



Facultad de Ciencias Económicas y Empresariales

Percepción ciudadana sobre el servicio de venta minorista en línea: Estudio de análisis en Twitter

Autor: Lourdes Artacho Sierras

Director: Jenny Alexandra Cifuentes Quintero

MADRID | Abril 2023

Resumen

La presencia en redes sociales es crucial para el desarrollo y crecimiento de las empresas, ya que les brinda exposición al público y acceso a la percepción de la marca. La gestión adecuada de estas plataformas se ha vuelto fundamental debido a la evolución tecnológica y al aumento de su uso para interactuar con los clientes. Este cambio ha diversificado los canales de comunicación y ha permitido a los usuarios expresar opiniones públicas. Es así como las redes sociales se han convertido en un factor influyente para retener y atraer clientes.

Este trabajo de fin de grado tiene como objetivo analizar la percepción y los sentimientos de los usuarios hacia la reconocida marca minorista Mercadona. Para ello, se realizó un exhaustivo análisis del estado del arte para identificar las técnicas más adecuadas. En el modelado de tópicos, se determinó que la técnica LDA es ampliamente empleada, por lo que se seleccionó para este estudio. En cuanto al análisis de sentimientos, se encontró que las técnicas más utilizadas incorporan reglas de decisión basadas en diccionarios, y en este proyecto, en particular, se optó por implementar la estrategia VADER.

De esta forma, este trabajo se basa en el análisis de un conjunto de 115,594 tweets en los cuales se menciona la palabra “mercadona”. A través del modelado de tópicos, se identificaron tres temas principales en los tweets relacionados con Mercadona: la experiencia de compra y variedad de productos, factores empresariales y de liderazgo, e interacciones con clientes y productos destacados. En el primer tema, se abordaron aspectos como la comodidad de compra, la calidad de los productos, los precios y la satisfacción general. En el segundo tema, se discutió sobre estrategias comerciales, gestión empresarial y decisiones de los líderes de Mercadona. Finalmente, en el tercer tema, se destacaron las interacciones entre los clientes y la empresa, incluyendo recomendaciones de productos, promociones y atención al cliente.

El análisis de sentimientos revela una tendencia negativa en los tweets sobre Mercadona, aunque la mayoría son neutrales. Se observa una evolución temporal del sentimiento general y de cada tópico, destacando marzo de 2020 durante el confinamiento como un momento relevante. Por su parte, los consumidores expresaron malestar por la escasez de productos de primera necesidad, las largas colas y la aglomeración en los supermercados. También se mencionaron aspectos empresariales y de liderazgo, como el presidente de Mercadona, Juan Roig, aunque este tema fue menos relevante en las conversaciones en Twitter.

Abstract

The presence on social media is crucial for the development and growth of companies as it provides them with exposure to the public and access to brand perception. Proper management of these platforms has become essential due to technological advancements and the increased use of social media for customer interaction. This shift has diversified communication channels and allowed users to express public opinions. This is how social media has become an influential factor in customer retention and attraction.

This undergraduate thesis aims to analyze the perception and sentiments of users towards the renowned retail brand Mercadona. To achieve this, an exhaustive analysis of the state of the art was conducted to identify the most suitable techniques. In topic modeling, it was determined that the LDA technique is widely employed, so it was selected for this study. As for sentiment analysis, it was found that the most commonly used techniques incorporate rule-based decision-making using dictionaries, and in this particular project, the VADER strategy was implemented.

Thus, this work is based on the analysis of a dataset of 115,594 tweets in which the word "mercadona" is mentioned. Through topic modeling, three main themes were identified in the tweets related to Mercadona: the shopping experience and product variety, business and leadership factors, and interactions with customers and featured products. The first theme addressed aspects such as shopping convenience, product quality, prices, and overall satisfaction. The second theme discussed business strategies, management, and decisions made by Mercadona's leaders. Finally, the third theme highlighted interactions between customers and the company, including product recommendations, promotions, and customer service.

The sentiment analysis reveals a negative trend in the tweets about Mercadona, although the majority are neutral. There is a time evolution of the overall sentiment and each topic, with March 2020 during the lockdown standing out as a relevant moment. Consumers expressed discontent over the scarcity of essential products, long queues, and overcrowding in supermarkets. Business and leadership aspects, such as Mercadona's president, Juan Roig, were also mentioned, although this topic was less relevant in Twitter conversations.

Agradecimientos

A mi padre y a mi madre por siempre creer en mí y levantarme en los momentos difíciles. A mis hermanas por alegrarme los días y apoyarme incondicionalmente. A Isa, mi mayor apoyo. A mi directora del proyecto, Alexandra, por estar siempre disponible para ayudarme. A mis compañeros de clase, por ser grandes amigos y empujarme a dar lo mejor de mí.

Índice general

1. Introducción	1
1.1. Motivación	1
1.2. Objetivos	4
1.2.1. Objetivo General	4
1.2.2. Objetivos Específicos	4
1.3. Organización de la Memoria	4
2. Analítica de textos orientada a entender la percepción de los consumidores: Revisión de la literatura	5
3. Metodología	12
3.1. Recopilación y exploración de los datos	13
3.2. Preprocesado de datos	13
3.3. Análisis Exploratorio de N-Gramas	16
3.4. Modelado de tópicos	16
3.5. Análisis de sentimientos	20
4. Resultados	22
4.1. Análisis exploratorio inicial	22
4.2. Análisis exploratorio de N-gramas	26
4.3. Modelado de tópicos	29
4.4. Análisis de sentimientos	33
5. Conclusiones	38
Bibliografía	43

Índice de figuras

1.1. Proporción de empresas que emplearon las redes sociales en España durante el periodo 2014-2021	2
1.2. Usuarios activos y Facturación Mundial de Twitter	3
3.1. Metodología de análisis implementada.	13
3.2. Cuota de mercado de las grandes cadenas de supermercados en España en 2022. Fuente: Statista (Statista, 2022a)	14
3.3. Algoritmo LDA	19
4.1. Evolución temporal del número de tweets	23
4.2. Distribución de tweets por número de retweets	24
4.3. Distribución de tweets por idioma	25
4.4. Wordcloud de HASHTAGS con mayor frecuencia	25
4.5. Distribución de emoticonos	26
4.6. Nube de palabras tweets Mercadona	27
4.7. Unigramas más relevantes	28
4.8. Bigramas más relevantes	28
4.9. Trigramas más relevantes	29
4.10. Evaluación de la coherencia según el número de tópicos.	30
4.11. Mapa de la distancia intertópica	31
4.12. Frecuencia de las categorías en el corpus de texto.	33
4.13. Representación gráfica de la distribución de la puntuación mediante un gráfico de cajas.	34
4.14. Distribución de la puntuación compuesta	35
4.15. Evolución temporal del sentimiento general	35
4.16. Evolución temporal del sentimiento, para cada uno de los tópicos	37

Índice de tablas

2.1. Publicaciones que explorarn el modelado de tópicos y el análisis de sentimiento en empresas de venta minorista.	11
4.1. Categorías, términos clave y n-gramas con mayor relevancia.	32

Acrónimos

<i>API</i>	Interfaz de Programación de Aplicaciones
<i>NB</i>	Naïve Bayes
<i>PLN</i>	Procesamiento de Lenguaje Natural
<i>NLTK</i>	Natural Language Toolkit
<i>LDA</i>	Latent Dirichlet Allocation
<i>O2O</i>	Online to Offline
<i>URL</i>	Uniform Resource Locator
<i>TF-IDF</i>	Fórmula de Frecuencia Inversa de Documento
<i>VADER</i>	Valence Aware Dictionary for Sentiment Reasoning

Capítulo 1

Introducción

1.1. Motivación

En la actualidad, la presencia en redes sociales es indispensable para el desarrollo y el crecimiento de cualquier empresa. De hecho, se considera una característica crucial para obtener una exposición al público general. Esta exposición genera una mayor accesibilidad a información sobre la percepción positiva o negativa de la marca, y por ello, es importante centrar esfuerzos en la gestión adecuada de estas plataformas. Cada vez, son más las empresas que utilizan este nuevo medio de comunicación con los clientes, adaptándose a la inevitable evolución tecnológica. De hecho, en España, el porcentaje de empresas con presencia en redes sociales ha aumentado hasta a un 67 % para el año 2021 (Ver Figura 1.1).

Es así, como este nuevo paradigma asociado a la irrupción de las redes sociales ha alterado la forma de comunicación de las empresas con sus clientes, así como las diferentes estructuras de sus procesos. De esta manera, los medios de comunicación entre empresa y consumidor se han diversificado añadiendo las redes sociales a los ya existentes, como la vía telefónica y el correo electrónico. A través de estas plataformas, los usuarios han podido hacer públicas sus quejas y comentarios en un espacio en el que las empresas son accesibles por cualquier persona en cualquier momento. Esta gran diversidad en conjunto con el carácter público de estas revisiones/opiniones, convierte a las redes sociales en un factor de gran influencia para la retención de consumidores actuales y la atracción de clientes potenciales.

Con el fin de contar con información actualizada, se hace necesario monitorizar constantemente la información que se obtiene de redes sociales, debido principalmente a las complejas dinámicas de las opiniones de los consumidores. Para ello, hace falta una infraestructura adecuada y personal especializado, suponiendo una importante inversión para las empresas. Es por esto que la presencia online estratégica y el uso de plataformas de redes sociales, como Facebook o Twitter, se han convertido en herramientas esenciales para obtener un crecimiento sostenible a largo plazo y un apoyo fiel y continuo por parte de los consumidores (Roman y Cuestas, 2008).

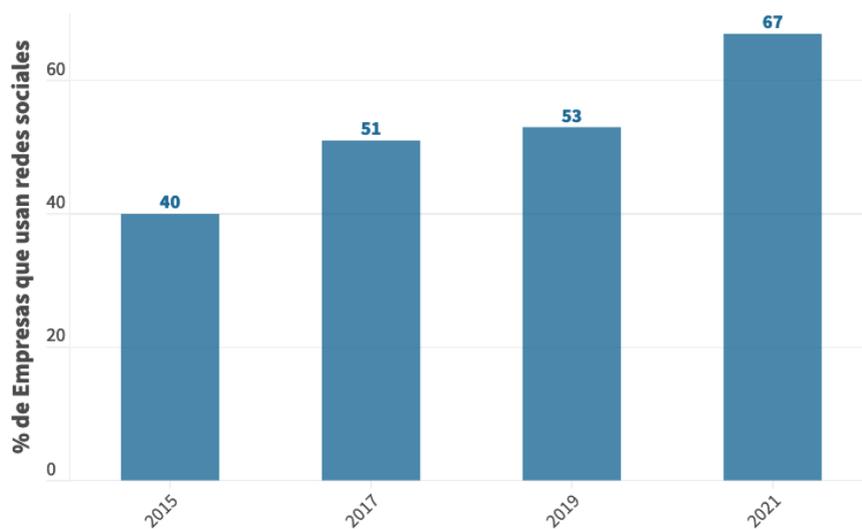


Figura 1.1: Proporción de empresas que emplearon las redes sociales en España durante el periodo 2014-2021

Fuente de Datos: (Statista, 2022b). Elaboración Propia

El objetivo buscado desde el sector empresarial consiste en crear una comunidad online de clientes, formada por personas con intereses similares que se comuniquen entre sí, acerca de la marca y la promoción entre el resto de usuarios de la red social. Así, la idea fundamental consiste en construir el concepto de lealtad de los clientes, el cual, es un elemento clave en el proceso de mercadeo de cualquier empresa (Park, Kee, y Valenzuela, 2009; Raacke y Bonds-Raacke, 2008). En este contexto, el uso de redes sociales permite la creación de estrategias de marketing digital más efectivas y con una gran proyección de evolución asociada al continuo crecimiento del volumen de clientes.

Actualmente, existe una gran variedad de proveedores de redes sociales, siendo los más importantes: Facebook, Youtube, Whatsapp, Instagram, Tiktok, Twitter y LinkedIn. Específicamente, Twitter es una de las redes sociales con mayor longevidad contando con una gran diversidad en las características de sus usuarios, y abarcando diferentes tramos de edad y ubicaciones geográficas en el mundo. Actualmente, cuenta con más de 360 millones de usuarios activos y acumula una facturación anual de más de 5000 millones de dólares para el año 2021 (Ver Figura 1.2). Esta plataforma incluye un servicio de microblogging que permite a los usuarios crear, escribir y leer mensajes cortos llamados *tweets*, los cuales tienen una longitud máxima de 280 caracteres. Asimismo, Twitter permite a los usuarios hablar libremente acerca de sus experiencias en cualquier ámbito, y compartir el contenido creado por otros usuarios a sus seguidores, posibilitando así una rápida difusión de los tweets en cualquier tema de discusión. Aunque existe la posibilidad de configurar la cuenta de manera privada, Twitter es una red social principalmente pública, lo cual permite la obtención de una gran parte de la información publicada. En este contexto, Twitter proporciona una elevada

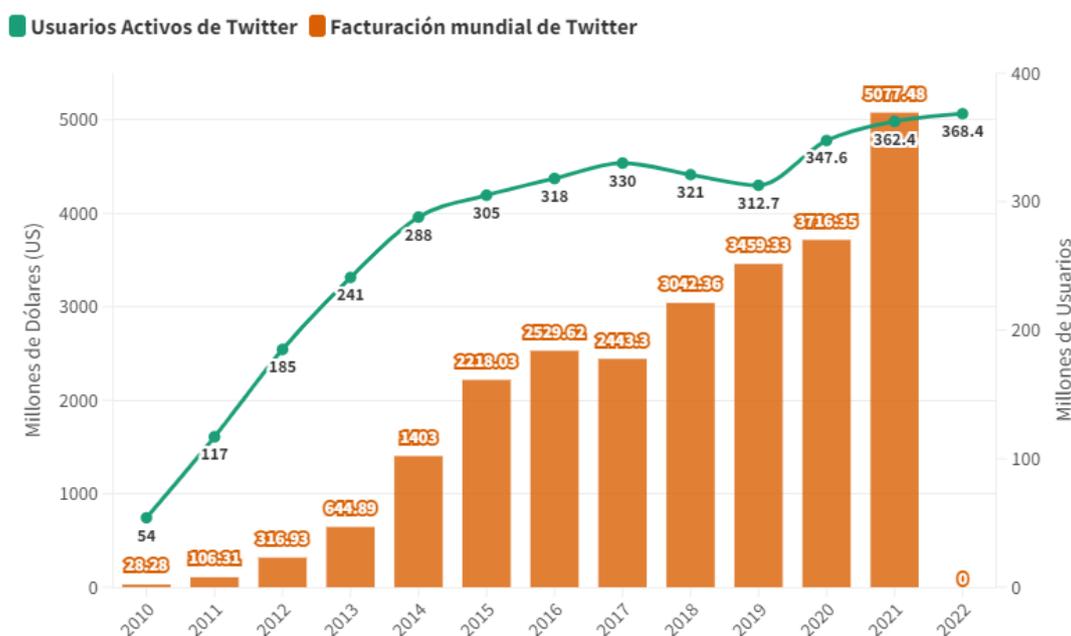


Figura 1.2: Usuarios activos y Facturación Mundial de Twitter
Fuente de Datos: (Bankmycell, 2022; Statista, 2022c). Elaboración Propia

cantidad de datos en tiempo real sobre las opiniones de sus usuarios, siendo muy útil para el análisis de información por parte de investigadores y empresas. De hecho, el volumen masivo de datos obtenido a través de este medio ha sido ampliamente estudiado en conjunto con estrategias automáticas de procesamiento de texto y de machine learning, permitiendo realizar minería de opiniones y de sentimientos asociados a la marca, así como su difusión en el diseño de estrategias de marketing digital (Pang y Lee, 2008).

De esta manera, esta plataforma se ha convertido en la principal fuente de datos en el análisis de la percepción ciudadana para una gran variedad de aplicaciones. De hecho, una encuesta realizada en Twitter en el año 2015, reportó que, desde el punto de vista empresarial, los consumidores prefieren comentar sus conflictos con las diferentes compañías a través de Twitter, en lugar de otras redes sociales como Facebook (Ibrahim y Wang, 2019), haciendo de Twitter la red social más usada por los consumidores para expresar sus opiniones y críticas respecto a la marca. Asimismo, un estudio realizado por (Nagy y Midha, 2014) determinó que el 80 % de los usuarios de Twitter consideraban normal mencionar marcas en sus tweets, y el 40 % admitió también alguna vez haber realizado una compra debido a una publicación en Twitter (Digital Marketing Institute, 2021).

Con base en la información presentada anteriormente, se puede ver que el uso cotidiano de las redes sociales también ha cambiado la forma de relacionarse entre empresas y consumidores, siendo estas esenciales en la valoración de la calidad del servicio a los clientes. Por su parte, Twitter reúne varias características que la convierten en una herramienta valiosa para la adaptación de las empresas a este nuevo medio de comunicación con los clientes.

Específicamente, el uso de esta fuente de datos junto a la implementación de tecnologías de análisis automático de texto, modelado de tópicos y análisis de sentimientos, presentan una alternativa de interés para extraer información significativa sobre las expectativas y la percepción de marca por parte de los consumidores.

1.2. Objetivos

1.2.1. Objetivo General

Analizar la percepción ciudadana sobre el servicio de venta de una de las empresas minoristas más importantes en España, de forma que se extraiga información relevante para determinar puntos fuertes y áreas de mejora en la empresa. Los medios que se utilizarán son la recopilación de datos de Twitter y la implementación de estrategias automáticas de minería de texto.

1.2.2. Objetivos Específicos

- Contextualizar la importancia que tiene el análisis de redes sociales en la extracción de información relevante sobre la percepción de los consumidores en el sector de venta minorista.
- Analizar las estrategias automáticas de minería y procesamiento de texto en redes sociales, identificando su importancia en la obtención de información relevante para los servicios de venta de diversas compañías.
- Extraer y analizar información relevante sobre la percepción ciudadana de una marca asociada a empresas de venta minorista en España, mediante la adquisición de datos de redes sociales y la aplicación de técnicas automáticas de análisis de texto

1.3. Organización de la Memoria

Este Trabajo de Fin de Grado está dividido en cinco capítulos. El capítulo 2 resume el estado del arte en investigaciones dedicadas al análisis de redes sociales, entre ellas Twitter, y el desarrollo de técnicas como el modelado de tópicos y el análisis de sentimientos, en áreas de aplicación asociadas al análisis de las percepciones de marca por parte de los consumidores. El capítulo 3 expone la metodología empleada en el desarrollo de este estudio, dedicando una sección a la extracción y pre-procesamiento de los datos, otra al modelado de tópicos y una última al análisis de sentimientos. El capítulo 4 detalla los principales resultados siguiendo la metodología presentada anteriormente, y, por último, el capítulo 5 detalla las principales conclusiones obtenidas.

Capítulo 2

Analítica de textos orientada a entender la percepción de los consumidores: Revisión de la literatura

El objetivo principal de este trabajo de fin de grado es proporcionar información valiosa sobre la percepción de los consumidores hacia una de las principales empresas minoristas de España, a través del análisis de los datos obtenidos de Twitter. Las empresas minoristas han encontrado en esta plataforma una manera eficaz de comunicarse con sus clientes y de conocer sus opiniones y necesidades en tiempo real. A su vez, Twitter se ha convertido en una herramienta fundamental para las empresas minoristas que buscan mejorar su estrategia de marketing, ya que les permite recibir retroalimentación inmediata sobre la calidad de sus productos y servicios. En este contexto, el modelado de tópicos y el análisis de sentimientos son técnicas que se ha popularizado en los últimos años y que se han utilizado para comprender las emociones subyacentes de los consumidores hacia una marca o producto. Estos análisis permiten a las empresas minoristas obtener información sobre la percepción de sus clientes y entender las tendencias y la dinámica del mercado.

En los últimos años, las estrategias de minería de textos y de procesamiento de lenguaje natural (PLN) se han convertido en un campo de investigación activo y en constante evolución. En este sentido, numerosos autores han propuesto diversas estrategias para extraer información relevante de las redes sociales, con el fin de analizar las percepciones y opiniones de los usuarios. En este contexto, estas estrategias de análisis automático se han utilizado, por ejemplo, para evaluar la satisfacción del cliente y comprender las opiniones sobre productos y servicios, así como para identificar las tendencias de compra y los patrones de comportamiento de los consumidores, lo que permite a las empresas diseñar campañas de publicidad más efectivas y personalizadas.

En esta línea de investigación, se destaca el estudio realizado por (He, Zha, y Li, 2013) que llevó a cabo un análisis competitivo de redes sociales enfocado en el marketing y la publi-

cidad de las tres cadenas de pizzas más importantes de Estados Unidos: Pizza Hut, Domino's Pizza y Papa John's Pizza, utilizando Facebook y Twitter. Para ello, se recopilaron datos manualmente y se aplicaron técnicas de minería y análisis de texto usando SPSS Clementine y Nvivo 9 con el fin de analizar sus publicaciones. A partir del análisis, se identificaron cinco temas relevantes: pedidos y entregas, calidad de la pizza, comentarios sobre las decisiones de compra de los clientes, publicaciones de socialización casual y publicaciones de marketing. Los resultados del estudio indicaron que las temáticas abordadas con mayor relevancia fueron aquellas asociadas a los servicios de pedido y seguimiento en línea (alrededor del 31 %) y a la calidad de la pizza en sabor e ingredientes (28 %). Asimismo, las cadenas utilizaron las redes sociales para encuestar a los clientes y mejorar sus productos según los comentarios obtenidos, lo que demuestra la importancia del análisis de sentimientos y la minería de texto en el ámbito del marketing y la publicidad.

Por otra parte, el estudio llevado a cabo por (Ghiassi, Skinner, y Zimbra, 2013) recopiló un gran volumen de datos a través de la API de Twitter, en concreto, 10.345.184 tweets desde el 3 de mayo hasta el 8 de junio, procedentes de la cuenta de Justin Bieber, quien en ese momento era la persona con mayor actividad de tweets. Con el fin de analizar la reputación de la marca, se propuso una nueva estrategia para mejorar la precisión en la detección de tweets neutrales, que consistía en combinar el análisis de n-gramas con redes neuronales, usando la red neuronal DAN2. Este estudio permitió a los investigadores clasificar tweets que pudieran tener una ligera tendencia hacia lo negativo o lo positivo. Con esta metodología, sólo el 10 % de los tweets relacionados con Justin Bieber fueron determinados como neutrales. Además, se tuvieron en cuenta los emoticonos, algo que no se había considerado en estudios previos. De esta manera, aumenta la cantidad de datos que se pueden emplear en el análisis y se observa que también mejora la capacidad de clasificación de sentimientos en los tweets. Por otro lado, existe un coste debido al aumento de falsos positivos.

En una línea de aplicación alternativa, (Yu, Duan, y Cao, 2013) realizaron un análisis sobre una selección de 824 empresas de diferentes industrias, entre ellas la farmacéutica, venta minorista, software, entidad de crédito, sanitaria y hostelería. Para determinar si los medios de comunicación influyen en el desempeño del mercado de valores, se recopiló una muestra de 52.746 mensajes de varias redes sociales y medios convencionales. Para el análisis de sentimientos se empleó el algoritmo de NB y la librería NLTK para el preprocesamiento del texto. Los resultados mostraron que las redes sociales tienen más influencia que los medios convencionales. Específicamente, se encontró que los sentimientos analizados en Twitter y blogs tienen un efecto positivo en el riesgo, de la misma manera que los blogs tienen un impacto positivo en la ganancia. Además, se observó que la interacción entre el sentimiento de los tweets y el sentimiento de las noticias tiene un impacto significativamente negativo en los rendimientos, aunque no se encontró un efecto significativo en el riesgo. Estos resultados demuestran la importancia de monitorear la percepción de los clientes en los medios sociales para el éxito de una empresa.

Por su parte, en la investigación llevada a cabo por (Ibrahim, Wang, y Bourne, 2017), se analizan los datos de Twitter desde la perspectiva del impacto de la interacción de las marcas en las redes sociales en los consumidores. Para ello, se extrajeron 76.166 tweets de cinco empresas de venta en línea del Reino Unido y se realizó un análisis de sentimientos en cada marca usando Node Pro para la recolección de datos y SentiStrength para el análisis de sentimientos. El análisis permitió identificar 29 patrones de palabras similares que se mencionan constantemente, como “gracias”, “bueno”, “me gusta” y “amo”. Estas palabras tienen una connotación positiva y demuestran que los consumidores están satisfechos con el servicio que reciben. Además, se observó que las empresas de la competencia aparecían en la nube de palabras de cada empresa del sector de alimentación, lo que sugiere que los consumidores comparan entre marcas en un sector altamente competitivo.

En una línea de análisis similar, el estudio presentado por (Akundi y cols., 2018) se enfoca en analizar las principales marcas de smartphones, como el iPhone, Samsung y Huawei. Para ello, se utilizaron los hashtags con los nombres de las marcas en Twitter y se obtuvieron los tweets correspondientes. Se seleccionaron las 10 palabras más frecuentes para el análisis de opinión. Éste se realizó usando la API de Twitter en la extracción de los datos y el diccionario de sentimientos SENTIWORD 3.0.0. Como resultado, se observó que el sentimiento hacia Samsung y Huawei era más positivo en comparación con Apple, especialmente en cuanto a precio y diseño. Por otro lado, el sentimiento hacia el sistema operativo iOS fue más positivo en comparación con el sistema operativo de Android. En el análisis de la percepción de marca, (Ibrahim y Wang, 2019) se centran también en el análisis de 386.379 tweets, recopilados usando la API de Twitter, relacionados con 5 empresas líderes en el sector de la venta online en Reino Unido durante el Black Friday. A través de la técnica de modelado de tópicos, usando LDA, se identificaron los temas más discutidos en Twitter, siendo los más relevantes la entrega, el producto y el servicio al cliente. Además, se concluyó que estos dos últimos aspectos requerían de una mayor atención por parte de las empresas.

En el sector del comercio O2O, Shen, Chen, y Wang (2019) se utilizó la minería de texto bilingüe para analizar los datos de las redes sociales y entender las tendencias en el comercio O2O. La investigación se centró en la identificación de patrones de comportamiento de los consumidores en línea y en el mundo real, y cómo estas tendencias están influyendo en el sector. Para ello, se recopilaron 19.273 tweets en inglés y 17.721 en chino usando Advanced Search Tool. El análisis se realiza usando Netlingo, SMS Dictionary y Royal Standard English Dictionary. Además, se identificaron las palabras claves: “China”, “Alibaba”, “Baidu”, “DianPing”, “Meituan”, “Asia”, “ecommerce” y “service”, que se dividen en países o áreas, empresas, plataformas O2O y modelos de servicio. Se puede observar que la mayoría de keywords están relacionados con el comercio O2O en países asiáticos, por lo que se puede concluir que los usuarios angloparlantes están interesados en ese mercado. El mercado de fast fashion también ha sido ampliamente estudiado, como se refleja en el análisis de sentimientos realizado por (Pantano, Giglio, y Dennis, 2019). En este estudio se analizan tres marcas

anónimas de Londres, denominadas como Retailer A, B y C, recopilando un total de 9.652 tweets con el API de Twitter. Cada empresa adopta diferentes estrategias de marketing, lo que se traduce en distintos resultados y, por tanto, un contenido generado diferente por los usuarios en las redes sociales. Se realiza un análisis de sentimientos usando la función de clasificación de Wolfram Mathematica, cuyos resultados muestran que la marca C no tiene tweets negativos y tiene la mayor proporción de tweets positivos, mientras que la marca B presenta la mayor cantidad de tweets negativos.

En el mercado de ropa y calzado deportivo, un estudio realizado por (Rasool, Tao, Marjan, y Naveed, 2019) se enfoca en las marcas líderes Adidas y Nike. Se recopilaron 99.850 tweets usando como palabras clave “Nike” y “Adidas” mediante Tweepy, una biblioteca de Python que permite interactuar con la API de Twitter de una manera más sencilla. Posteriormente, se realizó un análisis de sentimientos para ambas marcas usando el algoritmo de NB. El resultado mostró que Adidas tenía más reseñas positivas que Nike. Los tweets neutros denotaron la satisfacción de los consumidores con ambas marcas, constituyendo más del 60 % de los tweets. Además, se realizó una nube de palabras en la que se encontraron términos tanto positivos como negativos, en la que ambas marcas mencionaban a su competencia. Finalmente, en un estudio reciente, (Mishra, 2022) usan el algoritmo LDA para extraer las temáticas de discusión de los tweets relacionados con empresas de venta minorista en línea, incluyendo Amazon, Flipkart, Snapdeal, BookMyShow, Myntra y Nykaa. A partir de una muestra de 25,173 tweets de consumidores obtenida usando la API de Twitter, se normalizan los tweets usando NLTK. A continuación, se identificaron cinco temas principales: realización de pedidos, interacción con atención al cliente, entretenimiento, fallos de servicio y fallos de proceso; siendo el más frecuente el relacionado con atención al cliente y su calidad.

En la tabla 2.1 se puede observar un resumen de los resultados obtenidos en las investigaciones analizadas. En general, se muestra que se utilizó un volumen de tweets que varió entre 20.000 y 80.000, los cuales fueron extraídos utilizando como filtro las marcas de interés y en una franja temporal determinada. En cuanto a las estrategias de análisis, se utilizaron técnicas automáticas orientadas a modelar las categorías de discusión y a analizar los sentimientos subyacentes con el objetivo de realizar un análisis de la competencia o determinar los temas más relevantes para los consumidores de la marca.

En cuanto a las herramientas utilizadas para la extracción de datos, se observa que se utilizó principalmente la API de Twitter. En términos del modelado de tópicos, la estrategia más utilizada fue LDA debido a su facilidad de interpretación y a que identifica los temas sin necesidad de realizar un etiquetado manual previo. Por su parte, para el análisis de sentimientos, las estrategias más utilizadas fueron los algoritmos basados en diccionarios y en estrategias de machine learning como el clasificador NB. Por una parte, el algoritmo NB se considera una estrategia efectiva, robusta y escalable capaz de manejar grandes cantidades de datos, lo que lo hace una opción popular en el análisis de sentimientos. Por otro lado, los diccionarios de sentimientos también son ampliamente utilizados ya que son sencillos de

implementar e interpretar y no necesitan entrenamiento previo.

Cita	Empresa	tweets	Filtros	Objetivo	Tarea	Estrategia	Sector	Resultado principal
(He y cols., 2013)	Pizza Hut, Domino's Pizza, Papa Jhon's Pizza	135	@pizzahut, @dominos y @PapaJohns del 1 de Octubre hasta el 31 de Octubre de 2011	Análisis competitivo en redes sociales	Análisis competitivo	SPSS Clementine and Nvivo9 (Text Mining Analisis)	Cadenas de pizza de Estados Unidos	5 temas relevantes: pedidos y entregas, calidad de la pizza, comentarios sobre las decisiones de compra de los clientes, publicaciones de socialización casual y publicaciones de marketing
(Ghiassi y cols., 2013)	Justin Bieber	10.345.184	"Justin Bieber" del 6 de Mayo a las 7am hasta el 8 de Junio a las 7am en 2013	Análisis de la marca mejorando la precisión de las publicaciones neutras e incluyendo emoticonos	Análisis de sentimientos	Twitter API (obtención de datos), Lexalytics(diccionario sentimientos), DAN2 y SVM (Análisis de sentimientos)	NA	Se encontraron 181 expresiones de sentimientos generales de Twitter y 6 propias de la marca Justin Bieber
(Yu y cols., 2013)	824 empresas	24.505	Nombre de las empresas del 2011/07/01 hasta 2011/09/30	Determinar el efecto de las redes sociales y medios convencionales en el valor de las acciones	Análisis de sentimientos	NB y NLTK (Análisis de sentimientos)	farmacéutica, venta minorista, software, entidad de crédito, sanitaria y hostelería	Se determinó que las redes sociales tenían más influencia que los medios convencionales. Los sentimientos analizados en Twitter y blogs tienen un efecto positivo en el riesgo. Por otro lado, los blogs tienen un impacto positivo en la ganancia, mientras que los foros tienen un impacto negativo
(Ibrahim y cols., 2017)	Amazon, John Lewis, Tesco, Asda, Argos	76.166	[nombre marca] lang:en desde:2016-04-10 hasta:2016-04-11]	Analizar el impacto en los consumidores en la interacción de las marcas en Twitter	Análisis de sentimientos	NodeXL Pro (Recolección de datos), SentiStrength (análisis de sentimientos)	Venta minorista Reino Unido	Se encontraron 29 patrones similares de palabras que eran mencionadas constantemente, como 'thanks', 'good', 'like' y 'love'. Tiene connotaciones positivas y muestras que los consumidores están satisfechos. Las empresas de la competencia aparecen en la nube de palabras de cada empresa en el sector de alimentación, lo que demuestra que los consumidores comparan entre marcas
(Akundi y cols., 2018)	iPhone, Samsung, Huawei, IOS, Android	51.404	#iPhone, #Samsung, #Huawei, #IOS, #Android	Análisis de opinión acerca de los materiales y tecnologías de cada smartphone	Análisis de sentimientos	Twitter API (obtención de datos), SENTIWORD 3.0.0 (diccionario de sentimientos)	Cadena de suministro de Smartphones	El sentimiento hacia Samsung y Huawei era más positivo que el de Apple, sobre todo por el precio y el diseño. Sin embargo, el sentimiento hacia el sistema operativo IOS era más positivo que el de Android

Cita	Empresa	tweets	Filtros	Objetivo	Tarea	Estrategia	Sector	Resultado principal
(Ibrahim y Wang, 2019)	Amazon UK, Argos, Asda, John Lewis, Tesco	386.379	[nombre marca] lang:en since:2016-11-20 until:2017-01-20]	Analizar los temas más relevantes para los consumidores respecto a empresas de venta minorista durante el Black Friday.	Modelado de tópicos y Análisis de sentimientos	Twitter API (obtención de datos), LDA (modelado de tópicos), SentiStrength (análisis de sentimientos)	Venta minorista Reino Unido	Se determinó que los temas asociados a la entrega, el producto y el servicio al cliente eran los más discutidos en Twitter, siendo los dos últimos los que mayor mejora necesitaban.
(Shen y cols., 2019)	Alibaba, Baidu, DianPing, Meituan	36.994	“O2O”, “China”, “Alibaba”, “Baidu”, “DianPing”, “Meituan”, “Asia”, “ecommerce” y “service” del 1 de Enero de 2014 hasta el 31 de Diciembre de 2015	Analizar las tendencias del comercio O2O en Inglés y Chino	Análisis de tendencias	Advanced Search Tool (obtención de datos); Netlingo, SMS dictionary y Royal Standard English Dictionary (filtrado de texto)	O2O	Se puede observar que la mayoría de keywords están relacionados con el comercio O2O en países asiáticos, por lo que los usuarios angloparlantes están interesados en ese mercado.
(Pantano y cols., 2019)	3 marcas anónimas A, B y C	9652	nombre de la marca como hashtag o palabra en un tweet, solo en inglés	Explorar el contenido generado por los consumidores online para mejorar el marketing inteligente	Análisis de Sentimientos	Twitter API (obtención de datos), la función de clasificación de Wolfram Mathematica (clasificación de sentimientos)	Fast fashion	La empresa C no muestra ningún tweet negativo y, además, tiene la mayor proporción de tweets positivos. Por otro lado, la empresa B tiene la mayor proporción de tweets negativos.
(Rasool y cols., 2019)	Adidas, Nike	99.850	“Adidas”, “Nike” en fechas irregulares de 2018	Análisis de la opinión pública acerca de las dos marcas más populares en el sector	Análisis de sentimientos	Tweepy (obtención de datos), NB (análisis de sentimientos)	Ropa deportiva	Se observa que Adidas tiene más reseñas positivas que Nike. Los tweets neutros muestran la satisfacción de los consumidores con ambas marcas y son más del 60 %. También, en la nube de palabras de ambas marcas se menciona a la otra marca.
(Mishra, 2022)	Amazon, Flipkart, Snapdeal, BookMyShow, Myntra y Nykaa	25.173	“amazon”, “flipkart”, “snapdeal”, “bookmyshow”, “myntra” y “nykaa” de Enero de 2015 hasta Octubre de 2019	Extraer tópicos de tweets acerca de empresas de venta minorista online	Modelado de tópicos	API Twitter (obtención de datos), NLTK (normalización de texto), LDA (extracción de tópicos)	Venta minorista online	Se determinan 5 tópicos: realización del pedido, interacción con atención al cliente, entretenimiento, fallo de servicio y fallo de proceso.

Tabla 2.1: Publicaciones que exploran el modelado de tópicos y el análisis de sentimiento en empresas de venta minorista.

Capítulo 3

Metodología

La metodología seguida para alcanzar los objetivos planteados en este Trabajo de grado se especifican en la Figura 3.1:

1. **Recopilación y exploración de datos:** En primer lugar, se seleccionó la empresa minorista a estudiar, así como las palabras clave, hashtags o cuentas relacionadas para extraer los datos de Twitter. Una vez recopilados, se explorarán las variables y la cantidad de tweets obtenidos.
2. **Preprocesamiento:** Una vez realizada la adquisición de los tweets, se les aplicarán diversos algoritmos de preprocesamiento de texto. Específicamente, se utilizarán técnicas como la eliminación de ruido, la lematización, la eliminación de duplicados, la tokenización y la reducción de términos.
3. **Análisis N-Gramas:** Se realizará un análisis descriptivo de los unigramas, bigramas y trigramas encontrados en los tweets y se estimará su correspondiente relevancia. Este análisis de n-gramas proporcionará información valiosa para un primer análisis exploratorio.
4. **Modelado de Tópicos:** Se realizará el modelado y la identificación de los tópicos de discusión encontrados en el conjunto de datos analizados. Una vez que se hayan identificado los tópicos, se analizará la frecuencia de aparición de cada uno de ellos en el conjunto de datos para poder entender mejor las tendencias de discusión en Twitter.
5. **Análisis de sentimientos:** Este análisis permitirá evaluar el sentimiento subyacente a la opinión de los usuarios, en relación a la marca seleccionada y sus productos. Estos resultados son de gran importancia para el análisis, pues permiten identificar los aspectos específicos de los productos o servicios que generan opiniones positivas o negativas en los usuarios, lo que será útil para el análisis de fortalezas y debilidades de la marca.

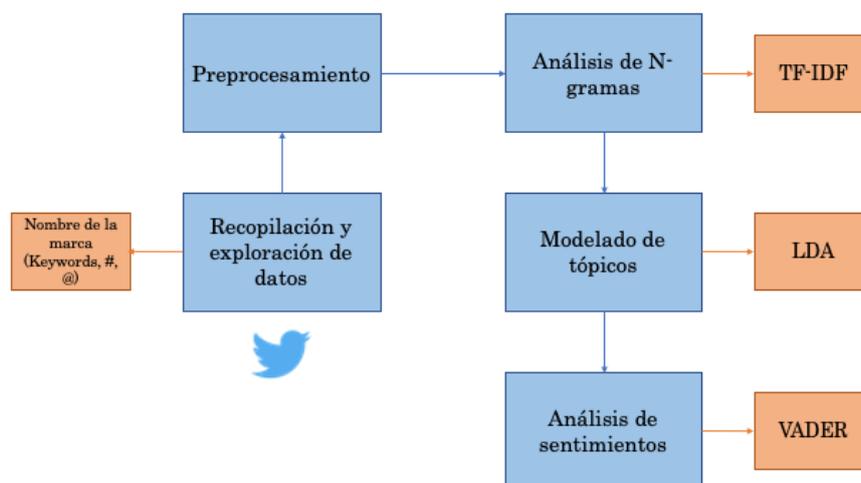


Figura 3.1: Metodología de análisis implementada.
Elaboración Propia

3.1. Recopilación y exploración de los datos

Para llevar a cabo el presente proyecto, es necesario obtener los datos necesarios para realizar el análisis. En primer lugar, se procederá a seleccionar la marca a estudiar en el sector de venta minorista. Para ello, se ha elegido la cadena de supermercados más importante en España, Mercadona, de acuerdo con la información de cuota de mercado mostrada en la Figura 3.2. Las palabras clave definidas para la recolección de los tweets son “mercadona” y el periodo de análisis comprende desde el 01-01-2017 hasta el 22-11-2022. De esta manera, se podrá estudiar el efecto que ha tenido la pandemia de COVID-19 en la marca. Además, se obtendrán únicamente los tweets procedentes de España, ya que el análisis se centra en este país. Todos estos elementos son esenciales para extraer los datos del API de Twitter.

Una vez obtenida esta información, se podrá proceder con la extracción de datos mediante el uso de la biblioteca la librería *academictwitter* de R y las palabras clave definidas previamente, filtrando únicamente los tweets en español. De esta manera, se garantiza que se estarán analizando únicamente aquellos tweets que sean relevantes para el análisis, lo que permitirá obtener resultados más precisos y fiables. Es importante destacar que la correcta selección de las marcas y palabras clave es un factor crítico para el éxito del proyecto, ya que de ello dependerá la calidad y relevancia de los datos obtenidos.

3.2. Preprocesado de datos

El preprocesamiento de los datos es esencial para lograr resultados precisos en las tareas de modelado de tópicos y el análisis de sentimientos. Para ello, es necesario normalizar los

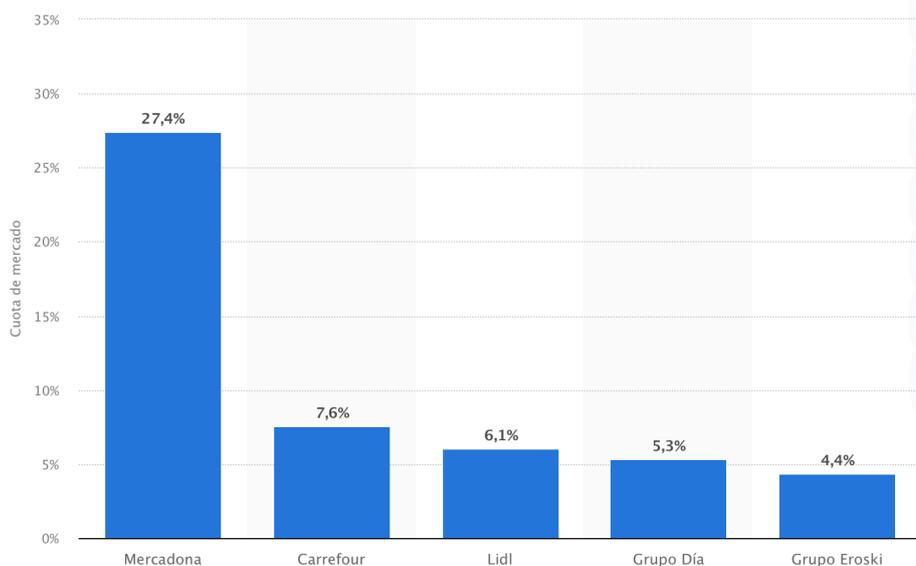


Figura 3.2: Cuota de mercado de las grandes cadenas de supermercados en España en 2022. Fuente: Statista (Statista, 2022a)

textos mediante la implementación de algoritmos que simplifiquen el análisis de un conjunto de palabras. En este proyecto de fin de grado, se seguirán los siguientes pasos: eliminación del ruido (como URLs, menciones de usuarios, hashtags, signos de puntuación, emojis, números, palabras con letras triplicadas y la expresión de risa ‘jaja’, espacios y caracteres extraños), lematización, revisión de errores y duplicados, tokenización y, finalmente, una reducción de términos. Este proceso de preprocesamiento es crucial para obtener una lista de palabras independientes, lo que permite calcular la frecuencia de cada palabra y analizarlas de manera individual, sin importar el orden o el significado semántico de la palabra en cada tweet. A continuación, se detallarán con mayor profundidad cada uno de los tratamientos aplicados sobre los datos, con el objetivo de brindar una explicación más exhaustiva del pre-procesado.

- Eliminación del ruido:** En el primer paso del preprocesamiento, se transforman los datos a minúsculas y se eliminan elementos irrelevantes para el análisis, como URLs, menciones de usuarios, hashtags, signos de puntuación, emojis, números, espacios, caracteres y líneas innecesarias. Además, se realiza una normalización de expresiones de risa, como “ajjajja” a “jaja”, y de palabras con letras triplicadas. Es importante destacar la eliminación de palabras vacías, las cuales no contribuyen al significado de una oración, sino que solo brindan soporte gramatical, como preposiciones. Para ello, se utiliza el diccionario de palabras vacías proporcionado por el paquete NLTK, al cual se agregan manualmente otras palabras, que no tienen información significativa del contenido, para filtrarlas. Este proceso reduce el ruido y características irrelevantes en los datos, mejorando la calidad de los resultados del análisis.
- Lematización:** La lematización es una técnica importante en el procesamiento de tex-

to que tiene como objetivo reducir el tamaño del vocabulario total al transformar las palabras a su forma base o lema. Esta técnica agrupa diferentes formas de una misma palabra, como verbos en diferentes tiempos o personas, en una sola forma base. Por ejemplo, las palabras “correr”, “corriendo”, “corrieron” y “corre” tendrían el mismo lema: “correr”. En este análisis, se utilizará el modelo preentrenado “es_core_news_sm” de SpaCy en español para llevar a cabo la lematización. La lematización no solo ayuda a reducir el tamaño del vocabulario, sino que también mejora la eficiencia de los algoritmos de procesamiento de texto. Al utilizar los lemas en lugar de las formas originales de las palabras, los algoritmos pueden realizar cálculos de forma más rápida y precisa. Además, la lematización ayuda a mejorar la precisión del análisis de texto, ya que permite agrupar palabras similares y realizar comparaciones más significativas entre ellas.

- **Ajuste de errores y duplicados:** Después de llevar a cabo el proceso de lematización en los datos de Twitter, se requiere una revisión minuciosa de la lista completa de cambios efectuados por el lematizador. La detección de errores es crucial y, de encontrarse, deben ser corregidos para evitar posibles inexactitudes en el análisis posterior. Además, se busca unificar palabras que comparten significados similares para que todas se ubiquen en la misma raíz y así reducir el tamaño del diccionario tanto como sea posible. Aunque el ajuste manual no siempre es obligatorio, puede contribuir significativamente a mejorar la calidad de los resultados del análisis. Una vez que se han definido los lemas finales para cada tweet, es necesario llevar a cabo una eliminación de duplicados para garantizar la integridad y calidad de los datos de análisis. Este proceso ayuda a evitar la inclusión de datos repetidos y a simplificar la lista de términos, lo que a su vez mejora la eficiencia del análisis y la calidad de los resultados obtenidos.
- **Tokenización:** La tokenización es una etapa crucial en el procesamiento de texto para el análisis de lenguaje natural. Su objetivo principal es dividir un texto en fragmentos más pequeños y significativos llamados tokens, que pueden ser caracteres, palabras o subpalabras. Los tokens se pueden obtener mediante técnicas como la eliminación de signos de puntuación y espacios en blanco. Esto permite transformar el lenguaje natural en una lista formalizada de palabras, que es más fácil de procesar por algoritmos de PLN.
- **Reducción de términos:** Para evitar posibles términos irrelevantes en el léxico utilizado en el modelado de tópicos de las publicaciones de Twitter, se han establecido ciertas condiciones para que un término aparezca. Así, se ha decidido excluir aquellos términos que aparecen en menos del 2 % o en más del 90 % de los documentos analizados. De esta manera, se evitan palabras poco comunes o demasiado comunes que no aportan valor al análisis y que puede que no hayan sido eliminadas en el prepro-

cesamiento como stopwords. Con estas condiciones, se espera obtener un léxico más preciso y relevante para el análisis posterior de los datos.

3.3. Análisis Exploratorio de N-Gramas

El análisis exploratorio de n-gramas es una fase esencial utilizada en el PLN para modelar y analizar textos. Consiste en dividir el texto en secuencias de n palabras, llamadas n-gramas, y medir la relevancia de la aparición de estas secuencias. Estos modelos permiten convertir el texto en vectores con características numéricas, lo que facilita su procesamiento por parte de los algoritmos. En este proyecto, se analizan n-gramas de una, dos o tres palabras consecutivas, es decir, 1-gramas (unigramas), 2-gramas (bigramas) y 3-gramas (trigramas), junto con su medida de relevancia. Así, con el fin de medir la importancia de los n-gramas, se utiliza la herramienta TF-IDF.

El método TF-IDF consta de dos componentes importantes: la frecuencia de un término (TF) y la frecuencia inversa del documento (IDF). TF, definido en la ecuación 3.1, indica la frecuencia con la que aparece un término específico en un documento, y se calcula dividiendo el número de veces que se utiliza la palabra clave en un documento (tweet) por el número total de palabras en dicho documento. Por otro lado, IDF, definido en la ecuación 3.2, indica con qué frecuencia se utiliza una palabra clave en un conjunto de documentos. Este parámetro correctivo tiene como objetivo compensar la repetición de palabras comunes, como artículos o palabras con funciones similares, y dar más relevancia a las palabras significativas.

La puntuación TF-IDF se obtiene al multiplicar los valores de TF e IDF (ver ecuación 3.3). Aquellas palabras con una puntuación más alta son consideradas más importantes y relevantes para el documento, en contraste con las palabras que obtienen una puntuación más baja. En definitiva, el análisis de n-gramas junto con el método TF-IDF nos permite obtener un análisis descriptivo inicial de los temas y conceptos relevantes en un texto.

$$TF_i = \frac{\log_2(f(i, j) + 1)}{\log_2(L_j)} \quad (3.1)$$

$$IDF_i = \log_2\left(1 + \frac{N_D}{f_i}\right) \quad (3.2)$$

$$TF - IDF_i = TF_i \times IDF_i \quad (3.3)$$

3.4. Modelado de tópicos

El modelado de tópicos es una técnica de análisis de texto que se utiliza para identificar los temas principales o tópicos que aparecen en un conjunto de documentos. Es especialmente

útil cuando se tienen grandes cantidades de texto sin etiquetar y se busca descubrir los temas comunes presentes en los documentos. En términos generales, estos métodos analizan las palabras y descubren temas latentes midiendo sus interconexiones.

Estas estrategias se consideran un enfoque de aprendizaje automático no supervisado, lo que significa que no es necesario entrenar el modelo con documentos previamente etiquetados. El objetivo principal del modelado de tópicos es identificar los temas más relevantes en los documentos y asignar a cada uno de ellos una distribución de probabilidad, lo que permite analizar los patrones de tópicos en un conjunto de documentos y entender mejor la temática del texto analizado.

En el modelado de tópicos, existe una gran diversidad de algoritmos disponibles. Uno de los más populares y sencillos, como se ha visto en el capítulo 2, es el modelo LDA, cuya representación gráfica se muestra en la Figura 3.3. Este algoritmo parte de la premisa de que cada documento está compuesto por una combinación de tópicos y que cada tópico se compone de una mezcla de palabras clave. En otras palabras, LDA permite identificar los tópicos latentes en un conjunto de documentos al analizar la distribución de palabras y su interconexión en el corpus. Para entender su funcionamiento, se definirán los siguientes términos:

- Palabra (W): Elemento básico de vocabulario.
- Documento (d): Conjunto de N palabras.
- Corpus (C): Conjunto de d documentos.
- Tópico (K): temas extraídos de los documentos.
- α : Hiperparámetro de la distribución de Dirichlet. Controla la distribución de tópicos por documento.
- β : Hiperparámetro de la distribución de Dirichlet. Controla la distribución de palabras por tópico.
- θ : Distribución de tópicos para el documento d
- ϕ : Distribución de palabras en el tópico K

De esta manera, en el proceso de modelado, para cada documento d en el corpus C :

1. Se utiliza una distribución de probabilidad con un parámetro η , el cual controla la diversidad de las palabras que se pueden seleccionar para un tópico. De esta manera, se selecciona aleatoriamente un conjunto de palabras para cada documento. $d \sim \text{Poisson}(\eta)$

2. Para asignar tópicos a un documento en el modelo de tópicos LDA, se utiliza la distribución de Dirichlet. Esta distribución se elige aleatoriamente con un parámetro α , el cual controla la concentración de los tópicos en el documento. Si el valor de α es alto, se espera que el documento tenga una mayor diversidad de tópicos asignados. $\theta_d \sim \text{Dir}(\alpha)$
3. Se utiliza la distribución de Dirichlet para modelar la distribución del vocabulario en los tópicos. El parámetro η se utiliza para controlar la concentración de la distribución de palabras del tópico. Cuando se utiliza un valor alto de η , se obtiene un tema con una variedad más amplia de palabras. $\phi_k \sim \text{Dir}(\beta)$
4. Así, para cada documento d_i :
 - a) Se asigna un tópico a cada palabra del documento mediante una distribución θ_d de los tópicos presentes en el documento. $z_n \sim \text{Multinomial}(\theta_d)$
 - b) Se selecciona aleatoriamente una palabra para cada tópico en el documento utilizando una distribución multinomial. $w_n \sim p(w_n | z_n, \beta)$

Teniendo en cuenta las consideraciones descritas anteriormente, el objetivo del LDA es encontrar el número de tópicos que maximice la probabilidad de generar una mejor distribución de instancias/temas (θ) y temas/palabras (β). Para definir el parámetro asociado al número de tópicos, se requiere utilizar una métrica que permita evaluar la consistencia de las categorías generadas. De esta manera, el valor de coherencia de tópicos es una métrica cuantitativa que se utiliza para evaluar la similitud/distinción entre los diversos tópicos generados. Esta métrica se basa en la hipótesis de distribución lingüística que establece que las palabras en contextos similares tienden a tener significados similares.

La puntuación de coherencia mide la similitud semántica entre las palabras con las puntuaciones más altas en cada tópico, lo que permite evaluar la calidad y la interpretabilidad de los tópicos generados por un modelo de LDA. Una puntuación de coherencia más alta indica tópicos más comprensibles para los humanos.

La fórmula para calcular la puntuación de coherencia es:

$$coherence(T) = \sum_{(V_i, V_j)} score(V_i, V_j) \quad (3.4)$$

Donde (V_i, V_j) es una pareja de palabras del tópico. La función de puntuación se obtiene mediante el algoritmo Umass, que calcula la probabilidad de co-ocurrencia de las palabras.

$$score(V_i, V_j) = \frac{\log D(V_i, V_j) + \epsilon}{D(V_j)} \quad (3.5)$$

Donde $D(V_i, V_j)$ representa el número de documentos que contienen la pareja de palabras, $D(V_j)$ representa el número de documentos que contienen la palabra V_j , y ϵ es un factor de

suavizado que asegura que el resultado de la operación sea un número real. La evaluación de la coherencia de los tópicos es esencial para determinar la calidad del modelo y garantizar una interpretación adecuada de los resultados obtenidos en el análisis de tópicos.

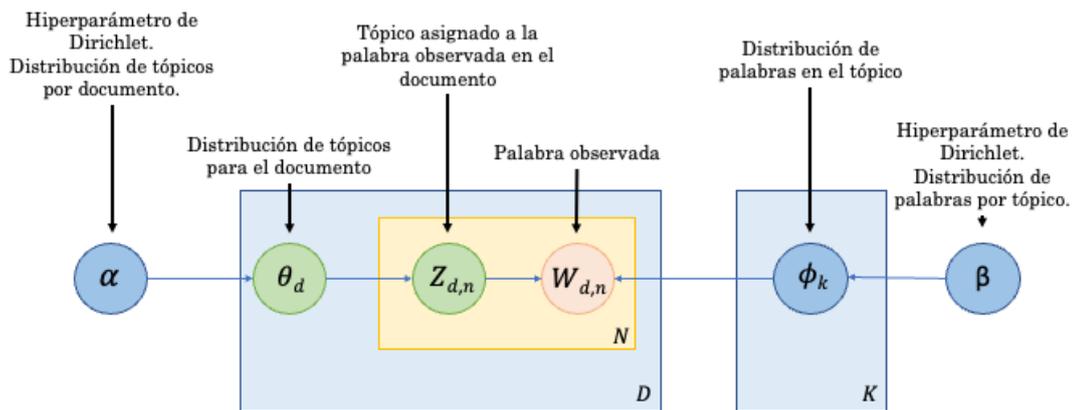


Figura 3.3: Algoritmo LDA
Elaboración Propia

Además de evaluar los valores de coherencia, se llevará a cabo un análisis de la distancia intertópica de los resultados obtenidos. Este cálculo implica el uso de la divergencia de Jensen-Shannon, combinada con un método de escalado multidimensional. Para visualizar de manera efectiva estos resultados, se empleará la biblioteca PyLDAvis, la cual permite representar los tópicos como círculos en un plano bidimensional. Los centros de los círculos se determinan en función de la distancia intertópica calculada, con el propósito de proyectar las distancias entre los tópicos en dos dimensiones.

En esta visualización, se codifica la prevalencia general de cada tópico utilizando el área de los círculos, de manera que los tópicos se ordenan en función de su prevalencia en un orden decreciente. Esto brinda una representación visual clara y intuitiva de la importancia relativa de cada tópico en el conjunto de datos analizados. Al codificar la prevalencia de esta manera, se facilita la identificación rápida de los tópicos más prominentes y permite al analista enfocar su atención en aquellos que poseen una mayor presencia en los textos analizados.

La visualización generada por PyLDAvis aporta una valiosa perspectiva en el análisis de tópicos, ya que permite una interpretación visual completa y una comprensión más profunda de las relaciones entre los tópicos. Al mapear los tópicos en un plano bidimensional, se revelan las distancias y similitudes entre ellos, lo que facilita la identificación de agrupaciones temáticas y la exploración de cómo se relacionan entre sí.

En este contexto, un resultado óptimo involucrará una distribución de tópicos (círculos) con muy pocas o ninguna intersección. Esto significa que los tópicos estarán claramente diferenciados y separados entre sí en el plano bidimensional. La ausencia de intersecciones entre los círculos indica una mayor coherencia y distinción entre los tópicos identificados en el análisis. Por el contrario, si los círculos se superponen o intersectan significativamente, puede indicar una mayor ambigüedad o superposición temática en los textos analizados.

3.5. Análisis de sentimientos

El análisis de sentimientos es una metodología ampliamente utilizada en PLN que facilita la detección de las opiniones, emociones y subjetividad expresadas en un texto. En este proyecto, se utiliza esta técnica para llevar a cabo la detección de polaridad en los tweets, lo que permite clasificarlos en sentimientos positivos, negativos o neutrales. Cada una de estas categorías se asocia a la emoción que caracteriza las publicaciones en Twitter. Los resultados obtenidos pueden ser esenciales para evaluar la reputación y la opinión que se tiene de la marca, y pueden ser utilizados para tomar decisiones que mejoren la percepción del público hacia la compañía. Para analizar el conjunto de tweets adquirido, en este caso, es necesario realizar un preprocesamiento diferente del texto. Aunque este proceso mantiene la eliminación de URLs, menciones, y hashtags, los signos de puntuación y las diferencias entre mayúsculas y minúsculas no se van a procesar, al ser fundamentales en la detección de los valores de polaridad.

Para llevar a cabo la clasificación de sentimientos, se utiliza la herramienta VADER. VADER es una herramienta de análisis de sentimientos binario que utiliza un enfoque de diccionario, y que contiene 7.518 unigramas, incluyendo signos de puntuación, palabras coloquiales, inicialismos, acrónimos y emoticonos. VADER mide tanto la intensidad del sentimiento como su polaridad (negativo o positivo) asignando una puntuación a cada palabra. Esta puntuación se mide en una escala de -4 a +4, donde -4 es lo más negativo y +4 lo más positivo. El punto medio 0 representa un sentimiento neutral. La puntuación total del análisis de sentimiento está comprendida entre -1 y 1, ya que se aplica una normalización. VADER es ampliamente utilizado y preferido como herramienta de análisis de sentimientos debido a sus heurísticas avanzadas (Hutto y Gilbert, 2014), caracterizadas por:

1. Puntuación: la utilización de signos de puntuación, como los signos de exclamación, incrementa la fuerza del sentimiento sin alterar el significado del mismo.
2. Mayúsculas: enfatizar una palabra relevante mediante el uso de mayúsculas aumenta la intensidad del sentimiento sin cambiar la dirección semántica.
3. Modificadores de grado y polaridad: Tanto los modificadores de grado como la conjunción "pero" juegan un papel crucial en la intensidad y polaridad de los sentimientos.

Los modificadores de grado pueden influir en la intensidad del sentimiento, ya sea aumentándola o disminuyéndola. Por otro lado, la conjunción "pero" señala un cambio en la polaridad de los sentimientos, con el sentimiento predominante en el texto que sigue a dicha conjunción.

4. Negación: cuando se analiza el trigramma anterior a un elemento léxico relacionado con una negación, se observa que en aproximadamente el 90 % de los casos, dicho elemento cambia la polaridad del texto.

Además de las 5 heurísticas incorporadas, VADER también puede ser personalizado con nuevas palabras y expresiones en el diccionario para mejorar la precisión del análisis de sentimientos. La herramienta ha demostrado ser efectiva en una variedad de dominios y contextos, incluyendo redes sociales, noticias y análisis de opiniones de productos (Elbagir y Yang, 2019; Pano y Kashef, 2020).

Capítulo 4

Resultados

En este capítulo, se exponen los resultados obtenidos mediante el procedimiento realizado siguiendo la metodología descrita en el Capítulo 3. El preprocesamiento del texto se ha analizado utilizando R, mientras que el modelado de tópicos y el análisis de sentimientos han sido implementados en Python. Los códigos usados para el análisis y el desarrollo de los modelos ha sido incluido en el repositorio de github: <https://github.com/lourdes5/TFGAnalytics>

4.1. Análisis exploratorio inicial

Previo a la ejecución del análisis de sentimientos y la modelización de tópicos, resulta crucial realizar una exploración preliminar de los datos con el fin de comprender su estructura y obtener una visión general de la distribución de palabras dentro del corpus objeto de análisis. Con este propósito, se ha recopilado un conjunto completo de tweets asociados a la marca Mercadona durante los últimos cinco años, abarcando el período de 2017 a 2022. Esta selección temporal busca examinar el efecto de la pandemia de COVID-19 en la percepción de los consumidores hacia Mercadona, considerando los tweets durante la pandemia, así como los anteriores y posteriores a ella. En total, se han obtenido 57,108 tweets en los que se menciona a Mercadona en España. Este análisis descriptivo proporcionará información valiosa para el análisis subsiguiente y permitirá elaborar conclusiones precisas sobre la percepción de los usuarios acerca de la marca en estudio.

En un primer paso, se llevó a cabo un análisis inicial para determinar la cantidad de tweets diarios que mencionaban la palabra clave “Mercadona” dentro del rango de tiempo seleccionado. El propósito de este análisis fue identificar posibles tendencias al alza o a la baja durante el período de estudio. Se buscó obtener información sobre la frecuencia de menciones relacionadas con “Mercadona” y evaluar si había variaciones significativas en la cantidad de tweets a lo largo del tiempo. Esta exploración inicial permitió obtener una visión general de la actividad en las redes sociales en relación con la palabra clave específica y establecer

una base para un análisis más detallado de los datos recopilados. Así, en la Figura 4.1, se presenta un análisis del nivel de actividad en la publicación de tweets durante el período de tiempo seleccionado (2017-2022). Aunque se aprecia una relativa estabilidad en el volumen de publicaciones durante la mayor parte del período analizado, existen dos periodos que se destacan de manera significativa. El primero de ellos abarca desde marzo hasta junio de 2020, periodo en el cual se observa un aumento considerable en la cantidad de tweets, siendo el 14 de marzo el día con el pico más destacado, con un total de 650 tweets registrados. Estas fechas coinciden precisamente con la imposición del estado de alarma, que se extendió desde el 14 de marzo hasta el 21 de junio. El incremento en la cantidad de tweets durante el 14 de marzo se debe a la controversia generada por el desabastecimiento en los supermercados y las largas colas ocasionadas por el pánico generado al comienzo del estado de alarma. El segundo pico se debe a una tendencia generada el 12 de marzo de 2022, a raíz de una publicación en la que se critican los precios elevados de Mercadona, un supermercado español, en comparación con los supermercados extranjeros. Estos eventos provocaron una mayor participación y discusión en las redes sociales sobre el tema relacionado con Mercadona.

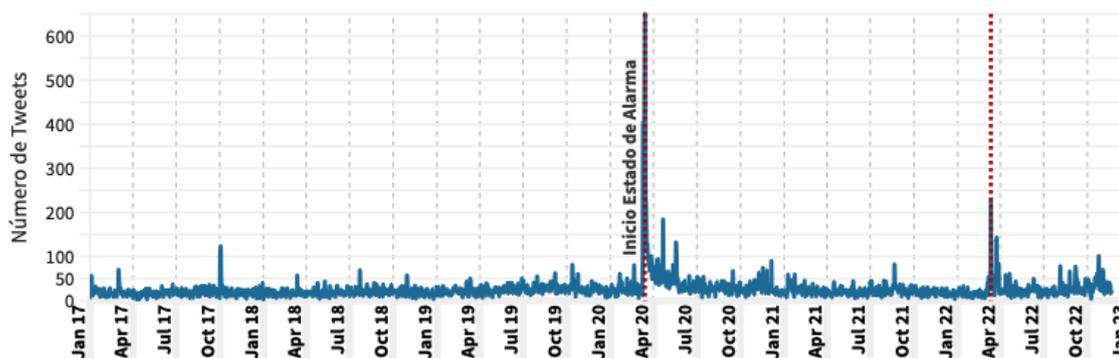


Figura 4.1: Evolución temporal del número de tweets
Elaboración Propia

A continuación, se llevó a cabo un análisis de la distribución de los retweets en los tweets recopilados. Como es posible observar en la Figura 4.2, la gran mayoría de los tweets tienen entre 0 y 9 retweets (99,07%). Por su parte, únicamente un 0,46% obtuvieron entre 10 y 19 retweets, y solamente un 0,11% de tweets superaron los 100 retweets. Estos resultados indican que un bajo porcentaje de las publicaciones logran capturar gran atención por parte de los usuarios y/o generar un impacto significativo en las interacciones en línea.

Después de examinar detalladamente la cantidad de tweets por fecha y el número de retweets, se decidió realizar un análisis de los idiomas utilizados en los tweets recopilados. Aunque los tweets fueron obtenidos con localización en España, era importante considerar la posibilidad de que se hubieran escrito en otros idiomas. Como era de esperar, el español predomina constituyendo el 88.5% de los tweets. El 11.5% restante está formado por catalán, inglés, tweets etiquetados como 'qme', que incluyen enlaces y menciones, y 'zxx', que solo

Distribución de retweets

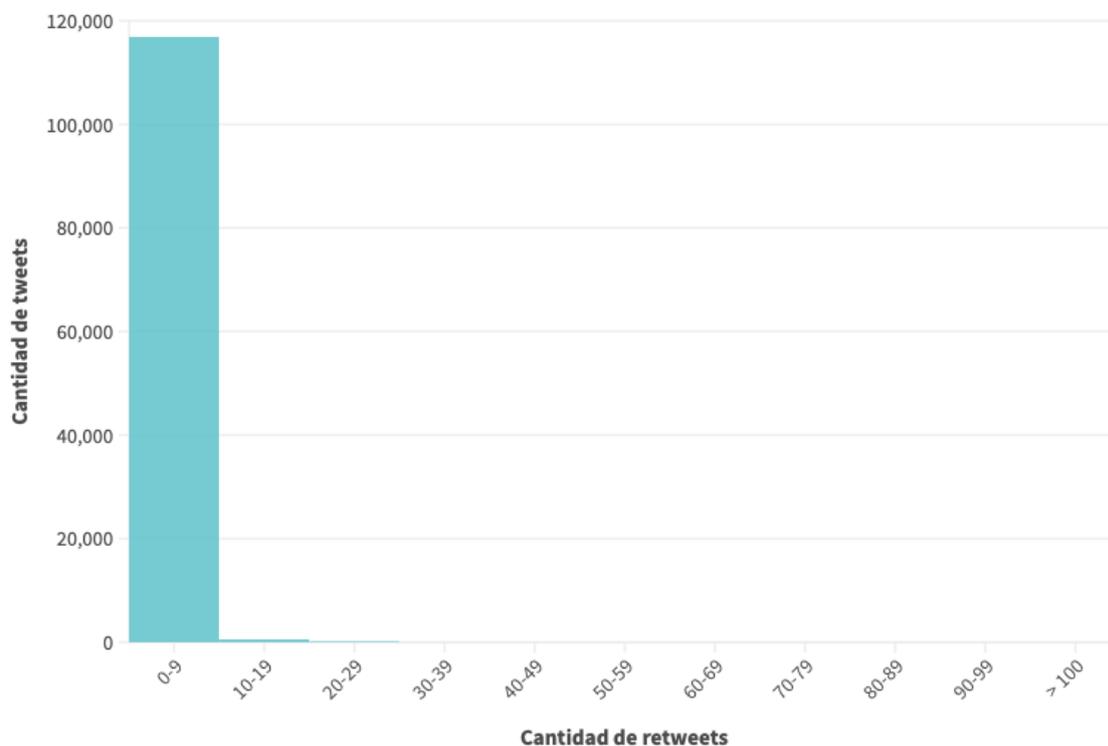


Figura 4.2: Distribución de tweets por número de retweets
Elaboración Propia

contienen un enlace. Finalmente, también se identificaron tweets en portugués. En la Figura 4.3, se muestran estos 5 idiomas más relevantes tras el español. Con el objetivo de mejorar la precisión del análisis, se ha centrado el análisis exclusivamente en los tweets en español. En consecuencia, se procederá a eliminar los tweets redactados en otros idiomas.

En este punto, resulta relevante el análisis de los hashtags utilizados en los tweets recopilados con el fin de obtener una comprensión más profunda de los temas abordados por los usuarios. Al examinar los hashtags más frecuentes, se observó que #mercadona y #Mercadonna eran los más predominantes. Esta situación es comprensible, dado que se trata de la palabra clave seleccionada para la recopilación de los datos. Por lo tanto, se ha tomado la decisión de eliminar estos hashtags, ya que no aportan información relevante adicional al análisis en cuestión. Así, la nube de palabras, ilustrada en la Figura 4.4, proporciona una visualización efectiva de los 20 hashtags más frecuentes y relevantes en estas publicaciones. En primer lugar, se destaca el hashtag #hacendado, el cual adquiere una prominencia significativa en la nube de palabras. Este hashtag representa la marca propia de Mercadona, la cual goza de una gran popularidad. Por otro lado, se observa la presencia de hashtags relacionados con la pandemia, como #QuedateEnCasa, #coronavirus, #COVID19 y #mascarilla. Estos términos se vinculan con las medidas y productos adoptados por Mercadona durante el período

Cantidad de tweets por idioma

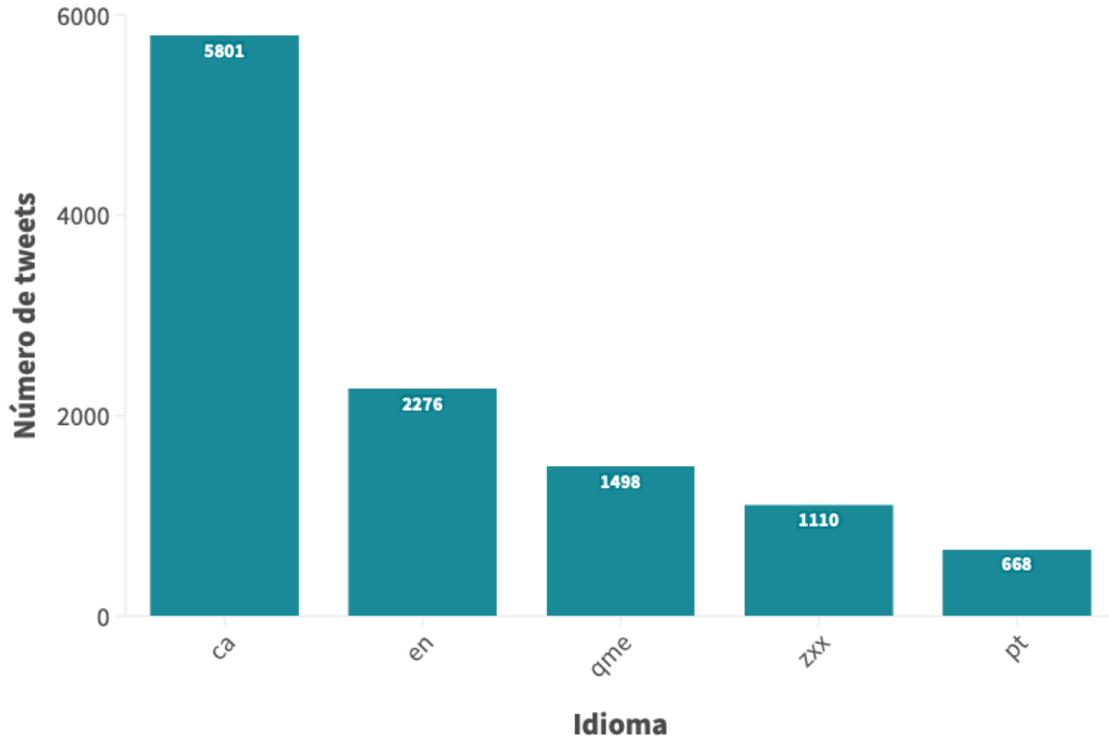


Figura 4.3: Distribución de tweets por idioma
Elaboración Propia

de COVID-19, destacando la alta demanda de mascarillas. Además, se menciona el hashtag #cerveza y otros relacionados. En conjunto, este análisis exhaustivo de los hashtags proporciona una visión más completa de los temas y preocupaciones que los usuarios expresan en relación a Mercadona.

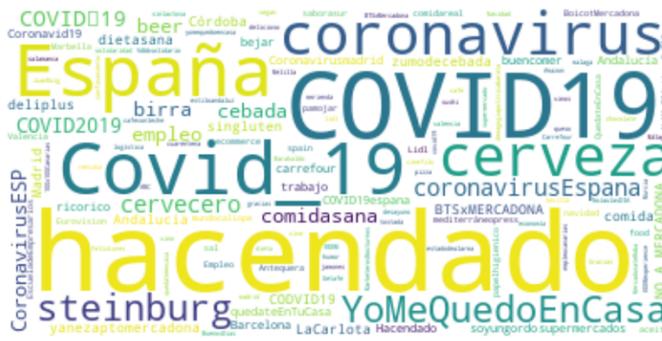


Figura 4.4: Wordcloud de HASHTAGS con mayor frecuencia
Elaboración Propia

Con el fin de obtener una descripción inicial de los sentimientos expresados por los usuarios hacia la marca, se ha analizado la frecuencia de los emoticonos presentes en los tweets recopilados, presentados en la Figura 4.5. En primer lugar, se destaca que los dos emoticonos

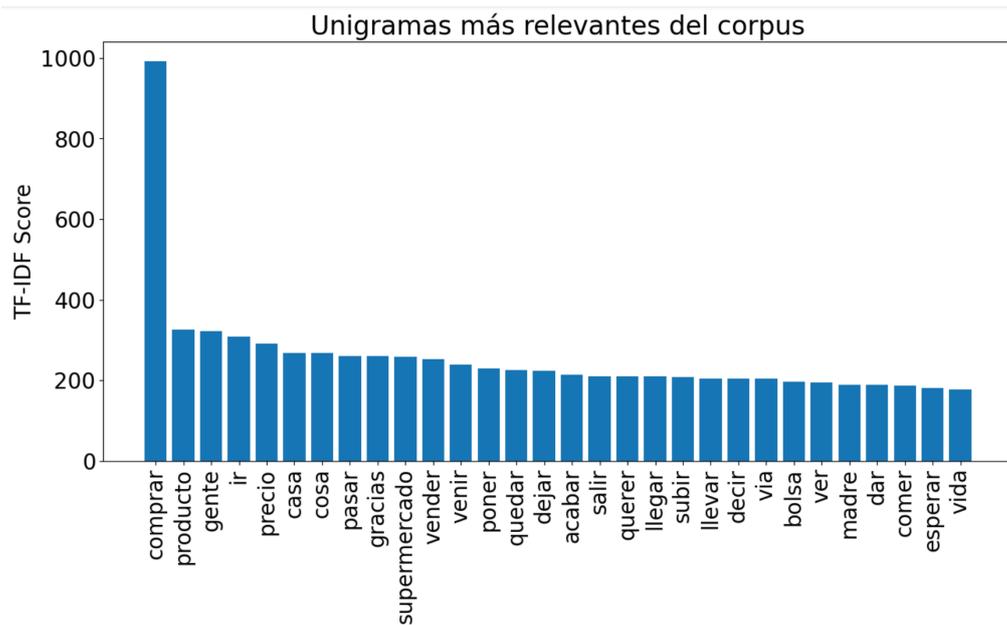


Figura 4.7: Unigramas más relevantes
Elaboración Propia

Además, el bigrama “subir precio” surge como un tema de discusión recurrente en los últimos tiempos, especialmente tras la pandemia, reflejando las preocupaciones sobre los aumentos de precios y su impacto en los consumidores. Por último, el bigrama “ir comprar” evidencia la intención y acción de los usuarios de acudir a Mercadona para realizar sus compras.

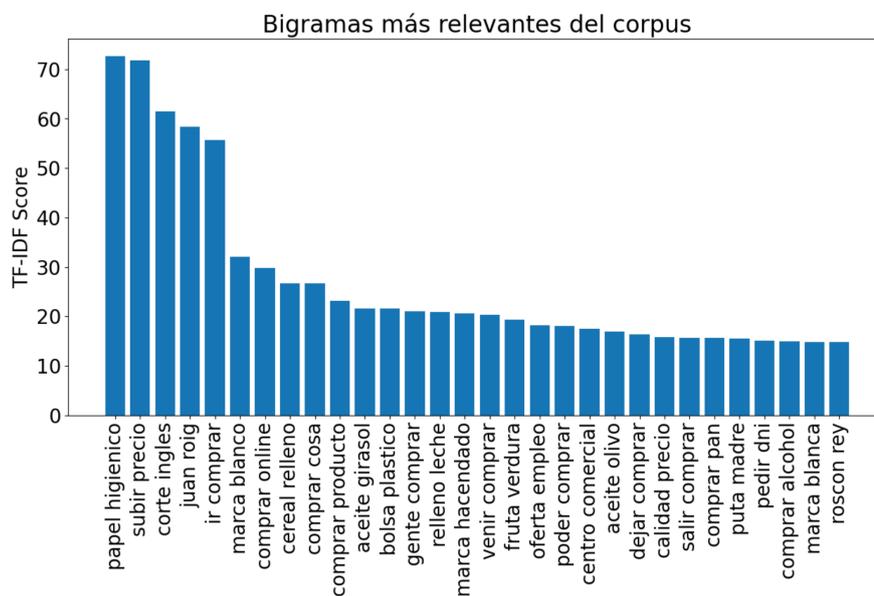


Figura 4.8: Bigramas más relevantes
Elaboración Propia

Por último, se ha realizado el cálculo del TF-IDF para los trigramas, que consisten en conjuntos de tres palabras, como se muestra en la Figura 4.9. Los trigramas con mayor puntua-

ción TF-IDF en los datos analizados proporcionan información relevante sobre los temas más destacados en las conversaciones sobre Mercadona en Twitter. Se observa que los usuarios mencionan productos específicos como “cereal relleno leche” y “cereal relleno chocolate”, lo que indica la importancia de estos productos en las interacciones de los usuarios. Además, se destaca la discusión en torno a la disponibilidad y compra de papel higiénico durante la pandemia, como se ve en los trigramas “comprar papel higiénico” y “quedar papel higiénico”, lo que refleja la relevancia de este producto durante un período de alta demanda. Por otro lado, algunos trigramas hacen referencia a productos particulares, como “helado huevo Kinder” y “empanada pollo seta”, lo que indica la popularidad de estos productos entre los usuarios de Mercadona. Además, se menciona al presidente de Mercadona, Juan Roig, en el trigma “juan roig presidente”, lo que demuestra el interés de los usuarios en el liderazgo y las decisiones de la compañía. Finalmente, los trigramas “pedir dni comprar” “relación calidad precio” sugieren la existencia de conversaciones relacionadas con la política de identificación requerida para realizar compras y la importancia percibida de la relación entre calidad y precio de los productos de Mercadona.

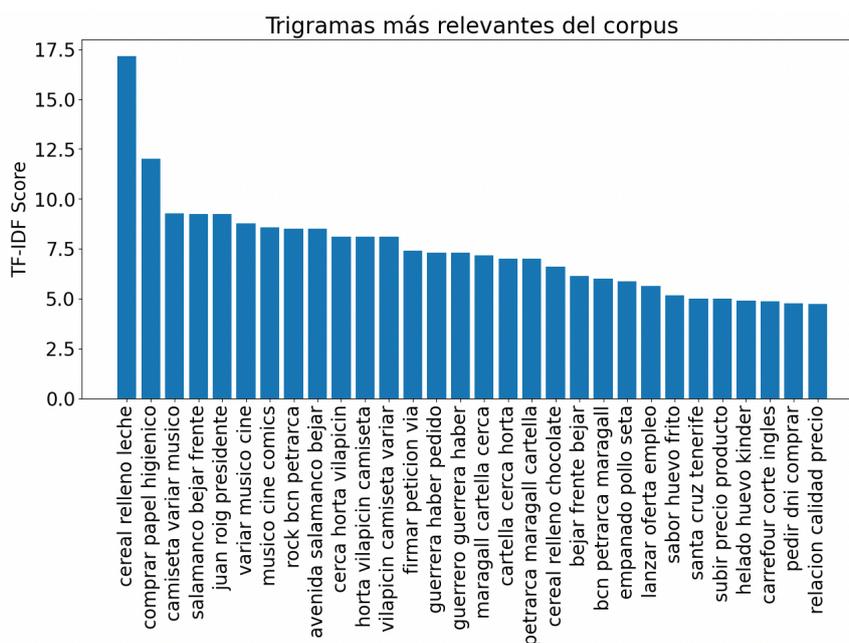


Figura 4.9: Trigramas más relevantes
Elaboración Propia

4.3. Modelado de tópicos

En esta sección, se presentan los resultados de modelar las categorías de discusión con base en los datos analizados. Como se detalló en la sección 3.4, uno de los primeros pasos es determinar el número apropiado de tópicos para el conjunto de datos analizado. Para lo-

garlo, se utilizó el índice de coherencia, el cual se calculó para distintos valores de número de tópicos. El objetivo fue identificar aquel valor que mostrara un pico significativo de crecimiento en el índice de coherencia, indicando así la presencia de tópicos más coherentes y distintos entre sí. En este caso, se observó que el valor $k = 3$ mostró un pico de crecimiento en el índice de coherencia, a partir del cual el índice comenzó a decrecer. Por lo tanto, se determinó que $k = 3$ era el número óptimo de tópicos para estos datos, ya que representaba una combinación adecuada de coherencia y diversidad temática.

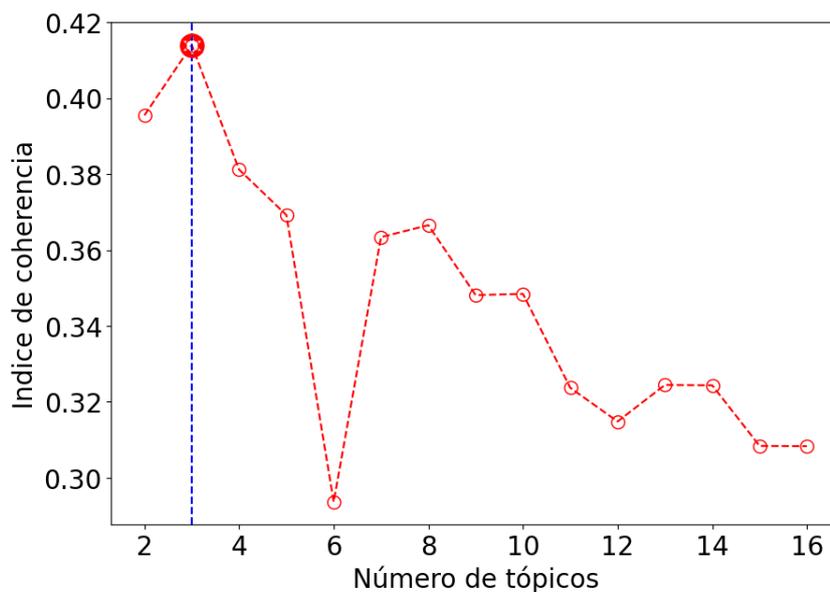


Figura 4.10: Evaluación de la coherencia según el número de tópicos.
Elaboración Propia

Aplicando un valor de $k = 3$, se lleva a cabo el análisis de la distancia inter-tópica mediante su representación visual, la cual se muestra en la Figura 4.11. El objetivo es encontrar un modelo que exhiba tanto un alto índice de coherencia como tópicos fácilmente interpretables, lo cual se refleja en un mapa donde los círculos que representan los tópicos no se superpongan en gran medida. A partir del análisis realizado, se observa que con $k = 3$ se obtiene la distancia intertópica mostrada en la Figura 4.10. En el mapa, se puede apreciar claramente la distinción entre los tres tópicos, lo que indica que el modelo logra generar tópicos discernibles entre sí.

Considerando el modelo generado, se presenta en la Tabla 4.1 una descripción detallada de los tópicos modelados, así como los 10 términos, bigramas y trigramas más importantes asociados a cada categoría. A partir de estos resultados, se puede apreciar que cada tópico se centra en aspectos específicos relacionados con Mercadona. Con base en las palabras claves modeladas, el primer tópico parece estar relacionado con la **experiencia de compra y la variedad de productos disponibles en Mercadona**. Se mencionan términos como “comprar”, “producto”, “precio” e “ir”, lo que sugiere discusiones sobre la adquisición de productos.

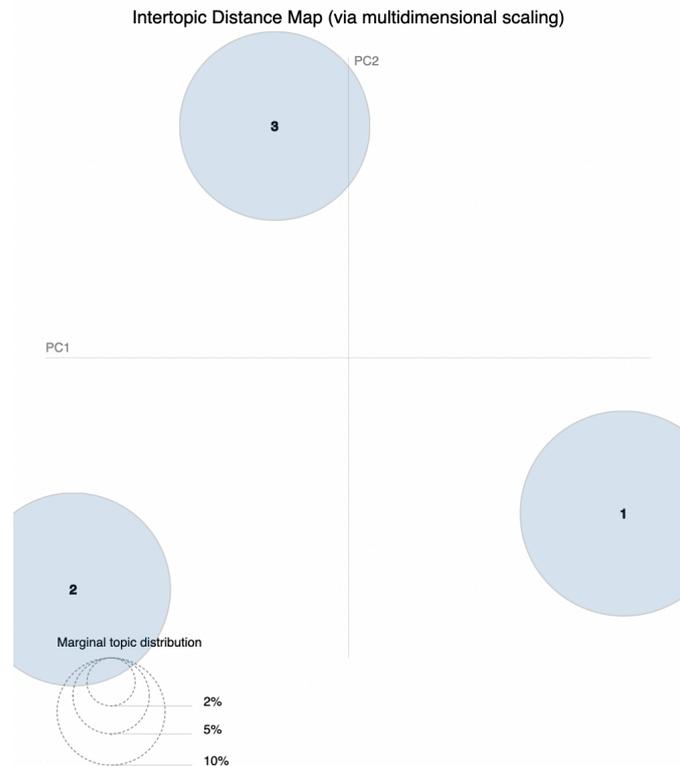


Figura 4.11: Mapa de la distancia intertópica
Elaboración Propia

Además, se hacen referencias a elementos específicos, como “cereal relleno”, “aceite girasol”, “marca blanca” y “marca Hacendado”, lo que indica la consideración de diferentes opciones y marcas disponibles en Mercadona. Además, se mencionan otros conceptos relacionados con la vida cotidiana y las actividades diarias, como “leche”, “casa”, “comer” y “salir”. Esto sugiere que los usuarios también comparten sus experiencias y hábitos de consumo relacionados con los productos de Mercadona en el contexto de su vida diaria. Finalmente, el tópico también aborda temas relacionados con los cambios de precios y las visitas frecuentes al supermercado, ya que se mencionan frases como “subir precio” e “ir comprar”. Esto puede indicar discusiones sobre la fluctuación de precios en los productos vendidos en Mercadona.

En el Tópico 2, se abordan temas relacionados con **Factores empresariales y de liderazgo en Mercadona**, así como con la comparación de la empresa con otros competidores del sector de supermercados. Se menciona a “Juan Roig” en varias combinaciones, como “roig”, “juan”, “juan roig presidente” y “lanzar oferta empleo”. Esto sugiere una asociación directa con el presidente de Mercadona, lo que indica que los usuarios están discutiendo temas relacionados con él y su papel en la empresa. Además, se mencionan otros elementos relacionados con la empresa y su funcionamiento, como “empresa”, “supermercado”, “centro logístico” y “puesto trabajo”. Estos conceptos indican que los usuarios están debatiendo sobre aspectos organizacionales, operativos y de empleo en Mercadona. Finalmente se hacen comparaciones con otros competidores del sector, como “Corte Inglés”, “Lidl” y

“Carrefour”, lo que sugiere que los usuarios están evaluando y contrastando a Mercadona con otras empresas de supermercados en términos de precios y otros aspectos.

Finalmente, el tópico 3 parece estar relacionado con las **interacciones entre los clientes y los empleados de Mercadona, así como algunos productos destacados**. Se mencionan palabras clave como “gracias”, “dejar”, “comprar”, “cliente”, “gente” y “trabajador”, lo que sugiere interacciones entre los clientes y el personal de Mercadona. Esto indica que los usuarios están discutiendo sobre su experiencia en la tienda, el trato recibido y la interacción con los empleados. Además, se mencionan productos específicos, como “papel higiénico”, “bolsa plástico”, “fruta verdura”, “helado huevo kinder” y “paquete papel higiénico”. Estos conceptos indican que los usuarios están hablando sobre la disponibilidad, la calidad o el precio de estos productos en particular. Finalmente, se mencionan situaciones específicas, como “cola entrar” y “subir precio”. Esto sugiere que los usuarios están compartiendo experiencias relacionadas con las colas para entrar al supermercado y discutiendo los aumentos de precios que han ocurrido.

Los resultados anteriores destacan claramente la importancia del tema de la subida de precios en los tres tópicos analizados. La presencia de la expresión “subir precio” en los bigramas y trigramas indica que este aspecto es relevante en las conversaciones relacionadas con Mercadona. Las variaciones en los precios pueden tener un impacto significativo en la percepción de los clientes sobre la relación calidad-precio de los productos ofrecidos.

Número de tópico	Tópico	Términos clave	Bigramas	Trigramas
1	Experiencia de compra y variedad de productos	comprar, producto, precio, ir, cosa, leche, casa, vida, comer, salir	subir precio, ir comprar, cereal relleno, marca blanca, aceite girasol, calidad precio, marca hacendado	cereal relleno leche, cereal relleno chocolate, empanada pollo seta, sabor huevo frito, pedir dni comprar, aceite oliva virgen, helado menta chocolate
2	Factores empresariales y de liderazgo en Mercadona	via, roig, esperar, trabajar, empresa, supermercado, poner, gente, juan, calle	corte ingles, juan roig, oferta empleo, comprar online, recogida alimento, centro logístico, lidl carrefour, puesto trabajo, subir precio	juan roig presidente, lanzar oferta empleo, carrefour corte ingles, centro logístico antequera
3	Interacciones con clientes y productos destacados	gracias, bolsa, dejar, comprar, cliente, gente, poner, carro, dar, trabajador	papel higiénico, bolsa plástico, comprar papel, ir comprar, cola entrar, fruta verdura, subir precio	comprar papel higiénico, quedar papel higiénico, rollo papel higiénico, helado huevo kinder, paquete papel higiénico

Tabla 4.1: Categorías, términos clave y n-gramas con mayor relevancia.

Por otro lado, se ha llevado a cabo un análisis detallado de la distribución de los tópicos en el conjunto de datos, y los principales resultados se presentan en la Figura 4.12. Este análisis proporciona una visión clara de la prevalencia de cada tópico en las conversaciones relacionadas con Mercadona. De acuerdo con la figura, se puede observar que el tópico 1,

denominado 'Experiencia de compra y variedad de productos', es el que cuenta con la mayor cantidad de reseñas asociadas. Esto sugiere que los usuarios de Twitter tienden a compartir sus experiencias de compra en Mercadona, así como la variedad de productos que encuentran en sus establecimientos. Esta temática se muestra como un punto focal de interés para los usuarios. Por su parte, el tópico 2, titulado 'Factores empresariales y de liderazgo en Mercadona', se posiciona en segundo lugar en cuanto a la cantidad de reseñas. Este tópico indica que los usuarios también se involucran en discusiones sobre diversos aspectos empresariales y el liderazgo dentro de la compañía. Es probable que se aborden temas como la gestión estratégica de la empresa, la visión del liderazgo y otros factores relacionados con el entorno empresarial de Mercadona.

Finalmente, aunque el tópico 3, denominado 'Interacciones con clientes y productos destacados' se ubica en el último lugar de la distribución, también ha generado una cantidad considerable de reseñas. Esto indica que los usuarios de Twitter se involucran activamente en conversaciones relacionadas con las interacciones entre Mercadona y sus clientes, así como los productos que destacan en el mercado.

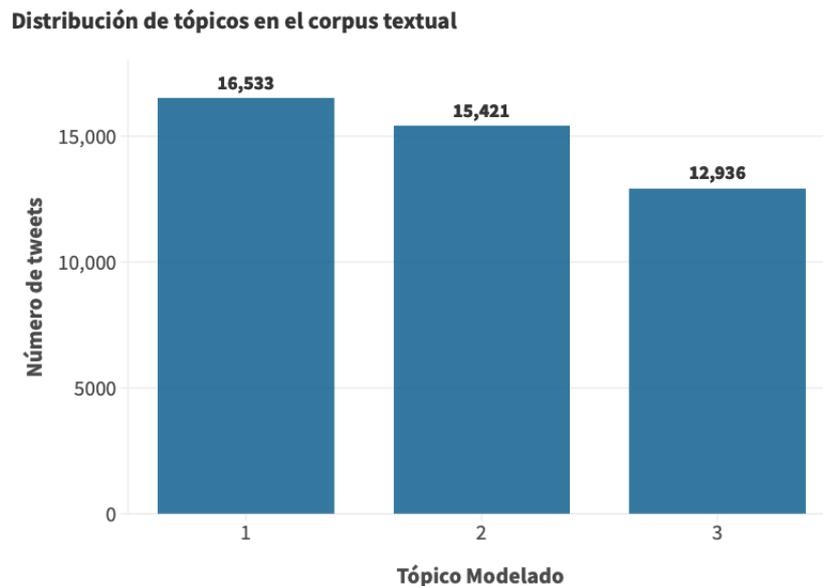


Figura 4.12: Frecuencia de las categorías en el corpus de texto.
Elaboración Propia

4.4. Análisis de sentimientos

Finalmente, se realizó un análisis del sentimiento de los tweets utilizando el algoritmo VADER. Los resultados de este análisis se presentan en la Figura 4.13, donde se muestra la distribución de las puntuaciones obtenidas después de eliminar los datos atípicos. En dicha figura, se puede observar que la mayoría de las opiniones expresadas por los usuarios tienden

a ser neutrales, aunque inclinadas ligeramente hacia un sentimiento negativo. La mediana de las puntuaciones obtenidas es de -0.07, donde 1 representa el puntaje máximo relacionado a una emoción positiva. Al examinar el primer cuartil, se puede apreciar que el 25 % de los datos presenta un puntaje por debajo de -0.3. Por otra parte, al excluir los valores atípicos del análisis, la puntuación más alta se sitúa por debajo de 0.5., lo que sugiere que no se registraron sentimientos positivos muy intensos en el conjunto de tweets analizados.

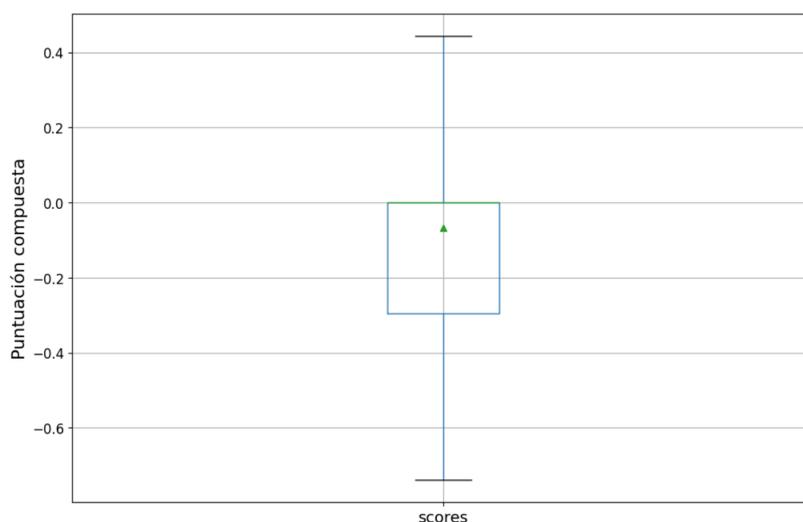


Figura 4.13: Representación gráfica de la distribución de la puntuación mediante un gráfico de cajas.

Elaboración Propia

Además, es importante destacar la distribución de las puntuaciones de sentimiento, tal como se muestra en la Figura 4.14. Este análisis nos proporciona información valiosa sobre la polaridad emocional de los tweets relacionados con Mercadona. En primer lugar, se puede observar que la mayoría de los tweets se sitúan en el rango de sentimiento neutro, con una puntuación de 0. Esto sugiere que gran parte de las conversaciones en Twitter relacionadas con Mercadona tienden a ser imparciales o no expresan una emoción claramente positiva o negativa. Sin embargo, al examinar más de cerca la distribución, se nota un mayor número de tweets con un sentimiento negativo en comparación con aquellos que tienen un sentimiento positivo. Específicamente, se observa un incremento significativo en la cantidad de tweets con una puntuación inferior a 0.25, lo que indica una tendencia hacia emociones negativas en estas publicaciones.

Además, es relevante analizar la evolución temporal del sentimiento expresado en los tweets a lo largo del periodo analizado, como se muestra en la Figura 4.15. Este análisis nos proporciona una visión más detallada de cómo las opiniones y emociones de los clientes han fluctuado en el tiempo. Como es posible observar en la figura, el sentimiento expresado por los clientes se mantiene alrededor del valor neutro de 0, pero inclinándose ligeramente hacia el lado negativo. Esta tendencia negativa se ha mantenido constante a lo largo del periodo

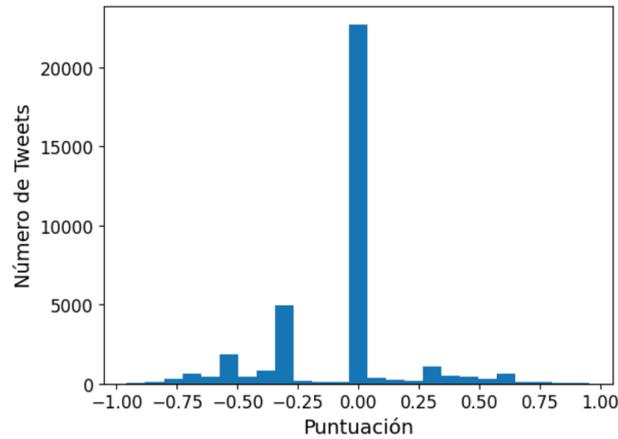


Figura 4.14: Distribución de la puntuación compuesta
Elaboración Propia

analizado, lo cual indica que existen preocupaciones o descontento persistente en las conversaciones relacionadas con Mercadona en Twitter. Es interesante destacar que se aprecia un pico negativo en las opiniones durante el mes de marzo de 2020. Este periodo coincide con la declaración del estado de alarma debido a la pandemia de la Covid-19, momento en el que se registraron numerosos comentarios y menciones sobre el desabastecimiento de los supermercados, siendo Mercadona uno de los más mencionados. Esta situación generó un impacto significativo en la percepción y experiencias de los clientes, lo cual se reflejó en un aumento de las opiniones negativas.

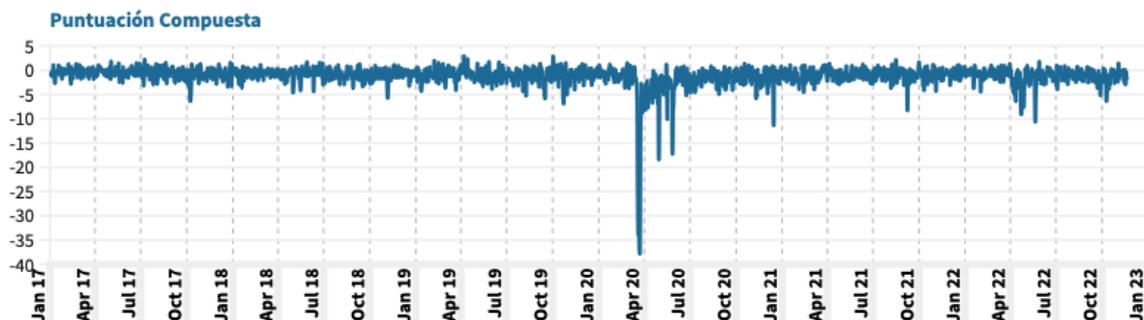


Figura 4.15: Evolución temporal del sentimiento general
Elaboración Propia

Siguiendo con el análisis, se centró en la variación temporal del sentimiento en relación a cada tópico identificado.(ver Figura 4.16). Comenzando con el tópico relacionado con la experiencia de compra y la variedad de productos, se puede observar que sigue una tendencia similar al sentimiento general del corpus de tweets. En este sentido, se evidencia el mismo pico negativo durante el mes de marzo de 2020, el cual ha sido mencionado anteriormente. Sin embargo, este evento tiene una estrecha relación con este tópico en particular. Durante ese periodo, los clientes expresaron una serie de quejas y preocupaciones relacionadas con

su experiencia de compra en Mercadona. Entre las principales inquietudes se encontraban las largas colas y la alta ocupación de los supermercados, lo que generaba incomodidad y malestar entre los compradores. Además, manifestaban su descontento por la falta de disponibilidad de algunos productos que se encontraban agotados, y no se contaba con suficientes existencias para satisfacer la demanda de los clientes. Estas circunstancias impactaron significativamente en el sentimiento expresado por los usuarios en relación a este tópico en particular.

Posteriormente, se realiza el análisis de la evolución del sentimiento asociado al tópico 2, el cual aborda los factores empresariales y de liderazgo en Mercadona (ver Figura 4.16). En relación a este tópico, se observa una disminución en la intensidad del sentimiento negativo en comparación con el tópico anterior. Si bien se presenta el mismo pico durante el año 2020, se puede apreciar que su intensidad es menor. Dentro de este contexto, las críticas se dirigieron principalmente hacia el director general de Mercadona y la empresa en sí, en relación a la gestión del crecimiento de la demanda tanto antes como durante el confinamiento. Los usuarios expresaron sus inquietudes y descontento en relación a cómo la empresa estaba lidiando con el aumento de la demanda y las dificultades de abastecimiento durante la pandemia. En la Tabla 4.1, al observar los trigramas y bigramas asociados a este tópico, se destacan menciones a empresas de la competencia, como Carrefour. Esto sugiere que los usuarios podrían haber comparado la gestión de Mercadona con otros supermercados en relación al problema del desabastecimiento y la capacidad de satisfacer la demanda de manera adecuada. No obstante, es importante destacar que este tópico no fue el centro de la discusión en la misma medida que el tópico anterior.

Por último, se realiza un análisis del tópico 3, el cual se relaciona con las interacciones entre los clientes y los productos destacados (ver Figura 4.16). En este tópico, se observa una mayor intensidad negativa durante el pico registrado en marzo de 2020, alcanzando una puntuación superior a -17.5. Además, durante ese período que abarca desde marzo hasta junio, se evidencia un rango pronunciadamente negativo adicional. Estos resultados pueden atribuirse a los problemas que surgieron para encontrar diversos productos de primera necesidad en los supermercados, como se refleja en los bigramas y trigramas de la Tabla 4.1. En dicha tabla, se destacan términos como “fruta verdura”, lo cual indica la falta de disponibilidad de estos productos esenciales. Asimismo, se menciona la expresión “cola entrar”, haciendo referencia a las largas filas que se formaban en los supermercados debido a las medidas de distanciamiento social implementadas. Asimismo, en repetidas ocasiones se hace mención al papel higiénico y su escasez, lo cual indica la preocupación y la atención que los clientes brindaron a este producto en particular durante ese período. Estos resultados respaldan la percepción de que la disponibilidad de productos básicos, como el papel higiénico, fue un tema de gran relevancia y discusión en las interacciones entre los clientes sobre la marca Mercadona.

Después de examinar cómo han evolucionado estos tópicos a lo largo del tiempo, se puede inferir que los clientes de Mercadona expresaron mayoritariamente un sentimiento negativo.

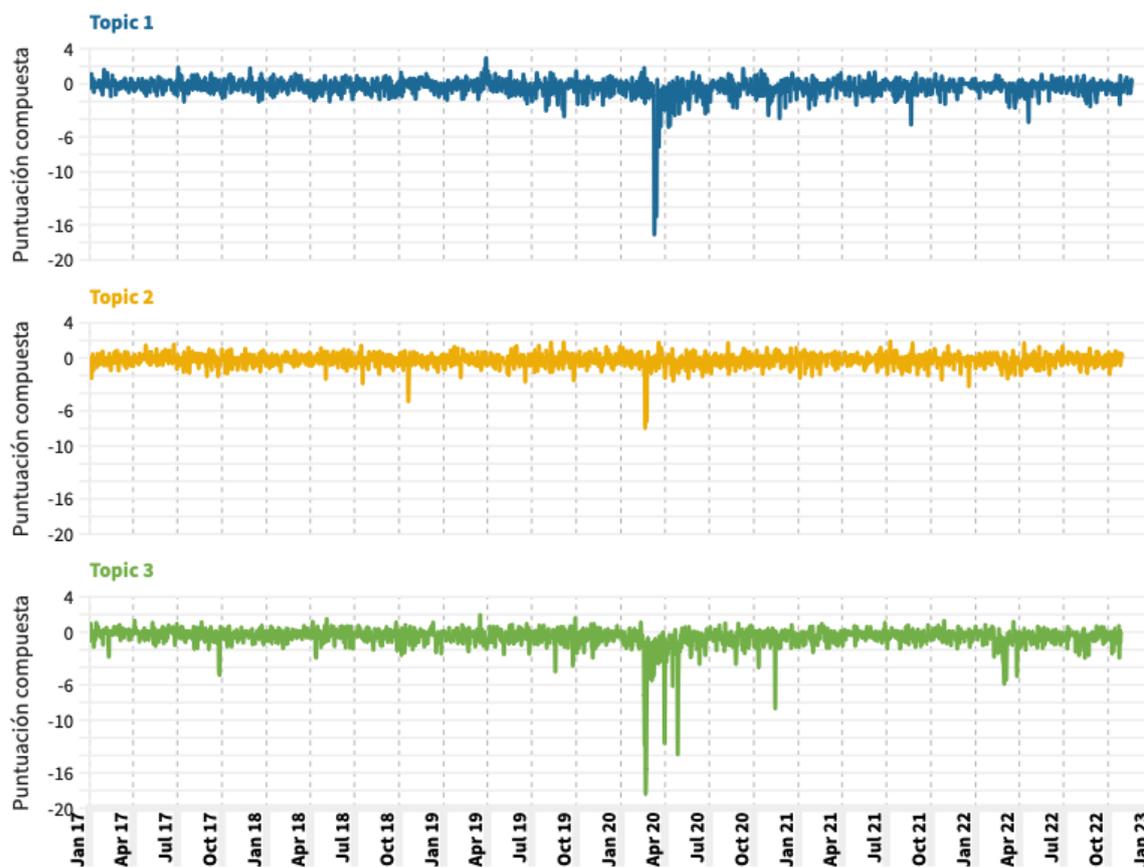


Figura 4.16: Evolución temporal del sentimiento, para cada uno de los tópicos
Elaboración Propia

El factor principal que afectó la opinión general estable que se tenía sobre Mercadona en España fue el estado de alarma, que abarcó el periodo entre marzo y junio de 2020. Durante este tiempo, el tópico que tuvo una mayor influencia en este suceso fue el tópico 3. El desabastecimiento y la falta de productos de primera necesidad fueron aspectos de gran importancia que causaron molestias e incertidumbre en los clientes. La escasez de productos básicos, como el papel higiénico, generó múltiples quejas y frustración. Además, las largas colas y la alta ocupación de los supermercados también fueron motivos de insatisfacción en la experiencia del usuario. Como consecuencia de este malestar generalizado, surgieron quejas dirigidas hacia los dirigentes de Mercadona y la empresa en general, aunque en menor medida en comparación con los aspectos relacionados con el desabastecimiento y la experiencia del cliente.

Capítulo 5

Conclusiones

En la actualidad, los clientes han adoptado las redes sociales como una plataforma fundamental para expresar sus opiniones sobre productos y servicios. Este cambio en el comportamiento de los consumidores ha generado una gran cantidad de datos valiosos, lo que ha llevado a las empresas a adaptar sus estrategias y procesos. En consecuencia, la identificación de los sentimientos de los clientes y la búsqueda de áreas de mejora para aumentar su satisfacción y lealtad se han convertido en una prioridad fundamental. No obstante, el creciente volumen de información generado plantea desafíos, requiriendo el empleo de métodos de análisis automático que permitan identificar los temas principales de discusión y comprender el sentimiento asociado a cada categoría. De esta manera, las empresas pueden aprovechar eficazmente los datos disponibles para tomar decisiones informadas y personalizar sus estrategias para satisfacer las necesidades cambiantes de sus clientes.

En el marco de este estudio de grado, se propuso como objetivo principal analizar la percepción de los ciudadanos en relación al servicio de venta de una de las empresas minoristas más destacadas en España, utilizando la plataforma de redes sociales Twitter. El propósito fundamental fue extraer información relevante sobre la forma en que los ciudadanos perciben la marca y el sentimiento subyacente en sus interacciones. Para lograr este objetivo, se realizó una revisión del estado actual de las técnicas de análisis avanzado de textos, especialmente diseñadas para el procesamiento de conjuntos de tweets. Durante este análisis, se identificaron las metodologías más apropiadas para desarrollar el trabajo, considerando tanto las características de los datos como los objetivos planteados en el estudio. En cuanto al modelado de tópicos, se observó que la técnica LDA es ampliamente utilizada debido a sus ventajas significativas, tales como la capacidad de procesar documentos de longitud variable y su implementación sencilla, sin requerir datos de entrenamiento extensos. Por lo tanto, se decidió emplear esta técnica para el análisis de los tópicos presentes en los tweets recopilados. En cuanto al análisis de sentimientos, se observó que las técnicas más ampliamente utilizadas incorporan reglas de decisión basadas en lexicones o diccionarios. En este trabajo de fin de grado, se decidió utilizar el modelo VADER, el cual se ajusta adecuadamente

a esta alternativa. El modelo VADER permite evaluar el sentimiento en textos cortos y en redes sociales, teniendo en cuenta las particularidades del lenguaje y la estructura de las oraciones. Por tanto, se consideró como una herramienta adecuada para analizar el sentimiento subyacente en los tweets relacionados con la empresa minorista estudiada.

Este estudio empleó una metodología de análisis que se inició con una selección de los tweets para su estudio, resultando en un total de 115,594 tweets que mencionaban la palabra “Mercadona”. Estos datos fueron sometidos a un proceso de limpieza y procesamiento para asegurar la calidad de la información analizada. A continuación, se llevó a cabo un análisis descriptivo preliminar con el objetivo de evaluar el contenido de los tweets y su evolución temporal a lo largo del periodo de estudio. Para identificar los términos más relevantes en el conjunto de tweets analizados, se utilizó la técnica de N-gramas mediante el cálculo del TF-IDF para cada unigrama, bigrama y trigramma. Esta técnica permitió obtener una visión descriptiva inicial de las palabras y combinaciones de palabras más frecuentes y significativas en los tweets relacionados con Mercadona.

Posteriormente, se implementó un modelo para identificar las categorías de discusión mediante la estimación del índice de coherencia para diferentes cantidades de tópicos. Este enfoque permitió determinar el número apropiado de tópicos que mejor representaba el corpus de tweets recopilados. Utilizando esta información, se generó un mapa de distancia inter-tópica para el valor del número de tópicos que mostraba el índice de coherencia más alto. Este mapa ofreció una visualización clara y organizada de la relación entre los distintos tópicos y su grado de similitud. Para concluir, se realizó un análisis adicional para examinar el sentimiento presente en los tweets. Se calculó una puntuación compuesta agregada en función del tiempo para el conjunto total de tweets y para cada uno de los tópicos individualmente identificados. Este análisis proporcionó una visión más detallada de los patrones temporales de los sentimientos expresados en los tweets analizados, ofreciendo información relevante sobre la evolución de la percepción y la opinión de los usuarios a lo largo del tiempo.

Durante el proceso de modelado de tópicos, se han identificado tres categorías principales que resultan altamente relevantes en el corpus de tweets relacionados con Mercadona. Estos tópicos se centran en la “Experiencia de compra y variedad de productos”, los “Factores empresariales y de liderazgo en Mercadona” y las “Interacciones con clientes y productos destacados”. En lo que respecta a la experiencia de compra y la variedad de productos, los usuarios expresan sus opiniones y comentarios sobre diversos aspectos, como la comodidad que experimentan al realizar sus compras en Mercadona, la calidad de los productos ofrecidos, la disponibilidad de una amplia gama de opciones y su grado general de satisfacción al interactuar con la empresa. Además, se ha observado que los precios son mencionados de manera recurrente, posiblemente debido a la controversia generada por las subidas de precios. El tópico relacionado con los factores empresariales y de liderazgo en Mercadona se enfoca en las menciones que se realizan acerca de aspectos como las estrategias comerciales implementadas por la empresa, la gestión empresarial en general y las decisiones tomadas

por los líderes de Mercadona, incluyendo al presidente Juan Roig. Estas discusiones exploran cómo dichas decisiones pueden influir en la percepción general de la marca. Por último, el tópico centrado en las interacciones de los clientes con productos destacados se concentra en las menciones y debates que se generan en torno a la relación entre los clientes y Mercadona, así como en los productos más destacados de la marca. Los usuarios comparten sus experiencias personales, recomiendan productos específicos, discuten sobre promociones especiales y expresan su nivel de satisfacción o insatisfacción, con respecto a la atención al cliente y la respuesta de Mercadona ante sus consultas o inquietudes.

A través del análisis de sentimientos, se obtiene información valiosa sobre la polaridad emocional de los tweets relacionados con Mercadona. Este análisis revela que la mayoría de las conversaciones en Twitter acerca de la empresa se encuentran en el rango de sentimiento neutro, sin una inclinación clara hacia lo positivo o lo negativo. No obstante, al examinar con más detalle la distribución de los sentimientos, se observa que hay más tweets con sentimientos negativos en comparación con los positivos. Específicamente, se destaca un incremento significativo en la cantidad de tweets que presentan una puntuación inferior a 0.25, lo que indica una tendencia hacia emociones negativas en dichas publicaciones.

Durante el análisis temporal del sentimiento general y de cada tópico, se observa un destacado pico negativo que se presenta desde marzo hasta junio de 2020, durante el período del confinamiento debido al estado de alarma. Durante esta etapa, los consumidores expresaron sus quejas y malestar a través de tweets, siendo la escasez de productos de primera necesidad el factor más influyente en el descontento de los clientes. En particular, el papel higiénico fue el producto más comentado y buscado por los consumidores. Además, la experiencia de compra de los clientes también desempeñó un papel relevante, ya que las largas colas y la aglomeración en los supermercados de Mercadona generaron una experiencia de compra desagradable. Aunque en menor medida, también se mencionaron aspectos relacionados con factores empresariales y de liderazgo en los tweets. En algunas ocasiones, se hizo referencia al presidente de Mercadona, Juan Roig. Durante esta controversia, su popularidad no fue positiva, aunque es importante destacar que este tópico no es el más relevante en las conversaciones sobre Mercadona en Twitter.

Después de analizar los resultados obtenidos, se pueden extraer importantes conclusiones. Es fundamental que la empresa tome en cuenta la experiencia durante la pandemia para mejorar su cadena de suministro y su relación con los proveedores. Si bien es cierto que la situación fue excepcional y provocó un pico de demanda muy alto, que resultó difícil de manejar, es necesario aprender de esta experiencia. Aunque los ciudadanos se dejaron llevar por el pánico y compraron cantidades exageradas, el inicio del estado de alarma era conocido, por lo que se podría haber previsto este aumento en la demanda y tomar medidas anticipadas. Es fundamental mejorar la planificación y la capacidad de adaptación ante situaciones inesperadas. Además, es necesario abordar el tema del aforo en los supermercados. Limitar la cantidad de personas que ingresan a los establecimientos por razones de seguridad y como-

didad es una medida que debería considerarse para evitar aglomeraciones y garantizar una experiencia de compra satisfactoria. En cuanto a las subidas de precios, es cierto que todas las cadenas de supermercados han ajustado sus precios, pero en el caso de Mercadona, donde la reputación estaba basada en ofrecer precios bajos, esta subida ha tenido un impacto significativo en su imagen en las redes sociales. Es importante encontrar un equilibrio entre ajustar los precios y mantener la confianza de los clientes. Por otro lado, es importante destacar los aspectos positivos de Mercadona, como su amplia variedad de productos y la calidad de su marca blanca “*Hacendado*”. Algunos productos, como los “*cereales rellenos de leche*”, han sido especialmente populares entre los usuarios de Twitter. Es fundamental potenciar esta marca blanca y seguir ofreciendo productos de alta calidad que sean bien recibidos en las redes sociales.

En el futuro, se podría enriquecer este estudio con un análisis exhaustivo de los productos virales destacados de Mercadona, considerando que varios de ellos han alcanzado una enorme popularidad en las redes sociales. Sería especialmente interesante investigar las características que contribuyen a la apreciación de estos productos por parte de los clientes, como su precio, calidad, originalidad y otros aspectos relevantes. Mediante este análisis, sería posible mejorar las estrategias publicitarias para nuevos productos, con el objetivo de generar mayor viralidad y aumentar los beneficios de ventas de la empresa.

Referencias

- Akundi, A., Tseng, B., Wu, J., Smith, E., Subbalakshmi, M., y Aguirre, F. (2018). Text mining to understand the influence of social media applications on smartphone supply chain. *Procedia Computer Science*, 140, 87–94.
- Bankmycell. (2022). *Total number of twitter users in the world (source: <https://www.bankmycell.com/blog/how-many-users-does-twitter-have>)*. (acceso Octubre 27, 2022) <https://www.bankmycell.com/blog/how-many-users-does-twitter-have>.
- Digital Marketing Institute. (2021). *20 surprising influencer marketing statistics*. (acceso Enero 13, 2023) <https://digitalmarketinginstitute.com/blog/20-influencer-marketing-statistics-that-will-surprise-you>.
- Elbagir, S., y Yang, J. (2019). Twitter sentiment analysis using natural language toolkit and vader sentiment. En *Proceedings of the international multiconference of engineers and computer scientists* (Vol. 122, p. 16).
- Ghiassi, M., Skinner, J., y Zimbra, D. (2013). Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with applications*, 40(16), 6266–6282.
- He, W., Zha, S., y Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, 33(3), 464-472. doi: <https://doi.org/10.1016/j.ijinfomgt.2013.01.001>
- 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).
- Ibrahim, N. F., y Wang, X. (2019). A text analytics approach for online retailing service improvement: Evidence from twitter. *Decision Support Systems*, 121, 37–50.
- Ibrahim, N. F., Wang, X., y Bourne, H. (2017). Exploring the effect of user engagement in online brand communities: Evidence from twitter. *Computers in Human Behavior*, 72, 321–338.
- Mishra, M. (2022). Customer experience: Extracting topics from tweets. *International Journal of Market Research*, 64(3), 334–353.
- Nagy, J., y Midha, A. (2014). The value of earned audiences: How social interactions amplify

- tv impact. *Journal of Advertising Research*, 54(4), 448–453. doi: 10.2501/JAR-54-4-448-453
- Pang, B., y Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2, 1–135.
- 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), 33.
- Pantano, E., Giglio, S., y Dennis, C. (2019). Making sense of consumers' tweets: Sentiment outcomes for fast fashion retailers through big data analytics. *International Journal of Retail & Distribution Management*, 47(9), 915–927.
- Park, N., Kee, K. F., y Valenzuela, S. (2009). Being immersed in social networking environment: Facebook groups, uses and gratifications, and social outcomes. *Cyberpsychology & behavior*, 12(6), 729–733.
- Raacke, J., y Bonds-Raacke, J. (2008). Myspace and facebook: Applying the uses and gratifications theory to exploring friend-networking sites. *Cyberpsychology & behavior*, 11(2), 169–174.
- Rasool, A., Tao, R., Marjan, K., y Naveed, T. (2019). Twitter sentiment analysis: a case study for apparel brands. En *Journal of physics: Conference series* (Vol. 1176, p. 022015).
- Roman, S., y Cuestas, P. (2008). 'the perceptions of consumers regarding online retailers' ethics and their relationship with consumers' general internet expertise and word of mouth: a preliminary analysis'. *Journal of Business Ethics*, 83, 641–656.
- Shen, C.-w., Chen, M., y Wang, C.-c. (2019). Analyzing the trend of o2o commerce by bilingual text mining on social media. *Computers in Human Behavior*, 101, 474–483.
- Statista. (2022a). *Cuota de mercado de las grandes cadenas de supermercados en españa en 2022*. (Enero, 2023) <https://es.statista.com/estadisticas/540894/porcentaje-de-ventas-de-los-grandes-supermercados-en-espana/>.
- Statista. (2022b). *Porcentaje de empresas que utilizaron las redes sociales en españa de 2014 a 2021*. (acceso Marzo, 2022) <https://es.statista.com/estadisticas/698728/porcentaje-de-empresas-que-hicieron-uso-de-las-redes-sociales-en-espana/>.
- Statista. (2022c). *Worldwide revenue of twitter from 2010 to 2021*. (acceso Octubre 27, 2022) <https://www.statista.com/statistics/204211/worldwide-twitter-revenue/>.
- Yu, Y., Duan, W., y Cao, Q. (2013). The impact of social and conventional media on firm equity value: A sentiment analysis approach. *Decision support systems*, 55(4), 919–926.