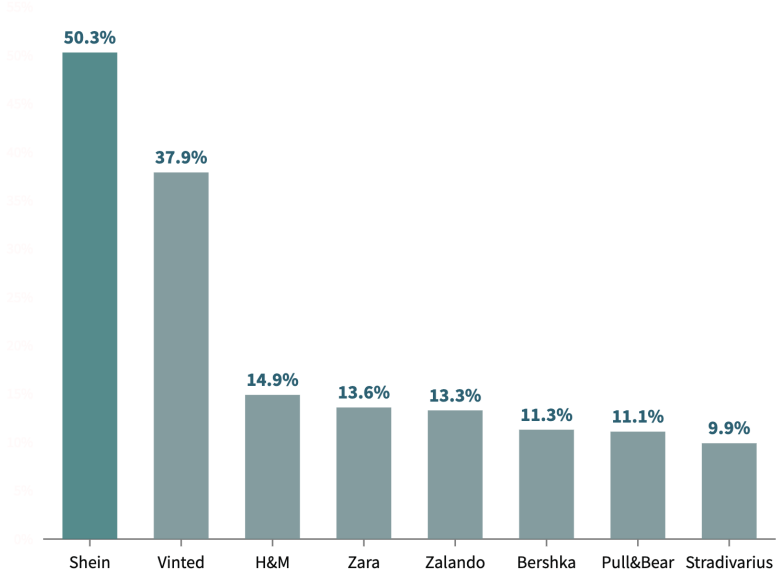


Figura 2.12: Apps de moda con mayor cuota de mercado en España
Fuente de Datos: Smartme Analytics (Smartme Analytics, 2021) . Elaboración Propia

2.4. Técnicas de minería de datos en el análisis de percepciones de marca

Hoy en día, la forma de hacer marketing se ha transformado complementamente como consecuencia de las nuevas tecnologías y las redes sociales. El marketing tradicional, ha perdido capacidad de atracción y se ha visto reemplazado en gran medida por lo que ahora conocemos como marketing digital (Zambrano et al., 2018). Esta nueva forma de hacer marketing se centra mucho más en el cliente y en sus emociones, lo que permite una mayor personalización de los productos y servicios gracias a la gran cantidad de datos e información que se puede obtener a través de la huella digital de internet.

Uno de los métodos mas útiles y más usadas para hacer marketing digital es la minería de datos, también conocido como *text mining*. Esta técnica, aprovecha a través de una serie de técnicas computacionales de procesamiento y análisis, las grandes cantidades de datos generados por la empresa para predecir el comportamiento de los consumidores y tratar de optimizar los productos y servicios que se ponen a su disposición. El objetivo de la minería de datos, es encontrar patrones de comportamiento y características similares entre los consumidores para poder aplicar estrategias de segmentación más precisas y cuyo patrón pueda ser replicado con nuevos conjuntos de usuarios con características similares (Yan, Wang, Wang, y Lin, 2011). Además, otras de las principales ventajas y utilidades de esta técnica aplicada al marketing son (Linoff y Berry, 2011):

- Prevención de fraudes y mejor visión del entorno competitivo
- Mejora del *Customer Relationship Management* (CRM) ó Gestión de la relación con el cliente , ya que permite conocer en profundidad los hábitos de consumo de los clientes para así, proporcionarles un servicio más personalizado.
- Previsión de tendencias y mejora de la toma de decisiones

Las fuentes de información y de extracción de datos en este campo son muy variadas. En particular, las opiniones generadas por la web en blogs y redes sociales se han convertido en un valioso recurso para la extracción de información relevante con fines de gestión de las relaciones con los clientes, seguimiento de la opinión pública y filtrado de textos (Mostafa, 2013). En concreto, Twitter ha sido elegido como la principal fuente de datos de redes sociales por diversos investigadores alrededor del mundo debido a varias razones. En primer lugar, es considerada la plataforma de microblogs más grande, popular y conocida del mundo (Mostafa, 2013), contando con más de 400 millones de usuarios activos al mes (Business of Apps - Twitter, 2022). Por otra parte, los usuarios utilizan la plataforma para expresar libremente sus opiniones y comentarios sobre diversos productos, servicios y marcas (Webster, 2010), y, finalmente, el acceso a los datos de Twitter es abierto, por lo que investigadores y empresarios acceden a un gran volumen de información mediante la interfaz de programación de aplicaciones (API) de Twitter. Estas condiciones proporcionan información sin precedentes, a gran escala, lo que permite abordar análisis de gran relevancia sobre percepciones ciudadanas en diversos campos.

En este contexto, este gran volumen de datos ha permitido alimentar estrategias de análisis automático con el fin de extraer información relevante para empresas e investigadores. Es así como las técnicas basadas en mineria de texto han sido aplicadas previamente para, por ejemplo, extraer sugerencias de las reseñas de productos de los consumidores, clasificar las reseñas de productos por parte de los consumidores en positivas y negativas, rastrear las tendencias de sentimiento en los foros de discusión en línea, determinar la insatisfacción de los consumidores con las campañas publicitarias en línea o diferenciar entre el contenido informativo y el emocional en redes sociales. En términos del análisis del impacto publicitario sobre la marca, se han realizado diversas investigaciones en el campo. En particular, en (Adeborna y Siau, 2014) se emplearon técnicas de minería de texto con información de Twitter, para conocer el rating o valoración que los usuarios daban a tres grandes aerolíneas: AirTran Airways, Frontier and SkyWest Airlines). Para llevar a cabo este estudio, se utilizó la técnica STR (*sentiment topic recognition*), que utiliza el modelo de temas correlacionados (*Correlated Topics Models* - CTM) junto al algoritmo de maximización de la expectativa variable (*Variational Expectation-Maximization* - VEM). Posteriormente, se atribuyó un puntaje a cada aerolínea en función de estos sentimientos encontrados, en términos de variables como el tiempo de embarque, el equipaje, los retrasos y las quejas de los clientes. Los resultados mos-

traron que los usuarios, en general, valoraban como mejor compañía a AirTran, seguida de Frontier. Por su parte, AirWest fue la empresa peor valorada por los usuarios.

Así mismo, en (Al-Hajjar y Syed, 2015), se implementaron estas técnicas para estudiar la reputación de marcas relevantes del sector tecnológico como Apple, Google, Samsung o Microsoft, entre otras. Para ello, se realizó la extracción de sentimientos y emociones de los tweets con fines de marketing digital. Así, para identificar el sentimiento asociado a los tweets, se implementó el diccionario *SentiWordNet* que extrae la polaridad asociada al texto (positiva o negativa), mientras que para la detección de emociones, se utilizó el diccionario NRC (National Research Council Canada's), que permite asociar listas de palabras a un total de 8 emociones (disfrute, anticipación, miedo, tristeza, sorpresa, enfado, asco y confianza). Finalmente, se llevó a cabo la combinación de las puntuaciones obtenidas en ambos procedimientos lo que permitió obtener una visión más completa de la percepción de los usuarios. Así, para validar la propuesta metodológica, se realizó una comparación de la precisión obtenida al realizar el análisis de sentimiento y de emociones de forma aislada frente a hacerlo de forma combinada, demostrando que el nivel de precisión medio es superior si se combinan dichas estrategias.

En la misma linea de investigación, en (Giri, Harale, Thomassey, y Zeng, 2018) se realizó un análisis de percepción de marca de dos grandes empresas de la industria, Zara y Levis, haciendo un análisis de sentimientos a través de la implementación de un algoritmo de clasificación conocido como *Naive Bayes*. Para ello, se realizó una extracción de datos de Twitter y se ha atribuido a los tweets, una emoción determinada, clasificándolos en función de su polaridad. Este estudio ha concluido que la mayor parte de los usuarios presentaban actitudes positivas hacia a ambas marcas lo que se traduce en buena reputación y percepción de marca. Por otra parte, en (Cosentino, 2019) se desarrolló un análisis del impacto de una campaña de marketing muy polémica en la compañia deportiva Nike, la cual generó un movimiento de sabotaje contra la marca en redes sociales. Esta investigación utilizó el análisis Latente de Dirichlet (*Latent Dirichlet Allocation* - LDA) como estrategia de modelado de tópicos y un enfoque basado en 4 diccionarios ("syuzhet" desarrollado en el laboratorio Nebraska Literary, "AFINN" desarrollado por Finn Årup Nielsen, "bing" creado por Minqing Hu and Bing Liu, and "NRC"), para el análisis de sentimientos de las conversaciones en Twitter que incluyeran el hashtag #BoycottNike. Los resultados revelaron que, a pesar de la reacción negativa inicial en Twitter en torno al movimiento, el sentimiento se volvió más positivo con el tiempo, incrementandose las ventas de la marca en un 31 %.

De forma similar, en (Liu y Burns, 2018) se llevó a cabo un análisis de 13 marcas de moda de lujo, entre las que se encontraban Luis Vuitton, Gucci, Prada, Burberry y Versace. Asimismo, se adicionó un análisis compativo con otras marcas no pertenecientes al sector de lujo. En este contexto, se identificaron entonces los tópicos de discusión de las publicaciones en Twitter usando LDA. Los resultados indicaron que las categorías identificadas a las marcas de lujo estaban asociadas a los shows de moda, los productos en sí, y el amor a la marca,

mientras que en las marcas de no lujo, los tópicos mas destacados estaban relacionados con información de producto, promociones y servicio al cliente, entre otras. En cuanto al análisis de sentimiento, se observó que los usuarios perciben las marcas de lujo un 10 % más positivamente que en el caso de las marcas no pertenecientes al sector. En el mismo sentido, se mostró que los consumidores perciben menos negativamente a las marcas de lujo que a las marcas de no lujo.

En este mismo año, se publicó otro estudio en (Barry, Valdez, Padon, y Russell, 2018), donde se buscaba analizar las campañas publicitarias de diferentes marcas de bebidas alcohólicas, mediante el análisis de mensajes en Twitter. El objetivo de este estudio se centraba en analizar el tipo de contenido de los mensajes de las marcas de bebidas alcohólicas, teniendo en cuenta que la mayoría de su público es gente joven, y una gran parte es menor de edad. Para llevar a cabo esta investigación, se analizaron 13 marcas muy conocidas de bebidas alcohólicas como Bacardí, Heineken, Malibu, Jack Daniels y Budweiser, entre otras. Específicamente, este estudio se centró en el proceso de modelado de tópicos, a través de LDA, con el fin de encontrar las categorías más importantes abordadas por dichas marcas al hacer sus campañas de marketing. El resultado fue que, en términos generales, todas las marcas asociaban las bebidas alcohólicas con diversión, fiesta, eventos y deportes, mientras que casi ninguna trasladaba mensajes de concienciación y moderación con el alcohol. Esta investigación, en particular, busca ser de utilidad para las instituciones de salud pública y reguladores a la hora de implementar medidas y supervisar la publicidad que se proporciona en los medios, y de esta manera proteger a los adolescentes de la exposición a las bebidas alcohólicas. En otro estudio posterior, reportado en (Babčanová, Šujanová, Cagáňová, Horňáková, y Hrablik Chovanová, 2021), se propuso un análisis de percepción de marca, esta vez sobre diferentes compañías del sector automobilístico, con el fin de extraer información acerca de su pocisionamiento en el mercado y sobre la opinión ciudadana alrededor de diversas marcas. Para realizar este estudio, se recopilaron tweets exclusivamente en inglés a nivel global y se implementó un análisis de sentimiento para categorizar las palabras según su caracter positivo o negativo. En esta ocasión, se utilizó un enfoque basado en el algoritmo VADER (*Valence Aware Dictionary and Sentiment Reasoner*) y los resultados obtenidos mostraron que Toyota era la marca con mejor percepción (mayor sentimiento positivo), seguida de VW y KIA. A su vez, las que peor percepción ciudadana mostraron fueron Peugeot y Citroen.

En 2022, se ha llevado a cabo por (Feldmeyer y Johnson, 2022), un estudio dónde se implementaron técnicas de análisis de texto, pero esta vez en lugar de tener como objetivo conocer la percepción de una marca, se pretendia extraer información de interés sobre la utilidad y percepción de los usuarios ante un ingrediente culinario, la cúrcuma. Este estudio, se realizó con el objetivo de encontrar oportunidades de desarrollo de este producto en el mercado, ya que la cúrcuma es una especia conocida por ser saludable y aportar gran bienestar, pues tiene propiedades para aliviar inflamaciones, dolores y mejorar problemas

digestivos. En este estudio, se recopilaron 20000 tweets que incluían la palabra "cúrcuma" (turmeric en inglés), y mediante un enfoque de modelado de tópicos junto al análisis estadístico BTM (*Biterm topic modelling*) , se realizó la identificación y agrupación de los tópicos encontrados en grandes categorías, de forma más resumida. Las categorías principales halladas fueron las siguientes: utilidad, beneficios para la salud, funcionalidad del ingrediente para realizar recetas, cuidado de la piel y bebidas. Finalmente, el estudio llevó a cabo un análisis de sentimientos para identificar mejor la actitud de los consumidores frente a este ingrediente, utilizando el diccionario AFINN, el cual puntúa las palabras con valores entre -5 y 5, según su polaridad positiva o negativa. Como resultado de este estudio, se obtuvo una calificación de sentimiento mayoritariamente positiva para el producto estudiado.

De esta manera, como puede verse en los estudios descritos anteriormente, existe gran variedad de investigaciones que han empleado técnicas de análisis automático de texto con el objetivo de extraer información relevante sobre las percepciones de los usuarios y las oportunidades de desarrollo de diversas marcas. Así, en la Tabla 2.1, se presenta la información resumida de la literatura explorada en este capítulo donde podemos observar que dentro de las técnicas de modelado de tópicos, el algoritmo más comunmente empleado es el LDA y en análisis de sentimientos, se destacan principalmente diccionarios como VADER, AFINN y NRC. Por una parte, LDA se destaca por su flexibilidad a la hora de identificar estructuras en los datos independientemente del volumen de datos, obteniendo buenos resultados para documentos largos y cortos. Además se hace énfasis en su bajo coste computacional en conparación con otros modelos propuestos en la literatura. Por otro lado, a la hora de realizar análisis de sentimientos, VADER permite, en comparación con otros diccionarios, identificar con facilidad los emociones contenidas en los datos ya que puede distinguir entre mayúsculas y minúsculas, interpretar signos de puntuación, emoticones y demás símbolos, reportandose resultados con mayor precisión en este análisis. Adicionalmente, este diccionario no requiere de un conjunto de datos de entrenamiento lo que hace que su implementación sea más simple e intuitiva. Los resultados previamente expuestos sobre el análisis de la literatura serán la base para la elección de las estrategias de análisis automático que se implementarán en el presente trabajo de fin de grado.

De cara a la recopilación de tweets, podemos observar que se usan principalmente las cuentas oficiales de las marcas (@) o mencionadas con el #marca. Como es posible ver en los resultados descritos en la Tabla 2.1, las técnicas de análisis de texto pueden ser de gran utilidad para conocer las opiniones de los usuarios ante determinados productos, o de la marca en general, de forma que las empresas puedan tomar decisiones más acertadas y llevar a cabo estrategias de mercadeo más eficientes. En la línea de estas investigaciones, en este proyecto, se recopilará la información de la marca Shein en España a través de Twitter, y se analizará la percepción e interés por la marca, a través de la implementación de estrategias de modelado de tópicos y análisis de sentimientos, de forma consistente a la literatura explorada en este campo.

Referencia	Marca	Número de Tweets	Método de Adquisición	Objetivo	Algoritmo	Resultados
(Adeborna y Siau, 2014)	Aerolíneas: AirTran Airways, Frontier and SkyWest Airlines)	AirTran: 452 tweets, Frontier Airlines: 499 tweets, SkyWest Airlines: 195 tweets	Conjunto aleatorio de tweets de: AirTran Airways, Frontier y SkyWest Airlines	Análisis de sentimiento	STR Model	Categorías modeladas:'En tiempo','Embarque cancelado','Equipaje perdido', 'Quejas del cliente'
(Al-Hajjar y Syed, 2015)	Sector tecnológico: Google, Microsoft, Apple, Samsung, GE, IBM, Intel, Facebook, Oracle y HP	1000 tweets de cada marca	Marcas como keywords	Análisis de sentimiento y Análisis de emociones	NRC y SentiWordNet	Exactitud: Análisis de sentimiento: 37.3 %, Análisis de emoción: 39.2 %, Análisis combinado: 52.6 %
(Giri et al., 2018)	Zara y Levis	702 tweets de Zara y 980 tweets de Levis	#Zara,#Levis	Modelado de tópicos, Análisis de sentimiento	Naive Bayes	Categorías modeladas: "alegría", "tristeza", "ira", "sorpresa", "miedo" y "asco". "disgust"
(Liu y Burns, 2018)	Luis Vuitton, Gucci, Prada, Versacce, entre otras	2 millones de tweets	No especificado	Modelado de tópicos, Análisis de sentimiento	LDA	Categorías modeladas para marcas de lujo: producto, afinidad a la marca y show de moda. Categorias en marcas de no lujo: producto, promoción, servicio al cliente, ubicación de tienda, y afinidad a la marca entre otras. Se obtuvo que las marcas de lujo generan un 10 % más de sentimientos positivos y un 10 % menos de sentimientos negativos con respecto a las marcas de no lujo
(Barry et al., 2018)	Marcas de bebidas alcohólicas: Bacardí,Absolut,Malibu, Heineken,Smirnoff, entre otras	19005 tweets	# de cada marca	Modelado de tópicos	LDA	11 temáticas asociadas a la marca específica: Jack Daniels con temática de vida rockera, Bacardi con una temática que asocia sus promociones a las fiestas, Heineken vinculó sus promociones al deporte, Malibu se centró en temas relacionados con la playa, y Grey Goose se centró en el lujo. 1 sola temática no asociada a la marca: faceta social del consumo de alcohol (socialización y fiesta).
(Cosentino, 2019)	Nike	79184 tweets	#BoycottNike,	Modelado de tópicos, Análisis de sentimiento	LDA, 4 diccionarios:"syuzhet", "AFINN", "bing", y "NRC"	Categorías modeladas: "boicotear Nike", " boicotear Kaepernick", "boicotear NFL y patrocinadores", "implicaciones monetarias de la campaña", "discusión política"
(Babčanová et al., 2021)	Toyota,Citroen, Peugeot,Skoda, KIA, VW	5279 tweets y 5117 retweets	#VW, #Peugeot, #Citroen, #Kia, #Skoda, #Toyota	Análisis de sentimiento	QDA-VADER	La marca mejor valorada fue Toyota (número de tweets, retweets, valor de las palabras positivas polarizadas), seguida de VW, KIA, Skoda, Citroën y Peugeot.
(Feldmeyer y Johnson, 2022)	Se analizó un ingrediente: la Cúrcuma para conocer su utilidad y beneficios	20000 tweets	Tweets que contengan la palabra cúrcuma	Modelado de tópicos, Análisis de sentimiento	BTM, AFINN	Categorías modeladas: Llamada de atención, Beneficios en salud de la cúrcuma, Platos con cúrcuma, Ingredientes combinados con la cúrcuma, Potenciar la salud de la piel y Bebidas con cúrcuma.

Tabla 2.1: Artículos representativos en el análisis de percepción de marca usando técnicas automáticas de analítica de textos .

Capítulo 3

Metodología

El presente capítulo describe la metodología abordada en el desarrollo de este trabajo de fin de grado. Como ya se ha mencionado previamente, el objetivo principal consiste en analizar las percepción publica de la conocida marca de moda Shein, en España, basándonos en la recolección y análisis de las opiniones de los usuarios de Twitter. El estudio desarrollado consta de 6 fases principales (Ver Figura 3.1), que se detallarán en el presente capítulo: 1) Obtención y preparación de los datos, 2) Preprocesamiento de datos, 3) Análisis descriptivo por medio de N-Gramas, 4) Modelado de tópicos y 5) Análisis de sentimiento.

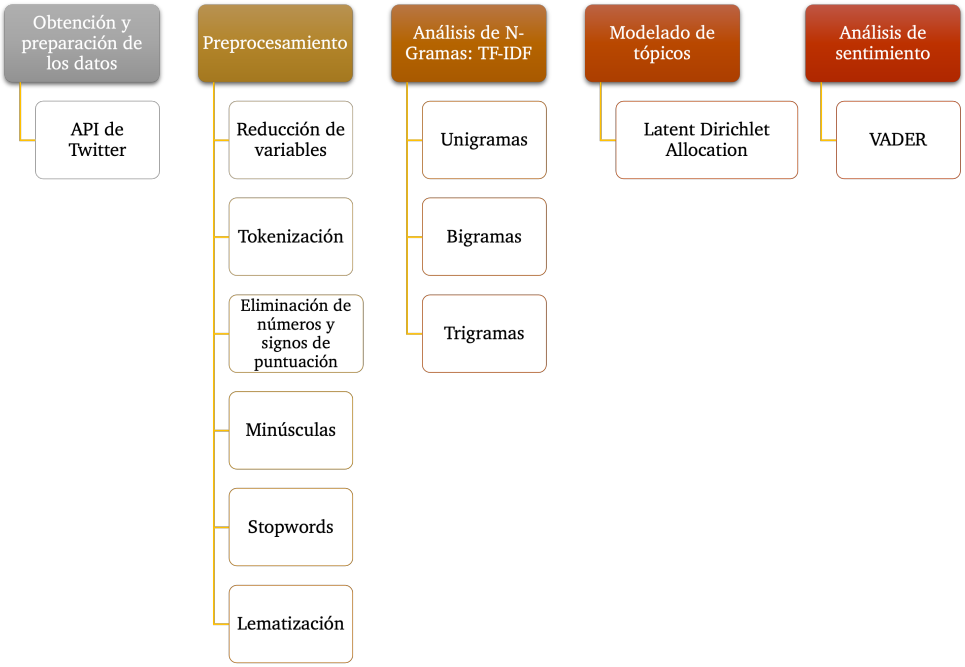


Figura 3.1: Metodología de análisis
Elaboración propia

3.1. Obtención y preparación de los datos

La primera parte del análisis presentado en este trabajo consiste en la recolección de los datos desde la red social Twitter. De esta manera, el objeto de estudio se centrará en tweets recopilados de las cuentas oficiales de la marca (@Shein), así como los hashtags empleados por la marca a la hora de hacer publicaciones. Para la recolección y descarga de estos datos, se realizó la solicitud de acceso a la interfaz de programación de aplicaciones (API) de la plataforma mediante la creación de una cuenta de desarrollador. Una API, es una herramienta que contiene un conjunto de protocolos usados para desarrollar e integrar el software de diversas aplicaciones, como en este caso, de Twitter. El acceso a los datos no está garantizado de forma pública, por lo que se requieren un conjunto de contraseñas (clave de la app, tokens) proporcionados por la plataforma a las cuentas de tipo desarrollador. Así, mediante este acceso, es posible consultar a la aplicación sobre el contenido de diversos tweets, incluyendo métricas de gran utilidad en análisis de redes sociales como el número de respuestas, retweets, contenido de menciones, hashtags, etc.

En este estudio, se realizó la descarga de todos los datos disponibles en el periodo comprendido entre Enero de 2017 hasta la fecha actual, Enero de 2022. La elección de esta ventana pretende evaluar la evolución de la marca desde un periodo anterior a su significativa evolución, la cual como se vio en el capítulo 2 incluye principalmente el periodo de pandemia donde la marca experimentó un gran crecimiento y la ha llevado a su actual posicionamiento en el mercado. Asimismo, como el alcance de este proyecto está enmarcado en el análisis de redes sociales en España, se delimitará la búsqueda a aquellos tweets que hayan sido georeferenciados en España. Finalmente, teniendo en cuenta los resultados del análisis de la bibliografia expuesto anteriormente, se utilizará como palabra clave de búsqueda el nombre de la marca (*shein*). Esta palabra clave engloba la mayor cantidad de hashtags y menciones asociados y referenciados a la marca.

Con el fin de mejorar la calidad de los datos recolectados, es importante realizar un proceso de limpieza y pre-procesamiento. Para ello, en primer lugar, se generó una tabla con la información recolectada y se realizó la selección de las variables que permitirán el análisis descriptivo de los datos. El dataset original, consta de 14 variables, sin embargo, puesto que no todas las variables aportan valor de cara al estudio descrito en este trabajo, se reducirá el conjunto de datos a 11 variables representadas en la tabla 3.1.

De esta manera, como resultado, se desarrollará el análisis con las variables indicadas en la tabla anterior, incluyendo además, dos variables adicionales: *año* y *mes*, obtenidas a partir de la variable *fecha*, ya que puede ser de utilidad de cara a realizar el filtrado, análisis y visualización sobre series temporales. Finalmente, teniendo en cuenta que el filtrado de los tweets por geolocalización en España no garantiza que el idioma de las publicaciones sea el español, se filtrarán los tweets teniendo en cuenta la variable adquirida *lang*.

Nombre de la variable	Significado	Tipo de variable
author_id	identificador del usuario emisor del mensaje	caracter
text	contenido del mensaje o tweet	caracter
fecha	fecha completa	fecha
lang	idioma del texto	caracter
reply_number	número de respuestas del tweet	entero
like_number	número de likes	entero
hashtags	hashtags empleados por el usuario	lista
users	cuenta del usuario	lista
retweet_number	número de retweets	entero

Tabla 3.1: Variables elegidas de los datos recolectados

Teniendo como base el resultado de la etapa de filtrado, descrita anteriormente, se realizó el pre-procesado de los datos resultantes, etapa considerada de gran importancia para el análisis automático de texto o *text mining*(Uysal y Gunal, 2014). Este proceso, requiere de la limpieza y preparación del texto para su posterior análisis. Para ello, en primer lugar, se llevará a cabo la *tokenización* de los datos. Tokenizar consiste en desglosar el texto en unidades más pequeñas, llamadas tokens, lo que permite trabajar con los datos de forma más sencilla. Este procedimiento, permite trabajar los datos en forma de lista de palabras individuales en lugar de un conjunto de palabras relacionadas, simplificando así el proceso de análisis. Un ejemplo de este proceso puede evidenciarse al tokenizar una oración como ”La capital de España es Madrid”, en la cual se obtendrá como resultado “La”,“capital”,“de”, “España”, “es”, “Madrid”.

Una vez convertido el texto en tokens, se deben eliminar aquellos caracteres especiales o signos no relevantes para el proceso de modelado de tópicos, pues no aportan información sobre el contenido de las temáticas y distorsionan el mensaje. Estos pueden incluir desde signos de puntuación (;, ? ,:) hasta caracteres especiales como @,&, %, etc. o incluso caracteres numéricos. En la misma línea, se eliminarán URLS, links, menciones y hashtags de Twitter, así como los emoticonos, con el fin de mantener solo aquella parte del texto que incluya contenido relevante sobre las temáticas discutidas en las publicaciones. A continuación, para asegurar que el texto está estandarizado, se normalizan todas las palabras a caracteres en minúsculas. De esta forma, se evitará confundir palabras escritas con distinto formato. Una vez convertido el texto a minúsculas, se realizará la eliminación de palabras vacías o *stop words* para mejorar la eficiencia del análisis y reducir el volumen del texto. Las *stop words* son palabras o caracteres que no aportan información de contenido, ni son relevantes en el análisis del texto como artículos (el, la, un, entre otros), conjunciones (y, e, ni, entre otras) o preposiciones (de, a, con, entre otras), entre otras. Finalmente, tras haber realizado la eliminación de

stop words, procederemos con la fase de lematización. Esto consiste en reducir las palabras a su forma base, o raíz, eliminando las conjugaciones y reduciendo así la complejidad del texto y su dimensionalidad. Por ejemplo, de la palabra "comí" o "comeré", su lematización sería "comer".

3.2. Análisis de N-Gramas

Tras haber realizado la adquisición y el preprocesamiento de los datos, se llevará a cabo una análisis exploratorio y descriptivo de las secuencias de palabras con mayor relevancia en el conjunto de datos recolectado. Para ello, se utilizará el análisis descriptivo de N-gramas, definiendo un N-grama como una secuencia consecutiva de elementos o en este caso, de palabras. El análisis de N-gramas es una técnica usada en minería de texto que permite la identificación y el análisis de patrones y relaciones entre los datos de texto. Existen diferentes niveles de análisis que se pueden realizar mediante la descripción de N-gramas: El análisis de unigramas, los cuales permiten idenitifcar la importancia en el conjunto de datos de un único elemento o palabra, los bigramas idenitifican secuencias consecutivas de dos palabras, los trigramas de tres palabras y así sucesivamente.

Mediante los resultados del análisis de N-gramas, se puede calcular una métrica de relevancia que se basa en parte en la frecuencia de los conjuntos de palabras dentro del texto. Este proceso, se lleva acabo como forma preliminar para entender los conceptos principales del conjunto de datos e identificar de forma exploratoria algunos de los tópicos de discusión. Sin embargo, es importante destacar que no solo se quiere calcular la frecuencia absoluta de cada N-grama, sino que se quiere brindar una medida de su relevancia dentro del conjunto de datos. La medida a utilizar se conoce como *Term Frequency - Inverse Document Frequency (TF-IDF)*. TF-IDF es definida como una métrica estadística que permite identificar la importancia de una palabra dentro de un documento o conjunto de documentos (en nuestro caso en una colección de tweets), y que se calcula mediante el producto de dos términos:

- **Frecuencia del Término (TF)**: mide la frecuencia normalizada con la que una palabra aparece en un texto. Se calcula como el número de veces en los que aparece la palabra t en el documento d, en relación con todos los demás términos de dicho contenido. Se considera una forma de medir la densidad de palabras clave.

$$TF(t,d) = f(t,d) \tag{3.1}$$

- **Frecuencia Inversa del Documento (IDF)**: este factor pretende reducir el peso de aquellos términos que se repiten mucho en el número total de observaciones y aumenta el valor de aquellas palabras con un menor frecuencia. Se calcula como el logaritmo

del número total de documentos N entre el número de documentos que contienen la palabra df_t.

$$IDF(t, N) = log\frac{N}{df_t} \tag{3.2}$$

Finalmente, la métrica TF-IDF será el resultado de la siguiente operación:

$$TF - IDF = TF(t, d) \times IDF(t, N) \tag{3.3}$$

Así, a partir de este cálculo, se obtiene una puntuación que refleja la importancia de una palabra dentro de un documento o conjunto de documentos. En términos de la interpretación de los resultados, si se obtiene una puntuación de TF-IDF elevada, entonces la palabra será muy relevante dentro del conjunto de datos analizado, mientras que una baja puntuación TF-IDF indicaría que la palabra es poco relevante para el conjunto de observaciones. Esta métrica por tanto, permite proporcionar un TF-IDF alto a aquellos términos con mayor frecuencia en un solo tweet y menor frencuencia de aparición en el total de tweets analizados.

3.3. Modelado de tópicos

El modelado de tópicos se define como una estrategia de análisis que consiste en identificar los tópicos o categorías principales de discusión que mejor representan el contenido del conjunto de datos recolectado.

En el estudio presentado en este trabajo, se realizará la implementación de una estrategia de modelado de tópicos conocida como *Latent Dirichlet Allocation (LDA)* . Este enfoque utiliza como base un modelo estadístico que considera al conjunto de observaciones como documentos, los cuales contienen una serie de tópicos o categorías, que son representados a su vez por un conjunto de palabras. Bajo estas consideraciones, esta técnica permite representar la combinación de tópicos y de palabras por medio de distribuciones de probabilidad, particularmente haciendo uso del modelo de Dirichlet (Onan, Korukoglu, y Bulut, 2016). Hay varias razones por las que LDA se ha convertido en una opción ampliamente implementada para el modelado de tópicos (Sepúlveda, 2016):

- LDA es un modelo probabilístico, lo que significa que puede proporcionar no solo los temas más probables para un documento, sino también la probabilidad de que cada una de las categorias haga parte del documento. Esta característica es de gran utilidad en casos en los cuales se requiera estimar la incertidumbre de las asignaciones de las temáticas.
- LDA es flexible y puede descubrir estructuras y patrones ocultos en datos textuales, como la presencia de múltiples temas superpuestos en un documento o en general, el

nivel de relación entre categorias.

- LDA es un método de ML no supervisado, lo que significa que no requiere ningún dato etiquetado de entrenamiento. Esta consideración lo hace particularmente útil cuando los datos etiquetados son escasos.

En resumen, LDA es un método útil y ampliamente utilizado para el modelado de tópicos, especialmente cuando el objetivo es descubrir la estructura compleja presente en una gran colección de documentos. El funcionamiento general de este modelo se muestra en la Figura 3.2:

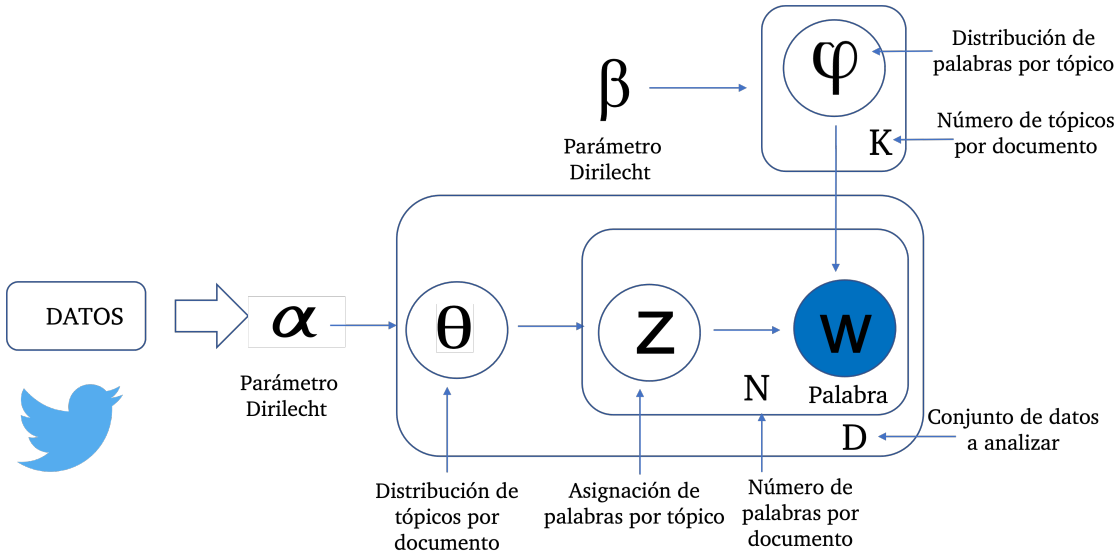


Figura 3.2: Representación gráfica del modelo LDA
Fuente de Datos: Matlab. Elaboración propia

Como se muestra en la figura, hay tres niveles principales de representación del modelo. D representa el conjunto de tweets a analizar, el cual es representado por un número N de palabras representadas por la variable w. Los parámetros α y β permiten modelar la distribución de probabilidad del número de tópicos por documento, y del número de palabras por tópico, respectivamente. Así, valores altos de α y β representarían una mayor proporción de tópicos en los documentos y de palabras por tópico, respectivamente. El parámetro K, por su parte, representa el número de tópicos por documento, el cual debe ser preestablecido en la implementación del modelo y ϕ la distribución de palabras por cada tópico representado. Finalmente, la variable Z caracteriza la asignación de cada palabra dentro del tópico y θ representa la distribución tópicos encontrados en cada documento. Así, en resumen, este enfoque busca modelar la distribución de probabilidad de que una palabra pertenezca a un tópico y de que un tópico pertenezca a un documento. El resultado del análisis permitirá obtener un conjunto de tópicos característicos del conjunto de datos, representados a su vez por el conjunto de palabras con mayor probabilidad de pertenecer a dicha categoría.

Como se ha descrito con anterioridad, el modelo recibe como entrada el número de tópicos K presentes en el documento. De esta manera, para elegir el valor de K que obtiene el mejor modelo, se utilizará el *índice de coherencia*. Esta métrica determina el nivel de cohesión textual, o lo que es lo mismo, el nivel de similitud semántica entre las palabras de un tópico. El índice de coherencia se obtiene a partir de distintas características del texto, como la repetición de palabras clave, el uso de pronombres, o la estructura de las oraciones entre otras. Esto se hace mediante la segmentación de las palabras más importantes y una medida de confirmación indirecta que utiliza la información mutua normalizada puntual (NPMI) y la medida de similitud del coseno. (Röder, Both, y Hinneburg, 2015). Cuanto mayor sea el valor del índice de coherencia, mayor será la calidad del texto.

Por otro lado, se debe calcular y representar la distancia intertópica. Esta métrica, mide la disimilitud entre los tópicos del texto y representa los tópicos en un espacio circular bidimensional, siendo el área de estos tópicos proporcional a la cantidad de palabras que pertenecen a cada tópico en el diccionario completo. El mapeado de estos círculos se realiza utilizando un algoritmo de reducción de dimensionalidad según las palabras que lo componen, de tal forma que los tópicos más cercanos entre sí, tendrán más palabras en común (Carson Sievert, 2014).

3.4. Análisis de sentimiento

El análisis de sentimiento es una técnica empleada en el procesamiento y el análisis automático de textos para identificar y extraer información de carácter subjetivo a partir de los datos. Su función principal es determinar el sentimiento asociado a un texto, siendo su uso más común el cálculo de la polaridad, la cual permite su clasificación como positivos, negativos o neutros. Este enfoque es de gran utilidad para muchas empresas pues permite mejorar la experiencia de los clientes mediante la extracción de información sobre sus percepciones ante un determinado producto o servicio o la identificación de sus posibles reacciones a determinadas publicaciones en redes sociales.

En nuestro caso particular, se busca desarrollar un análisis de sentimiento sobre los datos adquiridos de Twitter, asociados a la marca de moda Shein. El objetivo es identificar cómo se sienten los consumidores de la marca en España frente a sus productos, promociones, o cualquier variable relacionada con la experiencia e interacción con la marca. Como resultado del análisis descrito en el capítulo 2, pudo verse que la estrategía más usada en este campo es el uso de diccionarios. Un diccionario en términos de análisis de sentimientos, es una lista de palabras y frases, que junto con sus etiquetas indican un sentimiento o emoción asociado a cada palabra. Actualmente, se cuenta con una gran diversidad y disponibilidad de diccionarios que permiten llevar a cabo esta tarea, cada uno con su conjunto particular de palabras y etiquetas. En este contexto, el algoritmo elegido en este trabajo de grado es el diccionario

VADER *(Valence Aware Dictionary and Sentiment Reasoner)*.

VADER es una herramienta de análisis de sentimientos muy popular, basada en una serie de reglas y léxicos, específicamente diseñados para ser sensibles a los matices de los sentimientos expresados en redes sociales (Pano y Kashef, 2020). Esta característica se obtiene porque el análisis considera elementos adicionales al texto, como la presencia de emoticonos, caracteres especiales o las distinciones entre mayúsculas y minúculas, lo cual aporta mayor precisión. Este enfoque es usado principalmente cuando se analizan textos informales y breves (como el caso de un tweet o cualquier otro tipo de publicación en redes sociales). Asimismo, este diccionario es simple de usar y no requiere datos de entrenamiento, ya que incluye una lista de palabras y frases asociadas a un sentimiento mediante el uso de diversas puntuaciones.

Este diccionario permite calcular la polaridad (si el sentimiento es positivo o negativo) así como su respectiva intensidad (como de positivo o negativo es el sentimiento), proporcionando una puntuación de valencia a cada palabra. Esta puntuación asigna a las palabras un valor en una escala entre -4 y 4, donde -4 representa el sentimiento más negativo, 4 el más positivo y 0 el punto medio, caracterizando un sentimiento neutral. Para llevar a cabo este cálculo, VADER hace uso de un conjunto de cinco reglas (Hutto y Gilbert, 2014):

- La **puntuación**, como por ejemplo el uso de signos de exclamación (¡!), aumenta la magnitud de la intensidad sin modificar la orientación semántica. Por ejemplo, "¡¡¡La pelicula es mala!!" es más intenso que "La pelicula es mala."

- Las **mayúsculas**, en concreto el uso de TODAS MAYÚSCULAS para enfatizar una palabra relevante para el sentimiento en presencia de otras palabras sin mayúsculas. Esta característica aumenta la magnitud de la intensidad del sentimiento sin afectar la orientación semántica. Por ejemplo, "¡¡¡La pelicula es MALA!!" es más intenso que "La pelicula es mala."

- Los **modificadores de grado** influyen en la intensidad del sentimiento aumentándola o disminuyéndola. Por ejemplo, "La película es extremadamente buena" es más intenso que "La película es buena", mientras que "La película es ligeramente buena" reduce la intensidad.

- La **conjunción contractiva** "pero" señala un cambio en la polaridad de los sentimientos, siendo el sentimiento dominante del texto que sigue a la conjunción. Por ejemplo, "La película es buena, pero la sala del cine era muy incómoda" tiene un sentimiento mixto, siendo la segunda mitad, la que dicta la calificación general

- Al examinar el trigrama que precede a un elemento léxico cargado de sentimiento, se detecta casi el 90 % de los casos en los que la negación invierte la polaridad del texto. Una **frase negada** sería "La película no es tan buena".

Para analizar el grado de sentimiento, se usa la puntuación compuesta, calculada sumando las puntuaciones de valencia, ajustando dichas puntuaciones en función de las reglas descritas previamente y normalizándolas. De esta manera, el resultado final tendrá valores comprendidos entre -1 y 1. Por último, es importante destacar que para llevar a cabo el análisis de sentimientos, ciertos elementos del preprocesamiento serán modificados de tal manera que se facilite el cálculo de la puntuación compuesta descrita anteriormente. De esta manera, teniendo en cuenta la sensibilidad de VADER hacia las puntuaciones, las mayúsculas y algunas conjunciones, se mantendrán los caracteres especiales, emoticonos, y stop words y no se realizará la normalización a minúsculas. En este contexto, sólo se eliminarán hashtags, URLS, menciones y números.

Capítulo 4

Resultados

En este capítulo, se presentan los resultados obtenidos a partir del procedimiento llevado a cabo según la metdología descrita en el apartado anterior. El código utilizado para el desarrollo de este análisis automático de publicaciones de redes sociales, ha sido subido al repositorio github, siendo de acceso libre para su consulta donde el preprocesamiento del texto ha sido analizado en R y los modelos han sido probados en Python (Moreno, 2023).

4.1. Análisis Descriptivo

Antes de llevar a cabo el análisis automático de texto de los tweets asociados a la marca Shein, es crucial realizar un análisis descriptivo inicial para comprender la estructura de los datos, además de tener una primera idea de la distribución de las palabras en el corpus a analizar. Para ello, se han recolectado todos los tweets relacionados con la marca durante los últimos cinco años. Esta elección se debe a que se necesita una cantidad significativa de datos para obtener conclusiones relevantes y el volumen de tweets relacionados con la marca en los primeros años de creación de la marca es bajo. Esto se debe a que, en ese momento, aún no se podía apreciar el impacto de su reciente plan de expansión internacional, el cual comenzó en el año 2015, y se acentuó en el periodo de pandemia. En consecuencia, el número total de tweets obtenidos para el periodo de análisis comprendido entre 2017 y 2022 ha sido de 6.422, los cuales se han utilizado para realizar el análisis descriptivo inicial.

En primer lugar, se realizó un análisis de la frecuencia de los tweets obtenidos durante el periodo de tiempo seleccionado. Tal y como se puede apreciar en la Figura 4.1, el volumen de comentarios relacionados con la marca en Twitter ha experimentado un crecimiento significativo, especialmente a partir del año 2020. Es probable que este aumento esté relacionado con el crecimiento y posicionamiento de la marca en dicho periodo, como resultado de su plan de expansión y el punto de inflexión generado por la pandemia. Este crecimiento no solo ha afectado a Shein, sino que también ha tenido un impacto generalizado en la industria fast-fashion y el consumo de moda en línea.

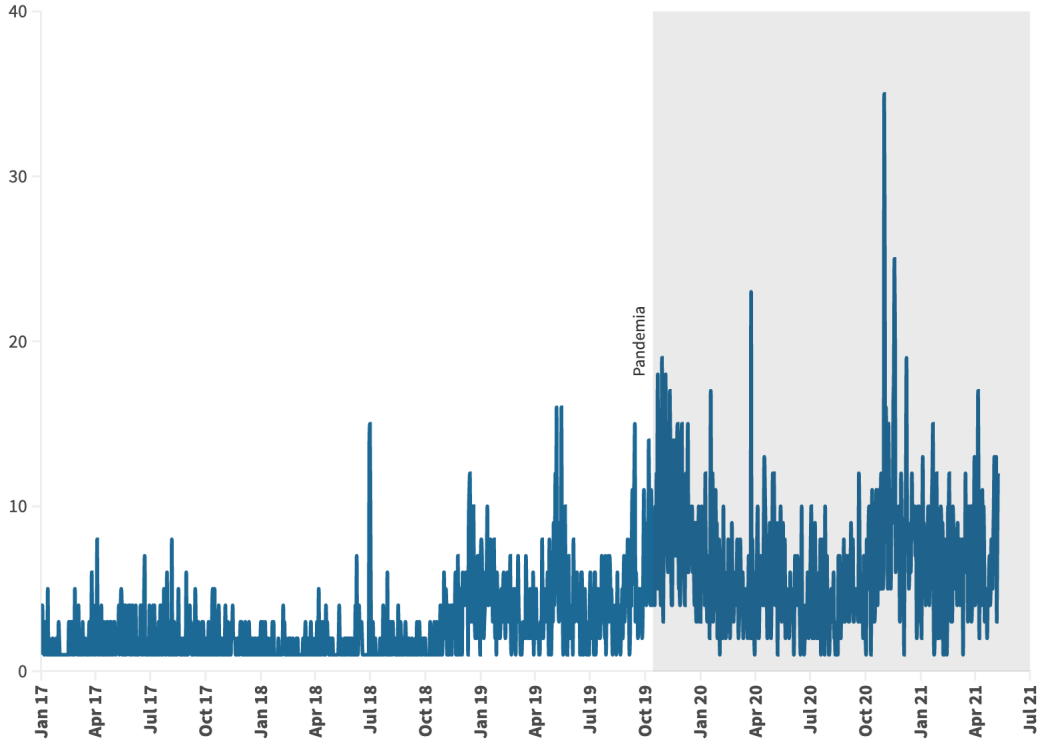


Figura 4.1: Número de tweets por fecha
Elaboración propia

Una vez se analizó la distribución temporal de los datos recolectados durante el periodo de análisis, se procedió a analizar el idioma de los tweets para asegurarnos de que nuestro enfoque se centrara en el mercado español. Para ello, se utilizó la variable "lang" proporcionada en la descarga de datos de la API de Twitter, que determina el idioma asociado a cada mensaje y filtramos por idioma español. Sin embargo, en nuestro conjunto de datos se puede apreciar aun así, una distribución donde aproximadamente el 85 % de los tweets están en español, mientras que el 15 % restante está en otros idiomas o se clasifican como indefinidos debido al uso de signos no reconocidos, como emoticonos o imágenes. De los tweets en otros idiomas, el inglés es el idioma más común, seguido de aquellos que contienen contenido no verbal, como emoticonos. De esta manera, se puede evidenciar que la gran mayoría de los tweets estan centrados en la población objetivo, por lo que se ha decidido filtrar los tweets en idioma español para continuar con el análisis. Asimismo, en la Figura 4.3 se puede observar los emoticonos más usados dentro del conjunto de datos, lo cual permite obtener una primera aproximación a las emociones expresadas por los usuarios que han empleado estos elementos como forma de expresión e interacción con la marca. La mayoría de los emoticonos que se muestran se pueden asociar con sentimientos positivos hacia la marca, como los que incluyen corazones, estrellas o incluso lágrimas de risa, que son los que con más frecuencia aparecen. Sin embargo, el quinto emoticono más frecuente si que podría estar asociado a comentarios más negativos o insatisfacción, como al retraso de un pedido o un pedido equivocado.

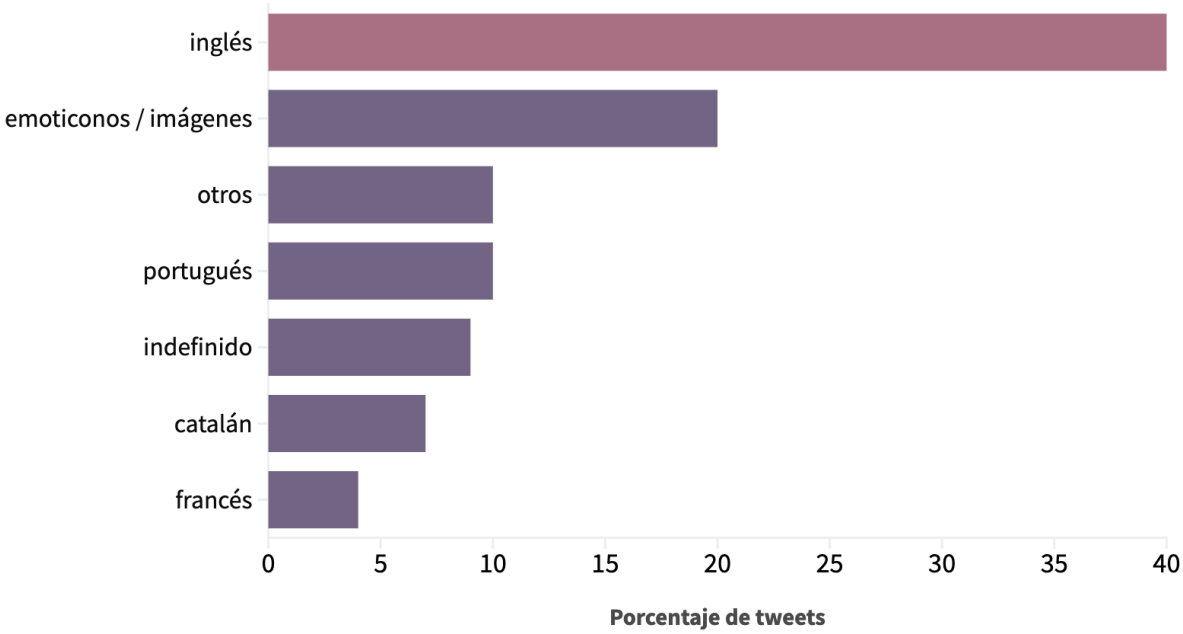


Figura 4.2: Frecuencia relativa de tweets por idioma identificado
Elaboración propia

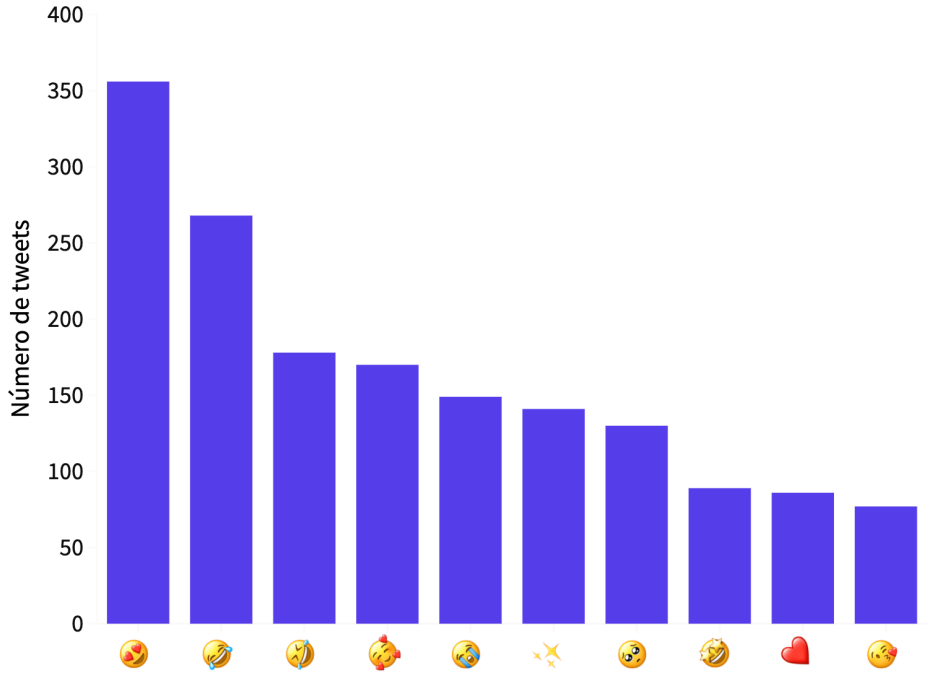


Figura 4.3: Emoticonos más utilizados
Elaboración propia

Por otro lado, es relevante identificar los hashtags más utilizados por la comunidad de Twitter a la hora de escribir tweets relacionados con la marca, ya que aquellos más utilizados

suelen implicar una mayor visibilidad a la hora de publicar contenido y suelen generar un mayor impacto sobre las acciones de la marca y el resto de usuarios. En este caso, de un total de 1280 hashtags diferentes identificados en los tweets recolectados, aquellos con mayor frecuencia de uso entre los usuarios han sido los relacionados con el propio nombre de la marca #shein. Este resultado es esperable debido a que se trata del propio nombre de la marca, que ha sido palabra clave para la adquisición y descarga de los tweets, por lo que para obtener un mejor análisis, se ha eliminiado este hashtag con el fin de estudiar que otros hashtags aparecen con gran frecuencia. Por tanto como se puede observar en la siguiente Figura 4.4, encontramos que los más empleados por los usuarios son #fashion, #ootd y #moda entre otros. En general, se puede observar que la gran mayoría de los hashtags empleados en los datos recolectados hacen referencia o bien directamente a la propia marca o elementos relacionados con el mundo de la moda como las prendas de vestir.

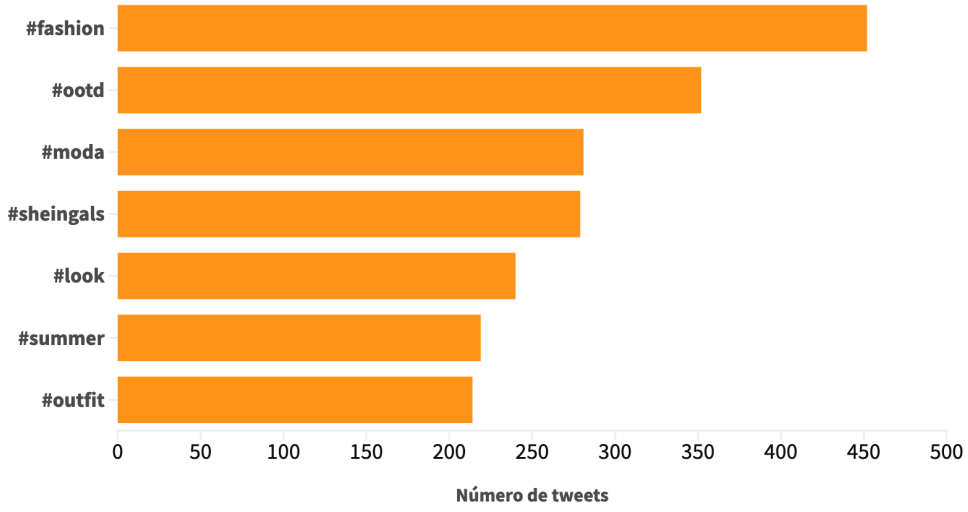


Figura 4.4: Hasgtags más utilizados
Elaboración propia

4.2. Limpieza y Análisis Descriptivo de los datos

Después de realizar un análisis preliminar de los datos para identificar su estructura y distribución, se ha llevado a cabo un proceso de limpieza de los mismos con el fin de simplificar el texto contenido en los tweets y preparar los datos para las tareas de modelado de tópicos y análisis de sentimiento. En primer lugar, se ha eliminado las *stopwords* o palabras vacías que no aportan información relevante al contenido. Entre ellas se encontraban preposiciones y conjunciones, que son las palabras más frecuentes en el texto. Además, se ha predefinido la eliminación de la palabra "Shein" del texto, ya que se utilizó como palabra clave para la descarga de los tweets y por tanto, todos los mensajes la contienen. Al convertirse en una

stopword, no aporta información relevante para el modelado de tópicos ni para el análisis de sentimientos. Posteriormente, se ha realizado una normalización de los datos, que incluye la eliminación de signos de puntuación, caracteres especiales y símbolos de emoticonos, con el objetivo de obtener un conjunto de datos homogéneo y más fácil de procesar. También se ha llevado a cabo un proceso de lematización para reducir las palabras a su raíz o lema, lo que nos permitirá agrupar términos similares y reducir la dimensión del vocabulario utilizado en el análisis. En esta etapa, se ha realizado además una revisión manual del proceso para corregir los casos en los que la librería no era capaz de identificar y lematizar correctamente los verbos. Esto resulta especialmente relevante en el análisis de textos en español, donde la librería empleada muestra algunas limitaciones. De esta manera, se ha logrado una mayor precisión en la agrupación de términos y en la reducción de la complejidad del vocabulario utilizado en el análisis.

Una vez realizado el análisis descriptivo y la limpieza de los datos, se ha procedido a realizar un análisis exploratorio de las palabras más frecuentes en el corpus analizado. Para ello, se ha generado una nube de palabras, la cual se muestra en la Figura 4.5. En esta nube de palabras, se pueden observar las palabras con mayor frecuencia en el corpus, donde el tamaño de la palabra refleja su frecuencia absoluta. Entre las palabras con mayor frecuencia en los tweets analizados, destacan aquellas relacionadas con la experiencia de compra, como "pedido", "paquete","llegar", "ropa" o "comprar". Además, se pueden apreciar unas pocas palabras relacionadas con opiniones y valoraciones sobre la marca y sus productos, como "precioso", "encantar", o "feliz". Finalmente, se observan palabras relacionadas con los códigos de descuento, los cuales son muy utilizados por la marca para incentivar la compra, como "descuento", o "código". En general, de forma preliminar, esta nube de palabras refleja algunos de los principales intereses de los usuarios de Shein en Twitter, lo que puede ser muy útil para entender la percepción de marca en redes sociales y mejorar la relación con los consumidores.

Después de realizar la limpieza de los datos, se ha utilizado la medida estadística TF-IDF para transformar el texto en vectores numéricos, permitiendo cuantificar la relevancia de las palabras en el texto no solo basándose en su frecuencia absoluta, sino también en su importancia en el corpus completo. Estas transformaciones han generado N-gramas, que incluyen tanto palabras individuales como combinaciones de dos o más palabras. En primer lugar, se han obtenido los unigramas de mayor relevancia en el corpus analizado. Así, la Figura 4.6 muestra las palabras que han obtenido una mayor puntuación TF-IDF. Se observa que entre las palabras más relevantes se encuentran términos como "pedido", "llegar" y "comprar", que están directamente relacionados con la compra y recepción de los pedidos de la marca. Además, destacan adjetivos con opiniones positivas sobre las prendas recibidas, como "precioso" y "encantar". Este análisis de los unigramas es consistente con la nube de palabras descrita anteriormente y proporciona una idea general de los temas que se discuten con mayor frecuencia en los tweets relacionados con la marca.

Figura 4.5: Palabras más frecuentes en los tweets
Elaboración propia

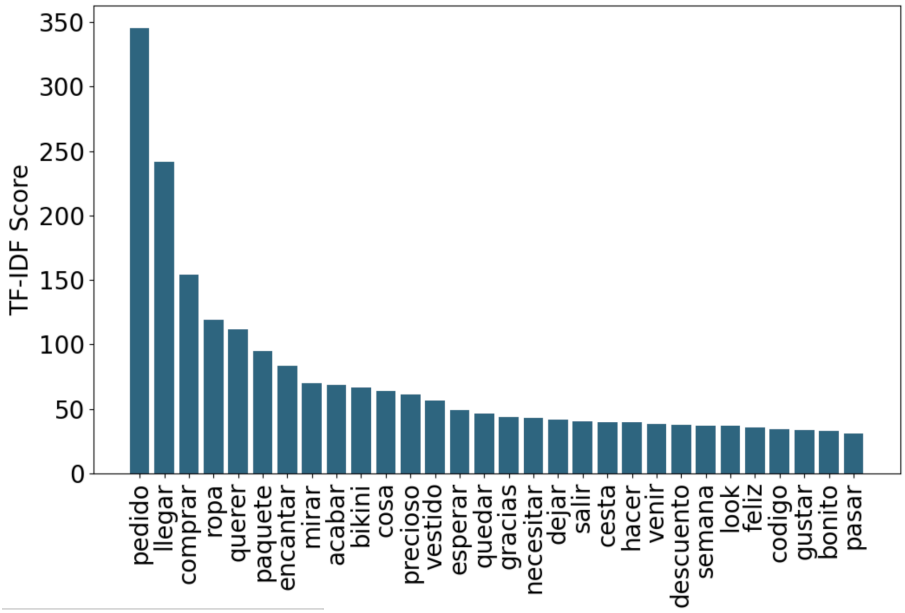


Figura 4.6: Unigramas con mayor TF-IDF
Elaboración propia

Después de obtener los unigramas más relevantes del texto, también se han analizado los bigramas y trigramas para observar las secuencias de 2 y 3 palabras que aparecen juntas con mayor importancia en el conjunto de tweets analizado. En la Figura 4.7 y la Figura 4.8 se pueden apreciar los bigramas y trigramas más relevantes, respectivamente. En primer lugar, se observa que el conjunto de palabras con mayor relevancia está relacionado con la recepción de los pedidos de la marca, como “querer llegar pedido” y “acabar llegar pedido”. En segundo lugar, de forma consistente con los resultados descritos anteriormente, se destacan

las secuencias de palabras relacionadas con el uso de códigos de descuento para la compra de productos de la marca, como “descuento usar código” y “usar código dejar”. Estas secuencias de palabras sugieren que los usuarios suelen compartir sus experiencias de compra, incluyendo la recepción de sus pedidos y la utilización de códigos de descuento, lo que puede ser útil para la marca en el desarrollo de estrategias de marketing.

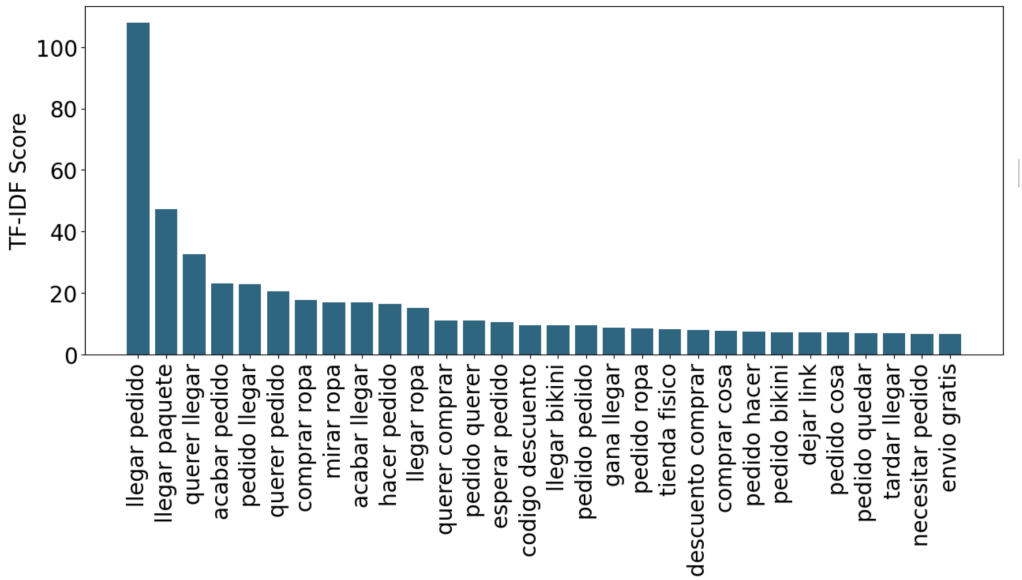


Figura 4.7: Bigramas con mayor TF-IDF
Elaboración propia

A partir de estos N-gramas obtenidos, es posible tener una idea preliminar de los posibles tópicos o categorías presentes en el corpus. Por ejemplo, las palabras relacionadas con comprar o pedir sugieren una categoría de discusión relacionada con la “Recepción de pedidos”. Asimismo, todas las palabras relacionadas con códigos de descuento sugieren que es un tema de discusión relevante para los usuarios de Shein. A pesar de esto, es importante tener en cuenta que la identificación precisa de categorías de discusión dependerá de la etapa de modelado discutida en la siguiente sección.

4.3. Resultados del modelado de tópicos

En la fase de modelado de tópicos, se ha llevado a cabo una tarea crucial para poder extraer información relevante del corpus de datos. La intención de esta tarea es encontrar patrones significativos y agruparlos en categorías representativas o tópicos. En este sentido, se ha utilizado un modelo muy conocido en la literatura denominado LDA o *Latent Dirichlet Allocation*. Este modelo se basa en la premisa de que cada documento puede estar compuesto por diferentes tópicos, y cada tópico se caracteriza por una distribución de palabras. El objetivo principal del modelo LDA es descubrir los tópicos latentes dentro del corpus y la

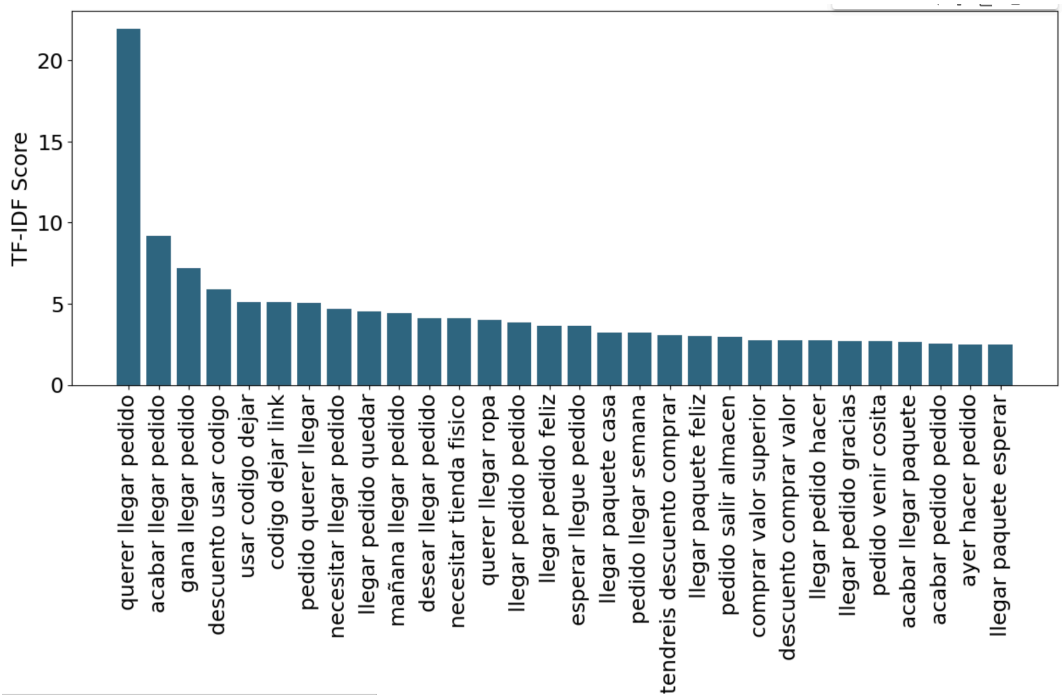


Figura 4.8: Trigramas con mayor TF-IDF
Elaboración propia

probabilidad de que un documento esté compuesto por esos tópicos. A través de este proceso, se puede obtener información relevante y útil de los datos, lo que facilita la comprensión de los mensajes emitidos por los usuarios y las opiniones expresadas sobre la marca.

Con el fin de llevar a cabo la implementación del modelo, es necesario determinar el número óptimo de tópicos K. Para ello, se analizará el índice de coherencia y la distancia intertópico del modelo para diferentes valores de K. El valor óptimo de K será aquel que permita obtener un valor alto de índice de coherencia junto con una distribución de tópicos interpretable (representada por la distancia en el mapa inter-tópicos). Así, en primer lugar, con el fin de analizar el índice de coherencia, se obtiene el K óptimo por el método del codo, el cual busca el valor que satisfaga que un incremento de K, no mejora sustancialmente el índice de coherencia. De esta manera, como puede verse en la Figura 4.9, el valor óptimo es $K = 5$, pues tiene una pendiente de crecimiento alta y un aumento del valor de K no mejora sustancialmente el índice de coherencia.

En cuanto al análisis de tópicos, también se ha llevado a cabo la evaluación de la distancia intertópica para determinar si los tópicos obtenidos son altamente interpretables. El objetivo de esta evaluación es encontrar un modelo en el que los tópicos, representados por círculos en la Figura 4.11, estén suficientemente separados entre sí y distribuidos en el mapa de manera que la información proporcionada por cada tópico sea fácilmente diferenciable. Tras la selección de un valor de $K = 5$, se ha observado que la distancia intertópica del modelo es buena, ya que los círculos no se solapan y están claramente separados unos de

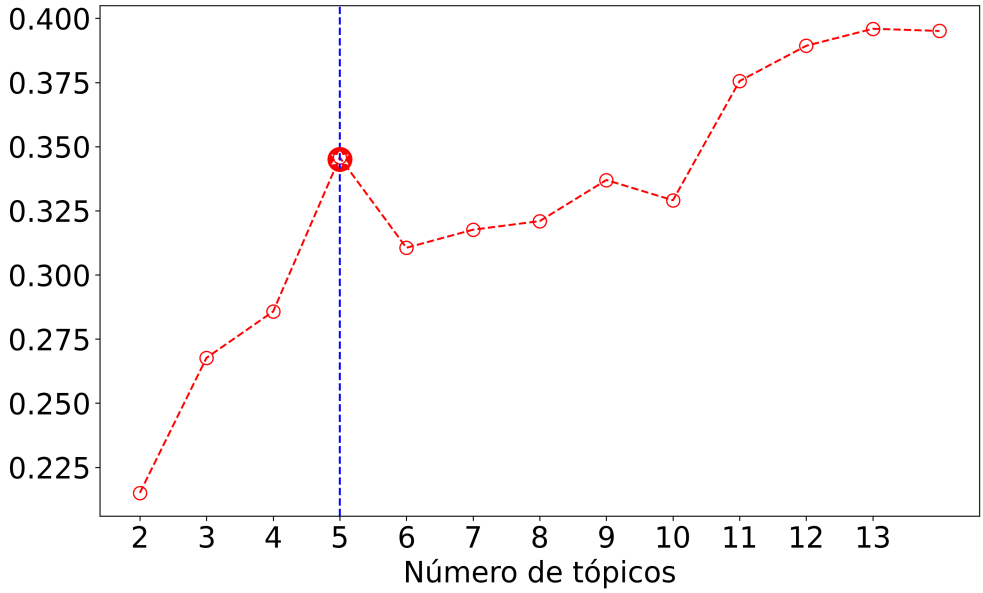


Figura 4.9: Modelo óptimo según el índice de coherencia
Elaboración propia

otros. Este resultado indica que la información proporcionada por cada uno de los tópicos es significativamente diferente, lo que a su vez implica que cada tópico contendrá información relativamente interpretable.

Como resultado, se presenta la tabla 4.1 con los 5 tópicos obtenidos mediante el modelo LDA, junto con su respectiva distribución de palabras clave y bigramas asociados. Se ha asignado a cada tópico un nombre de categoría representativo, acorde con los bigramas más relevantes y algunas de sus palabras claves. Aunque algunas palabras como “comprar” o “pedido” se repiten en varios tópicos debido a su alta frecuencia de aparición en diferente temáticas, en general, los tópicos tienen conceptos claramente diferenciables entre si.

De esta manera, en la primera categoría modelada se identifica un tópico asociado a los **Descuentos**. Las palabras clave que definen este tópico están asociadas a códigos y descuentos. Es importante destacar que la marca en cuestión es muy conocida por ofrecer atractivos descuentos y promociones, con el objetivo de incentivar el consumo de sus productos. Además, esta estrategia de marketing se ve reforzada gracias a su colaboración activa con diversas celebridades e influencers en las redes sociales, a quienes les proporcionan códigos personalizados para que los compartan con su comunidad. Por su parte, el segundo tópico que se ha identificado, denominado **Consumo**, está estrechamente relacionado con el primero, ya que la estrategia de ofrecer descuentos y promociones de forma constante lleva a los clientes a realizar compras de forma compulsiva y repetitiva. De esta manera, la marca logra fomentar la adicción a la compra en su sitio web y a su catálogo de productos. Además, se puede observar que los usuarios se sienten atraídos por la gran cantidad de opciones que ofrece la

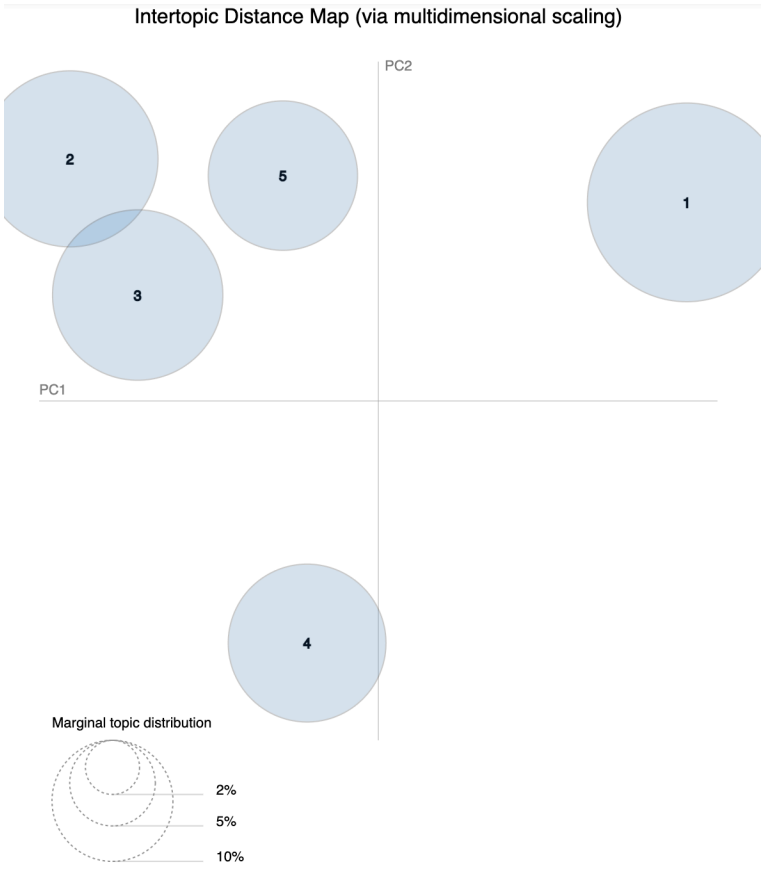


Figura 4.10: Distancia intertópica, modelo K=5
Elaboración propia

tienda en línea y la variedad de productos que están disponibles, lo que les lleva a consumir en grandes cantidades. En resumen, el tópico **Consumo** representa una tendencia a adquirir una gran cantidad de productos por parte de los usuarios de la marca, en gran parte motivada por la estrategia de descuentos que implementa la empresa.

El tercer tópico se centra en las prendas en sí y en las opiniones que los usuarios tienen sobre ellas, por ello ha sido catalogado como **Looks y prendas**. Aquí se pueden encontrar diferentes aspectos relacionados con las prendas como el tipo de prenda o su diseño. También se incluyen comentarios recibidos hacia las propias publicaciones de la marca sobre sus prendas. El cuarto tópico del análisis de la marca presenta dos subtemas interesantes que vale la pena explorar con mayor detalle. El primer subtema se relaciona con el canal de compra y la apertura de una tienda física. Se puede apreciar claramente una aparente demanda entre los usuarios por una experiencia de compra en vivo, lo que puede estar generando cierta frustración entre los consumidores de la marca que solo pueden comprar en línea. Es por eso que muchos usuarios comentan sobre la necesidad de una tienda física, lo que sugiere una oportunidad de negocio para la marca. El segundo subtema se centra en la compra de prendas de verano, concretamente de bikinis. Es notable la popularidad de esta prenda dentro del catálogo de productos de la marca, ya que dispone de una amplia variedad de modelos con

Tópico	Categoría	Palabra clave	Bigramas más relevantes
1	Descuentos	descuento, código, necesitar, encantar comprar	codigo descuento, descuento comprar, search codigo, descuento articulo, superior codigo cupon descuento
2	Consumo	comprar, cesta, ropa pedido, hacer, esperar	llenar cesta, cesta lleno, adicta comprar, comprar vicio, obsesionado comprar
3	Looks y prendas	precioso, vestido, look ropa, bolso, comprar	precioso look, ropa verano, precioso vestido, ropa hombre, precioso jersey
4	Canal de compra y bikinis	comprar, bikini, españa, comprar gustar, bonito	tienda fisica, necesitar tienda, madrid españa, mirar bikini, comprar bikini, bikini quedar
5	Recepción de pedidos	pedido, llegar, querer esperar, paquete,venir	llegar pedido, querer llegar, querer pedido, necesitar llegar, tardar llegar

Tabla 4.1: Palabras clave y bigramas por categoría

precios muy competitivos en comparación con otras marcas habituales. Por ello, los comentarios asociados a esta prenda aparecen con gran frecuencia en el análisis, y es uno de los principales productos por los que destaca entre los usuarios. Además, la marca ha logrado mantener una imagen fresca y juvenil asociada a la moda de verano y la playa, lo que ha contribuido a la popularidad de los bikinis y otros productos de verano. Por esta razón este cuarto tópico ha sido categorizado como **Canal de Compra y bikinis**.

El último tópico en el análisis de los comentarios de los usuarios está centrado en la **recepción de pedidos** y su logística. Los usuarios suelen compartir sus experiencias en cuanto al tiempo de entrega y la recepción de sus paquetes, expresando su satisfacción o insatisfacción en relación a la rapidez y eficacia de la entrega. Finalmente, se ha analizado la distribución de los tópicos dentro del corpus, como se muestra en la Figura 4.11. De esta forma, se puede observar que los tópicos 5 (Recepción de pedidos) y 3 (Looks y prendas) son los que tienen mayor cantidad de tweets categorizados siendo el tópico 2 (Consumo) el que tiene una menor frecuencia de tweets. Este resultado es coherente con el análisis preliminar, donde en la Figura 4.5, ya se observaba a priori que las palabras con mayor frecuencia tenían que ver con estas categorías, destacándose palabras como “pedido”, “llegar”, “paquete” o “ropa” entre otras.

4.4. Análisis de sentimiento

En esta última fase del análisis, se busca identificar las emociones contenidas en el texto con el objetivo de entender el sentimiento de los usuarios hacia la marca. Por tanto, la metodología empleada para analizar las emociones contenidas en el texto es VADER, descrito en la sección 3.4, un diccionario léxico diseñado para analizar sentimientos en redes sociales, que utiliza una combinación de reglas gramaticales y listas de palabras con valores asociados a los sentimientos positivos y negativos. Al utilizar este diccionario, se puede obtener

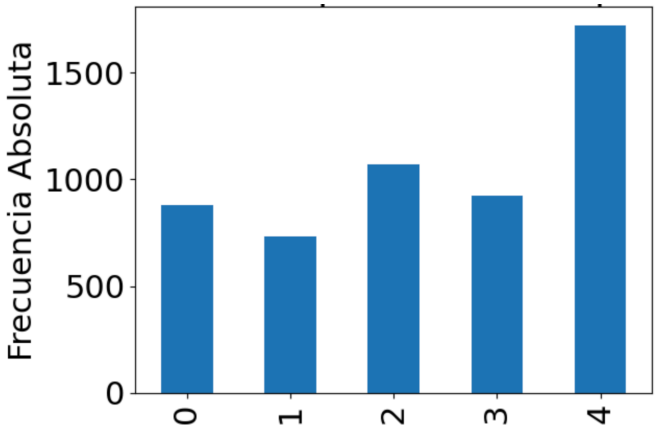


Figura 4.11: Distribución de tópicos en el corpus
Elaboración propia

información valiosa sobre el tono emocional de los comentarios, lo que permite comprender cómo se sienten los usuarios en relación a la marca y sus productos. Es importante mencionar que, el preprocesamiento inicial ha sido modificado con el fin de incluir mayúsculas o signos de puntuación, ya que estos elementos pueden ayudar a comprender mejor las emociones subyacentes en el texto.

De esta manera, se ha realizado un análisis de sentimientos de los comentarios del dataset, calculando una puntuación compuesta para cada uno. La Figura 4.12 muestra la evolución de los sentimientos expresados en los tweets a lo largo del tiempo. Aunque se observa una tendencia neutra donde las emociones negativas y positivas se equilibran, se destacan picos mínimos y máximos para fechas puntuales. Además, es importante señalar que a partir del año 2020 se observa un cambio significativo en el volumen de interacciones, coincidiendo con el inicio de la pandemia y el crecimiento de la marca en el país a través de su estrategia de expansión y la evolución en las compras en línea. Este mayor número de observaciones en esta etapa considera un aumento tanto en los comentarios positivos como negativos en relación con la marca. Es decir, se evidencia que la marca ha tenido una mayor visibilidad y ha atraído una cantidad importante de opiniones de los usuarios en el último período analizado.

Cabe destacar que se registró un pico especialmente negativo a mediados del año 2022 en los comentarios sobre la marca. Este fenómeno se puede atribuir a las investigaciones que se realizaron acerca de las condiciones laborales de los trabajadores de la marca, que se ha vuelto muy popular en el sector de la moda *fast-fashion*, con el objetivo de eclipsar a competidores como Zara. Algunos usuarios compartieron mensajes de ayuda encontrados en las etiquetas de las prendas de la marca, lo que llevó a una mayor visibilidad del tema. En línea con la gráfica que muestra la evolución de los sentimientos, en este año también se emitió un documental británico que mostraba la precariedad de las condiciones de los trabajadores, lo que dejó a la marca en una posición comprometida. Esto generó un amplio descontento y críticas hacia la marca, lo que pudo haber contribuido al pico negativo en los comentarios.

Es importante señalar que, a pesar de este incidente, la marca ha mantenido una presencia fuerte en el mercado, ha continuado expandiéndose en línea con su estrategia de crecimiento y no evidencia un sentimiento particularmente negativo a partir de esa fecha.

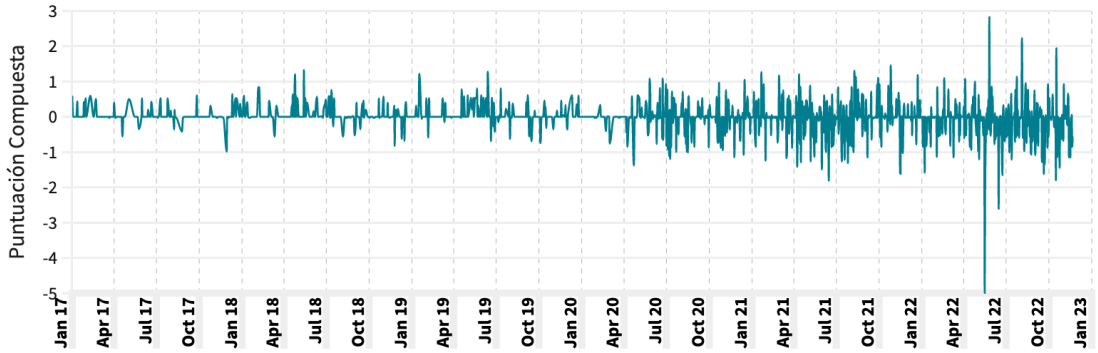


Figura 4.12: Evolución del sentimiento en el tiempo
Elaboración propia

Para complementar el análisis del sentimiento en las categorías de discusión desarrolladas, se ha realizado un análisis de sentimiento por cada uno de los tópicos encontrados. Este análisis proporciona una perspectiva más detallada de cómo los usuarios perciben cada uno de los temas. La Figura 4.13 muestra los resultados del análisis de sentimiento en cada uno de los tópicos. El análisis de sentimiento por cada uno de los tópicos con permite obtener una comprensión más profunda sobre las emociones de los usuarios en relación con las diferentes categorías halladas en el modelado de tópicos. Los resultados muestran un comportamiento neutral en el sentimiento subyacente a las publicaciones en Twitter para cada uno de los tópicos. Sin embargo, es interesante notar que el tópico de consumo y descuentos tenía una componente mayoritariamente positiva antes de la pandemia, y durante y después de la misma se observó un aumento en los comentarios tanto positivos como negativos. Por otro lado, la categoría de consumo demostró ser la que tiene menores valores de sentimiento, siendo más neutral para los usuarios. Además, se observó un componente positivo de gran importancia en las categorías de canal de compra y recepción de pedidos.

En última instancia, es evidente el aumento significativo de interacciones a partir del año 2020, tal y como se ha mencionado anteriormente. Este incremento es notorio en comparación con los períodos anteriores. Este fenómeno se debe a que, antes de la pandemia, la marca no era muy conocida, pero a partir de ese momento, con el auge de la venta en línea y la popularidad de plataformas como TikTok -en la que Shein invierte una gran cantidad de dinero en publicidad- la marca ha experimentado un aumento exponencial en popularidad y ventas. Si se comparan los tópicos, se puede observar que especialmente en el tópico 5, relacionado con la recepción de pedidos, existe una mayor cantidad de interacciones. Este hecho es coherente, ya que la mayoría de los comentarios en nuestro conjunto de datos se relacionan con los usuarios que reclaman a la marca por la llegada de sus pedidos y la ansiada espera

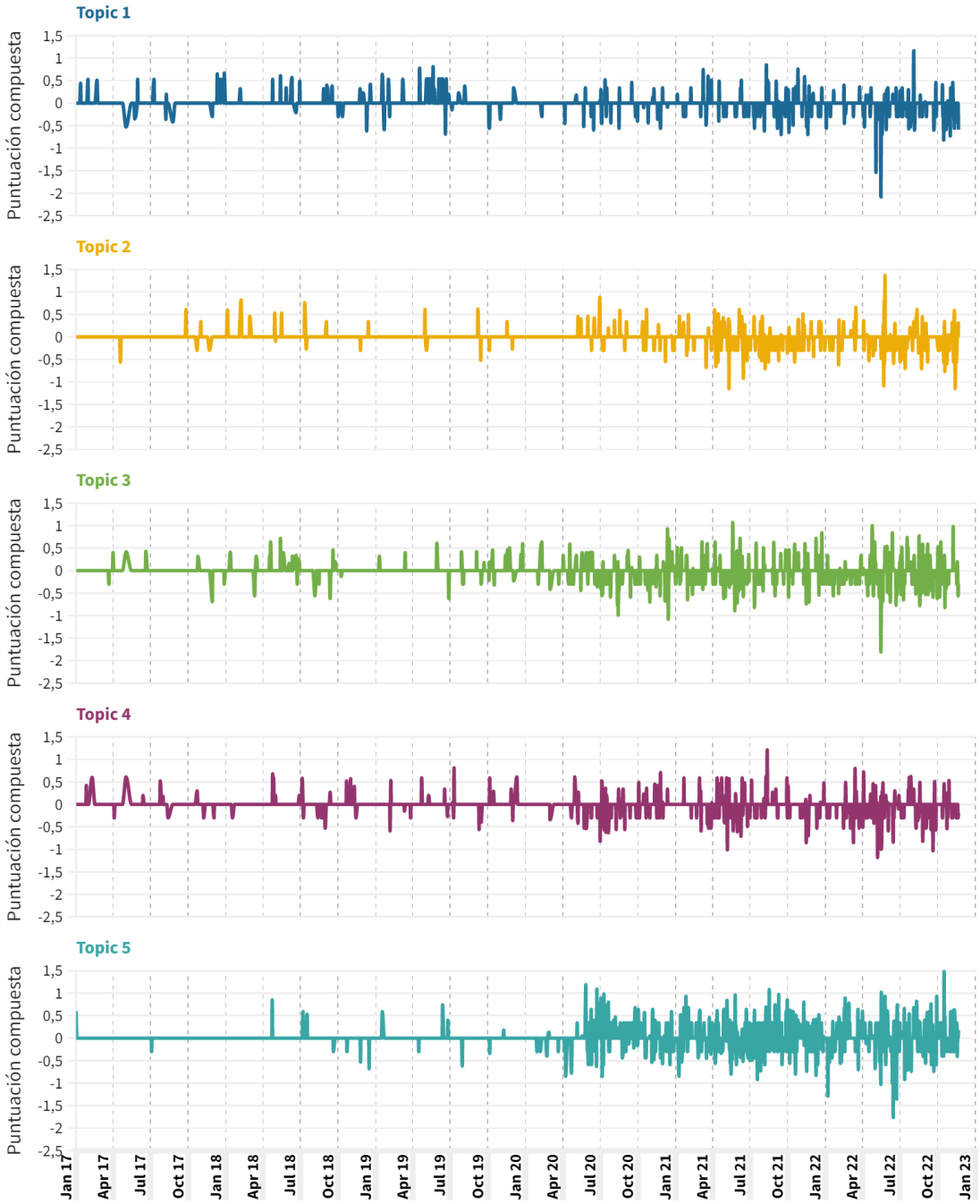


Figura 4.13: Evolución del sentimiento en el tiempo por tópico
Elaboración propia

por recibirlos, como se aprecia en la distribución de palabras por tópico representada en la Figura 4.11. Sin embargo, salvo los elementos destacados, debido la neutralidad encontrada en el análisis de sentimiento de cada tópico, no se pueden extraer conclusiones muy significativas acerca de si los usuarios muestran polaridad de tipo positivo o negativo en cada una de las categorías.

Capítulo 5

Conclusiones

En este trabajo se ha realizado el análisis de la percepción de la marca Shein en España a través de redes sociales es importante debido al creciente papel que estas plataformas juegan en la promoción y difusión de marcas y productos. Shein es una marca que ha experimentado un crecimiento significativo en España en los últimos años, especialmente durante la pandemia, lo que hace aún más interesante el análisis de la percepción de los consumidores hacia la marca debido a su reciente popoularidad. Además, en los últimos años, han surgido algunas críticas hacia Shein por su modelo de negocio, acusándola de fomentar la cultura de la moda rápida y el consumismo excesivo, lo que ha llevado a cuestionar su impacto social y ambiental. Teniendo en cuenta estas consideraciones, el objetivo del presente trabajo de grado está enfocado en el análisis de la percepción de los consumidores españoles hacia Shein, lo cual permite obtener una visión más clara de cómo la marca es percibida en este mercado y qué aspectos son valorados positiva o negativamente. Para ello, se han realizado un análisis de *social media*, en concreto, usando como fuente de datos Twitter donde se han obtenido aquellos comentarios asociados a la marca desde el comienzo de su periodo de expansión en 2017, hasta la actualidad.

En este contexto, se ha realizado una revisión exhaustiva de la literatura con el fin de analizar las metodologías más adecuadas para llevar a cabo este análisis. Como resultado, se ha determinado que la técnica LDA es la más utilizada para el modelado de tópicos en redes sociales, debido a su flexibilidad y capacidad para encontrar patrones y estructuras en los datos sin necesidad de datos de entrenamiento previos. Además, se ha identificado que las estrategias basadas en diccionarios, especialmente VADER, son la opción más utilizada para el análisis de sentimientos en datos extraídos de redes sociales, donde se requiere la interpretación de elementos no lingüísticos como signos de puntuación o emoticonos.

Teniendo en cuenta estos resultados, la metodología de análisis implementada en este trabajo ha constado de varias fases. Por un lado, la adquisición de los datos a través de un API de Twitter y su posterior pre-procesamiento, donde se han trasformado y preparado los datos de cara al modelado de tópicos. Tras el procesamiento, se ha implementado un análisis

descriptivo de los datos, donde se han obtenidos los N-Gramas calculando la frecuencia de los términos en el documento mediante la técnica TF-IDF, pudiendo así extraer aquellos unigramas, bigramas y trigramas de mayor relevancia. Esta parte del análisis permitió observar que las palabras más frecuentes en el texto se relacionan principalmente con la recepción y envío de pedidos de la marca. Este resultado no es sorprendente, ya que Shein es una marca exclusivamente online que opera desde una localización alejada (China) y tiene un modelo de negocio de fabricación bajo demanda, lo que conlleva a tiempos de envío extensos y a menudo incumplimientos con retrasos debido a problemas en las aduanas y otros factores. La alta frecuencia de estas palabras en los comentarios de los usuarios refleja la importancia que estos aspectos tienen para la percepción general de la marca.

En la parte del modelado de tópicos, se obtuvo el número óptimo K de tópicos mediante su respectivo índice de coherencia donde se obtuvo un total de 5 tópicos que representan las categorías más relevantes en el documento de forma interpretable. Estos tópicos fueron asignados como Descuentos, Consumo, Looks y prendas, Canal de compra y bikinis y Recepción de pedidos. La identificación de cada categoría se logró gracias a la agrupación coherente de palabras clave y bigramas por tópico. Es importante destacar que el tópico 5, relacionado con la Recepción de pedidos, presentó la mayor frecuencia en el corpus, lo que coincide con la tendencia de comentarios asociados a la marca. Por otro lado, el tópico 4 resultó ser interesante, ya que englobaba dos subtemas relevantes: la demanda de tiendas físicas y las prendas veraniegas, específicamente la palabra “bikini”.

Finalmente, se llevó a cabo un análisis de sentimiento para averiguar cómo se sienten los consumidores con respecto a los diferentes tópicos categorizados, donde se obtuvo un sentimiento de tipo neutral principalmente. Esto, implica que no existe un sentimiento particularmente positivo o negativo en las publicaciones en Twitter relacionadas con cada uno de los tópicos analizados. No obstante, se observó un cambio interesante en el tópico de consumo y descuentos, que antes de la pandemia tenía una inclinación mayoritariamente positiva pero durante y después de la misma, se registró un aumento en los comentarios tanto positivos como negativos. Por otra parte, la categoría de consumo mostró valores de sentimiento más bajos, con un componente neutro mayor por parte de los usuarios. También se encontró un componente positivo significativo en las categorías de canal de compra y recepción de pedidos.En este último tópico es donde se aprecia mayor volumen de interacciones y, como dato interesante, se observó que la intensidad del sentimiento se ve fuertemente incrementada a partir del año 2020, periodo de la pandemia, el cual coincide con el auge de la marca.

En conclusión, este estudio ha permitido analizar el comportamiento de los usuarios en redes sociales respecto a una marca de moda online. A través del análisis descriptivo de n-gramas, modelado de tópicos y análisis de sentimiento, se ha obtenido información valiosa sobre los intereses y percepciones de los usuarios. En general, la metodología de análisis implementada podría ser de gran utilidad para la marca a la hora de diseñar estrategias de marketing y mejorar la experiencia de usuario en su plataforma online. En cuanto a trabajos

futuros, se podría considerar la obtención de un mayor volumen de datos, ya que esto permitiría una mayor profundidad en el análisis de los tópicos y la obtención de sentimientos con una polaridad más clara. Una forma de lograr esto sería a través de la implementación de técnicas de georeferenciación de los tweets, lo que permitiría detectar un mayor número de tweets publicados en España y, por lo tanto, tener una visión más detallada de la opinión de los usuarios españoles sobre la marca. Asimismo, se podría considerar la inclusión de otras redes sociales en el análisis, lo que permitiría tener una visión más completa de la percepción de la marca en el panorama digital. Además, sería interesante explorar la posibilidad de realizar un análisis de sentimiento más profundo, incluyendo el uso de modelos de aprendizaje profundo para una mayor precisión en la detección de sentimientos. En resumen, estas líneas de trabajo permitirían profundizar en los hallazgos obtenidos en este trabajo y aumentar la cantidad de información relevante extraida de la percepción de la marca en línea.

Referencias

Adeborna, E., y Siau, K. (2014). An approach to sentiment analysis-the case of airline quality rating. En *18th pacific asia conference on information systems, pacis 2014*.

Adevinta. (2021). *Informe sobre la evolución y las tendencias en los hábitos de consumo. 2021-2022.* (acceso Septiembre 6, 2022) `https://www.adevinta.com/app/uploads/sites/2/2022/03/Pulso-Digital-Adevinta-2021.pdf`.

Al-Hajjar, D., y Syed, A. Z. (2015). Applying sentiment and emotion analysis on brand tweets for digital marketing. En *2015 ieee jordan conference on applied electrical engineering and computing technologies (aeect)* (p. 1-6). doi: 10.1109/AEECT.2015.7360592

Apparel Market Statista. (2022). *Revenue of the apparel market worldwide from 2013 to 2026.* (acceso Septiembre 26, 2022) `https://www.statista.com/forecasts/821415/value-of-the-global-apparel-market`.

Apptopia. (2022). *Worldwide and us download leaders 2021.* (acceso Septiembre 30, 2022) `https://blog.apptopia.com/worldwide-and-us-download-leaders-2021`.

Babčanová, D., Šujanová, J., Cagáňová, D., Horňáková, N., y Hrablik Chovanová, H. (2021). Qualitative and quantitative analysis of social network data intended for brand management. *Wireless Networks*, *27*(3), 1693–1700.

Barret, P., y Baumann, D. (2019). *Made in ethiopia: Challenges in the garment industry's new frontier. new york university's stern center for business and human rights.* (acceso Septiembre 6, 2022) `https://issuu.com/nyusterncenterforbusinessandhumanri/docs/nyu_ethiopia_final_online?e=31640827/69644612`.

Barry, A. E., Valdez, D., Padon, A. A., y Russell, A. M. (2018). Alcohol advertising on twitter—a topic model. *American Journal of Health Education*, *49*(4), 256–263.

Bhardwaj, V., y Fairhurst, A. (2010). Fast fashion: response to changes in the fashion industry. *The international review of retail, distribution and consumer research*, *20*(1), 165–173.

Business of Apps - Shein. (2022). *Shein revenue and usage statistics (2022).* (acceso Septiembre 6, 2022) `https://www.businessofapps.com/data/shein-statistics/`.

Business of Apps - Twitter. (2022). *Social app report 2022.* (acceso Octubre 19, 2022)

https://www.businessofapps.com/data/twitter-statistics/.

Carson Sievert, K. E. S. (2014). Ldavis: A method for visualizing and interpreting topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 63–70.

Castro, A. (2022). *Fast fashion hoy, el mundo colgando de un hilo mañana.* (acceso Septiembre 11, 2022) http://eduneuro.com/revista/index.php/revistaneuronum/article/viewFile/395/477.

Clothing Production Statista. (2018). *Where pay is lowest for cheap clothing production.* (acceso Septiembre 26, 2022) https://www.statista.com/chart/17903/monthly-minimum-wage-in-the-global-garment-industry/.

Cosentino, A. (2019). Risk and reward: An analysis of# boycottnike as a response to nike's colin kaepernick advertising campaign. *Elon Journal of Undergraduate Research in Communications*, *10*(1), 54–63.

Daxue Consulting. (2022). *Shein's market strategy: how the chinese fashion brand is conquering the west.* (acceso Septiembre 26, 2022) https://daxueconsulting.com/shein-market-strategy/.

del Arrabal Fernández Matilla, M. (2017). *Moda sostenible.anÁlisis de su naturaleza y perspectiva futura.* (acceso Septiembre 11, 2022) https://buleria.unileon.es/bitstream/handle/10612/7207/Fern%C3%A1ndez%20Matilla%2C%20Mar%C3%ADa%20Del%20Arrabal.pdf?sequence=1.

Dottle, R., y Gu, J. (2022). *The global glut of clothing is an environmental crisis.* (acceso Septiembre 11, 2022) https://www.bloomberg.com/graphics/2022-fashion-industry-environmental-impact/.

Eco Club. (2022). *5 times shein has copied designs from independent fashion brands.* (acceso 26 Septiembre, 2022) https://ecoclubofficial.com/shein-stealing-designs-independent-fashion-brands/.

Ecommerce Countries Statista. (2022). *Fashion retail e-commerce revenue worldwide in 2022, by country.* (acceso Septiembre 26, 2022) https://www.statista.com/forecasts/1305317/e-commerce-fashion-revenue-by-country-worldwide.

Ecommerce Worldwide Statista. (2022). *Fashion retail e-commerce revenue worldwide from 2019 to 2025, by region.* (acceso Septiembre 26, 2022) https://www.statista.com/forecasts/1305325/e-commerce-fashion-revenue-by-country-worldwide.

Feldmeyer, A., y Johnson, A. (2022). Using twitter to model consumer perception and product development opportunities: A use case with turmeric. *Food Quality and Preference*, *98*, 104499.

Giri, C., Harale, N., Thomassey, S., y Zeng, X. (2018). Analysis of consumer emotions about fashion brands: An exploratory study. En *Data science and knowledge engineering for sensing decision support: Proceedings of the 13th international flins conference (flins*

2018) (pp. 1567–1574).

Godoy, M. (2021). *Shein contra las cuerdas: el fabricante de las icónicas botas dr. martens y otros minoristas acusan a la plataforma china de infracción de marca registrada.* (acceso 26 Septiembre, 2022) https://www.businessinsider.es/fabricante-botas-dr-martens-acusa-shein-plagio-882257.

Guptal, D. (2022). *Business model of shein - how does shein make money?* (acceso 11 Septiembre, 2022).

Hutto, C., y Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. En *Proceedings of the international aaai conference on web and social media* (Vol. 8, pp. 216–225).

Kant, R. (2011). Textile dyeing industry an environmental hazard. *Scientific research publishing*.

Leading Fashion Statista. (2022). *Leading fashion and clothing retailers in europe in 2020, based on turnover in europe.* (acceso Septiembre 26, 2022) https://www.statista.com/forecasts/711107/turnover-of-clothing-retailers-in-european-union-eu#:~:text=This%20statistic%20presents%20the%20turnover,turnover%2C%20at%2012.9%20billion%20euros.

Linoff, G. S., y Berry, M. J. (2011). *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons.

Liu, X., y Burns, A. (2018). Understanding consumer-generated content about luxury brands using big data: An abstract. En *Academy of marketing science annual conference* (pp. 503–504).

Mazaira, A., Gonzalez, E., y Avendaño, R. (2003). The role of market orientation on company performance through the development of sustainable competitive advantage: the inditex-zara case. *Marketing Intelligence & Planning*.

McKinsey & Company. (2019). *The state of fashion 2019.* (acceso Julio 3, 2022) https://www.mckinsey.com/~/media/mckinsey/industries/retail/our%20insights/renewed%20optimism%20for%20the%20fashion%20industry/the-state-of-fashion-2018-final.pdf.

McKinsey & Company. (2021). *The state of fashion 2021.* (acceso Julio 3, 2022) https://www.mckinsey.com/~/media/McKinsey/Industries/Retail/Our%20Insights/State%20of%20fashion/2021/The-State-of-Fashion-2021-vF.pdf.

MHE Consumer. (2021). *Shein. turbo low cost mobile fashion.* (acceso Septiembre 26, 2022) https://www.mheconsumer.com/en/2021/06/shein-turbo-low-cost-fashion/.

Miranda, J. A., y Roldán, A. (2021). Inditex y la ventaja competitiva de la fast fashion española, 1985-2019. *Dimensioni e problemi della ricerca storica*(2), 155–178.

MMX, C. (2022). *Tipo de visitantes en genero y edad en shein, zara y el corte ingles.* (acceso

30 Septiembre, 2022).

Modaes.es. (2022). *Informe económico de la moda en españa, 2018.* (acceso Julio 1, 2022) https://www.modaes.com/files/000_2016/0001publicaciones/pdfs/informe_economico_2018.pdf.

Moi Global. (2022). *El modelo de negocio de shein.* (acceso Septiembre 11, 2022) https://moiglobal.es/el-modelo-de-negocio-de-shein/.

Moreno, M. (2023). *Text-mining-analysis-using-twitter.* Descargado de https://github.com/Mariamoreno17/Text-mining-analysis-using-Twitter

Mostafa, M. M. (2013). More than words: Social networks' text mining for consumer brand sentiments. *Expert systems with applications*, *40*(10), 4241–4251.

Niinimäki, K., Peters, G., Dahlbo, H., Perry, P., Rissanen, T., y Gwilt, A. (2020). The environmental price of fast fashion. *Nature Reviews Earth & Environment*, *1*(4), 189–200.

Onan, A., Korukoglu, S., y Bulut, H. (2016). Lda-based topic modelling in text sentiment classification: An empirical analysis. *Int. J. Comput. Linguistics Appl.*, *7*(1), 101–119.

Oxford Analytica. (2021). Ultra-fast fashion risks eroding sustainability gains. *Emerald Expert Briefings*(oxan-db).

Pano, T., y Kashef, R. (2020). A complete vader-based sentiment analysis of bitcoin (btc) tweets during the era of covid-19. *Big Data and Cognitive Computing*, *4*(4). Descargado de https://www.mdpi.com/2504-2289/4/4/33

Perry, P. (2017). Read this before you go sales shopping: the environmental costs of fast fashion. *The Conversation*.

Peters, G. M., Sandin, G., y Spak, B. (2019). Environmental prospects for mixed textile recycling in sweden. *ACS Sustainable Chemistry & Engineering*, *7*(13), 11682-11690.

Public Eye. (2021). *75-hour weeks for shein: Public eye looks behind the chinese online fashion giant's glitzy front.* (acceso Septiembre 6, 2022) https://www.publiceye.ch/en/media-corner/press-releases/detail/75-hour-weeks-for-shein-public-eye-looks-behind-the-chinese-online-fashion-giants-glitzy-front.

Quantis. (2018). *Measuring fashion 2018. environmental impact of the global apparel and footwear industries study.* (acceso Julio 3, 2022) https://quantis.com/wp-content/uploads/2018/03/measuringfashion_globalimpactstudy_full-report_quantis_cwf_2018a.pdf.

Remy, N., Speelman, E., y Swartz, S. (2016). *Style that's sustainable: A new fast-fashion formula* (Inf. Téc.). McKinsey Global Institute.

Research & Markets. (2021). *Global fast fashion market report 2021.* (acceso Septiembre 6, 2022) https://www.prnewswire.com/news-releases/global-fast-fashion-market-report-2021---market-is-expected-to-grow-at-a-cagr-of-5-3-from-2025-and-reach-211-909-7-million-in-2030--301414180.html.

Reuters. (2021). *Exclusive chinese retailer shein lacks disclosures, made false state-*

ments about factories. (acceso Septiembre 6, 2022) https://www.reuters.com/business/retail-consumer/exclusive-chinese-retailer-shein-lacks-disclosures-made-false-statements-about-2021-08-06/.

Röder, M., Both, A., y Hinneburg, A. (2015). Exploring the space of topic coherence measures. En *Proceedings of the eighth acm international conference on web search and data mining* (pp. 399–408).

Sepúlveda, B. C. (2016). Aplicación y evaluación lda para asignación de tópicos en datos de twitter. *es. En: pág*, *55*.

Shen, J. (2022). Analyzing on the going global marketing strategy–taking shein as an example. En *2022 international conference on creative industry and knowledge economy (cike 2022)* (pp. 225–229).

Smartme Analytics. (2021). *Top apps con más cuota de mercado en españa: retail, banca, moda y delivery.* (acceso Septiembre 30, 2022) https://marketing4ecommerce.net/top-apps-con-mas-cuota-de-mercado-en-espana-retail-banca-moda-y-delivery/.

Smith, P. (2022). *Global apparel market - statistics facts.* (acceso Julio 3, 2022) https://www.statista.com/topics/5091/apparel-market-worldwide/#dossierContents__outerWrapper.

Statista. (2021). *Top online stores in spain in 2021, by e-commerce net sales.* (acceso Septiembre 8, 2022) https://www.statista.com/forecasts/871159/spain-top-online-stores-spain-ecommercedb.

Tamayo, X. (2018). *La cara oculta de la industria de la moda, un análisis socio jurídico* (Tesis de Master no publicada). Universitat de Vàlencia.

The Guardian. (2021). *'worst of the worst': why is fast fashion retailer shein launching a reality show?* (acceso Septiembre 6, 2022) https://www.theguardian.com/fashion/2021/aug/29/fast-fashion-retailer-shein-design-reality-show.

UNFCC. (2018). *United nations climate change. un helps fashion industry shift to low carbon.* (acceso Septiembre 6, 2022) https://unfccc.int/news/un-helps-fashion-industry-shift-to-low-carbon.

Uysal, A. K., y Gunal, S. (2014). The impact of preprocessing on text classification. *Information processing & management*, *50*(1), 104–112.

Webster, T. (2010). *Twitter usage in america: 2010.* (acceso Julio 1, 2022) http://www.edisonresearch.com/twitter_usage_2010.php.

Yan, K.-Q., Wang, S.-C., Wang, S.-S., y Lin, Y.-P. (2011). Application of data mining for enterprise digital marketing strategy making. En *2011 third international conference on communications and mobile computing* (pp. 509–512).

Zambrano, G. N. A., Andrade, E. V. A., Cagua, L. A. A., Mera, S. P. C., Posligua, L. A. C., Zambrano, M. M. D., y Zambrano, L. (2018). Evolución del marketing tradicional al

marketing digital. *Comité científico revisores-correctores*, 64.