COMILLAS
UNIVERSIDAD PONTIFICIA

Facultad de Ciencias Económicas y Empresariales

El Tabaco en la Era Digital: Explorando la Percepción Ciudadana a Través del Análisis de Publicaciones

Autor: Sofía Mazarío Olivé
Director: Jenny Alexandra Cifuentes Quintero

MADRID| Abril 2024

Resumen

Enfrentar el consumo de tabaco como un problema de salud pública es un desafío significativo. Es fundamental que las autoridades sanitarias tengan acceso a información actualizada sobre las percepciones y actitudes ciudadanas respecto al tabaquismo para desarrollar y ajustar políticas de salud efectivas. Este acceso permite a las autoridades adaptarse mejor a las necesidades contemporáneas y optimizar las intervenciones en esta área crítica de salud pública.

En este contexto, la plataforma X (anteriormente conocida como Twitter) se ha mostrado como una fuente de datos valiosa para identificar tendencias y evaluar opiniones públicas en tiempo real, proporcionando una visión detallada y accesible de las actitudes de la ciudadanía. La efectividad de esta plataforma ha sido reconocida por investigadores de diversas partes del mundo que han empleado estrategias como el modelado de tópicos con LDA (Latent Dirichlet Allocation) y el análisis de sentimientos con VADER (Valence Aware Dictionary and Sentiment Reasoner) para recopilar y analizar datos. Este trabajo también adopta la herramienta de modelado de tópicos LDA y la técnica de análisis de sentimientos VADER para profundizar en las percepciones sobre el tabaco en X. La metodología propuesta en este trabajo abarca varios pasos, desde la selección del dataset y el tratamiento de los datos, hasta el análisis descriptivo de los términos más relevantes, la categorización de los tópicos de discusión y el análisis de sentimientos.

Como resultados, este trabajo de fin de grado ha identificado 4 tópicos principales: “Abandono de Productos de Nicotina”, “Opiniones sobre el Consumo de Nicotina y Alcohol”, “Dinámicas Sociales relacionadas con el Consumo de Tabaco” y “Control y Regulación del Uso de Tabaco y Marihuana”. Estos tópicos se presentan en orden descendente de discusión, siendo el abandono de la nicotina el más debatido y la regulación de su uso el menos mencionado. Por su parte, el análisis de sentimientos ha evidenciado que el tono predominante es neutral con un sesgo ligeramente positivo. Según las medias obtenidas en el diagrama de cada tópico, el tercero es el más negativo, mientras que el segundo es el más positivo, aunque ambos se mantienen más cercanos a una tono neutral.Las opiniones sobre el abandono de la nicotina varían de neutrales a ligeramente positivas, reflejando un debate sobre los beneficios de dejar ese hábito. En cuanto al consumo de tabaco y alcohol, existe una división cultural equilibrada entre la aceptación social y los riesgos para la salud. Las dinámicas sociales muestran una

aceptación moderada del tabaco, mientras que las actitudes hacia la regulación muestran una sociedad dividida entre la necesidad de control y la defensa de la libertad personal.

Abstract

Tackling tobacco use as a public health problem is a significant challenge. It is essential for health authorities to have access to up-to-date information on citizens' perceptions and attitudes towards smoking in order to develop and adjust effective health policies. This access allows authorities to better adapt to contemporary needs and to optimise interventions in this critical area of public health.

In this context, the X platform (formerly known as Twitter) has proven to be a valuable data source for identifying trends and assessing public opinion in real time, providing a detailed and accessible view of citizen attitudes. The effectiveness of this platform has been recognised by researchers from around the world who have employed strategies such as topic modelling with LDA (Latent Dirichlet Allocation) and sentiment analysis with VADER (Valence Aware Dictionary and Sentiment Reasoner) (Valence Aware Dictionary and Sentiment Reasoner) to collect and analyse data. This paper also adopts the topic modelling tool LDA and the sentiment analysis technique VADER to delve into the perceptions about tobacco in X. The methodology proposed in this paper covers several steps, from dataset selection and data processing, to descriptive analysis of the most relevant terms, categorisation of discussion topics and sentiment analysis.

As results, this thesis has identified 4 main topics: "Nicotine Product Abandonment", "Opinions on Nicotine and Alcohol Use", "Social Dynamics related to Tobacco Use" and "Control and Regulation of Tobacco and Marijuana Use". These topics are presented in descending order of discussion, with nicotine cessation being the most debated and regulation of nicotine use the least mentioned. The sentiment analysis showed that the predominant tone was neutral with a slightly positive bias. According to the averages obtained in the diagram for each topic, the third topic is the most negative, while the second is the most positive, although both remain closer to a neutral tone.opinions on nicotine cessation vary from neutral to slightly positive, reflecting a debate on the benefits of quitting. On tobacco and alcohol use, there is a balanced cultural divide between social acceptance and health risks. Social dynamics show a moderate acceptance of smoking, while attitudes towards regulation show a society divided between the need for control and the defence of personal freedom.

Índice general

Índice de figuras

Índice de tablas

Acrónimos

ADR	Reacciones Adversas a Medicamentos
AECC	Asociación Española Contra el Cáncer
API	Application Programming Interface
COP9	Conferencia de las Partes (COP) de la Convención Marco para el Control del Tabaco
IDF	Frecuencia Inversa de Documento)
INE	Instituto Nacional de Estadística
JMIR	Journal of Medical Internet Research
LDA	Latent Dirichlet Allocation
LSA	Análisis Semántico Latente
ML	Machine Learning
OMS	Organización Mundial de la Salud
PLN	Procesamiento de lenguaje natural
SEPAR	Sociedad Española de Neumología y Cirugía Torácica.
SVM	Support Vector Machine
TF	Frecuencia de Término
TF-IDF	Term Frequency-Inverse Document Frequency
t-SNE	Tt-distributed Stochastic Neighbor Embedding
UNCTAD	Conferencia de las Naciones Unidas sobre Comercio y Desarrollo
URL	Uniform Resource Locator
VADER	Valence Aware Dictionary and Sentiment Reasoner

Capítulo 1

Introducción

1.1. Motivación

La investigación sobre los efectos del tabaco se destaca como un tema de estudio de gran relevancia debido a su alcance en la salud pública y la economía a nivel mundial. Con aproximadamente 1.300 millones de personas fumadoras, de las cuales 820 millones son hombres y 480 millones mujeres, según datos de la Organización Mundial de la Salud (OMS), el tabaco se mantiene como un desafío persistente en la salud global. La Figura 1.1 muestra la prevalencia del consumo de tabaco desde el año 2000, proyectando que, a pesar de una disminución general, las tasas de fumadores en hombres se mantendrán por encima del 30 % para 2025, y las cifras en mujeres seguirán siendo significativas (>5 %) (OMS, 2023). A pesar de la tendencia decreciente, el consumo anual de cerca de 7 billones de cigarrillos evidencia el inmenso reto que aún representa el tabaquismo (X. Dai, Gakidou, y Lopez, 2022).

Este contexto se agrava al considerar las consecuencias mortales asociadas al tabaco. Anualmente, más de 8 millones de muertes son consecuencia de enfermedades relacionadas con el consumo de tabaco. La Asociación Española Contra el Cáncer (AECC) ha revelado datos alarmantes: el tabaco es responsable de ocho de cada diez casos de cáncer pulmonar a nivel mundial. De hecho, como puede verse en la Figura 1.2, el 82 % de los casos de cáncer de pulmón pueden atribuirse al tabaquismo, con los fumadores enfrentando un riesgo de veinte a veinticinco veces mayor en comparación con los no fumadores. Además, el peligro asociado al tabaco trasciende al cáncer de pulmón, ya que su uso incrementa el riesgo de desarrollar cáncer en diversas partes del cuerpo, incluyendo el 84 % de los cánceres de laringe y hasta un 50 % de los cánceres de vejiga y orofaringe. Estos datos refuerzan la necesidad de redoblar los esfuerzos de prevención y control para hacer frente a esta problemática de salud pública (OMS, 2023).

En España, el tabaquismo continúa siendo un asunto crítico de salud pública, con datos del Instituto Nacional de Estadística indicando que el 18 % de los adultos españoles eran fumadores en 2022 (INE, 2022). Esta cifra adquiere una dimensión más grave ante el informe

de la Sociedad Española de Neumología y Cirugía Torácica, que estima que el tabaco causa alrededor de 60,000 defunciones al año en España, lo que representa unas 160 muertes cada día (SEPAR, 2023). La enfermedad más letal asociada al tabaco, el cáncer de pulmón, presenta más de 30,000 nuevos casos anuales, con la mayoría de ellos diagnosticados en personas que fuman o que han dejado de fumar recientemente. Estos individuos tienen un riesgo hasta veinte veces mayor de desarrollar cáncer de pulmón en comparación con los no fumadores, lo que resalta el nexo entre el tabaquismo y esta enfermedad mortal (AECC, 2023). Estadísticas como estas evidencian la necesidad de políticas de prevención y control más efectivas para mitigar las consecuencias del consumo de tabaco en España (del corazón, 2022)

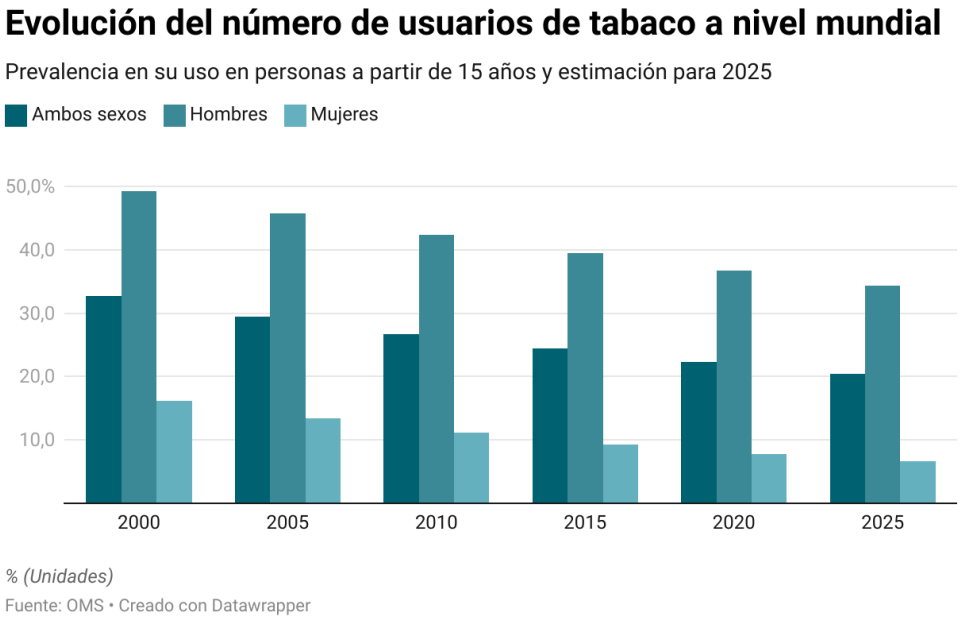

Figura 1.1: Evolución del número de usuarios de tabaco a nivel mundial Fuente de datos: OMS (OMS, 2022). Elaboración propia

Además, en el ámbito económico, la industria del tabaco figura entre las más lucrativas a nivel mundial. De hecho, en el año 2000 la industria tabaquera mundial facturó casi 400.000 millones de dólares anuales, mostrando la complejidad de la relación entre el bienestar público y los intereses económicos (Yach y Bettcher, 2000). A pesar de los crecientes esfuerzos por parte de diversos organismos de salud pública para reducir el consumo de tabaco, debido a sus efectos perjudiciales en la salud, esta industria sigue destacándose como un fuerte pilar económico. En el caso de España, que cuenta con una arraigada historia de consumo de tabaco, los datos revelan una tendencia a la baja desde 2012 (Ver figura 1.3). No obstante, el número de cigarrillos consumidos en 2020 sigue siendo elevado (más de 19000 millones de cigarrillos consumidos), destacando el reto continuo que el tabaquismo supone para la salud pública española y el estrés que ejerce sobre su sistema sanitario.

En respuesta a las preocupaciones de salud pública, el gobierno español ha puesto en

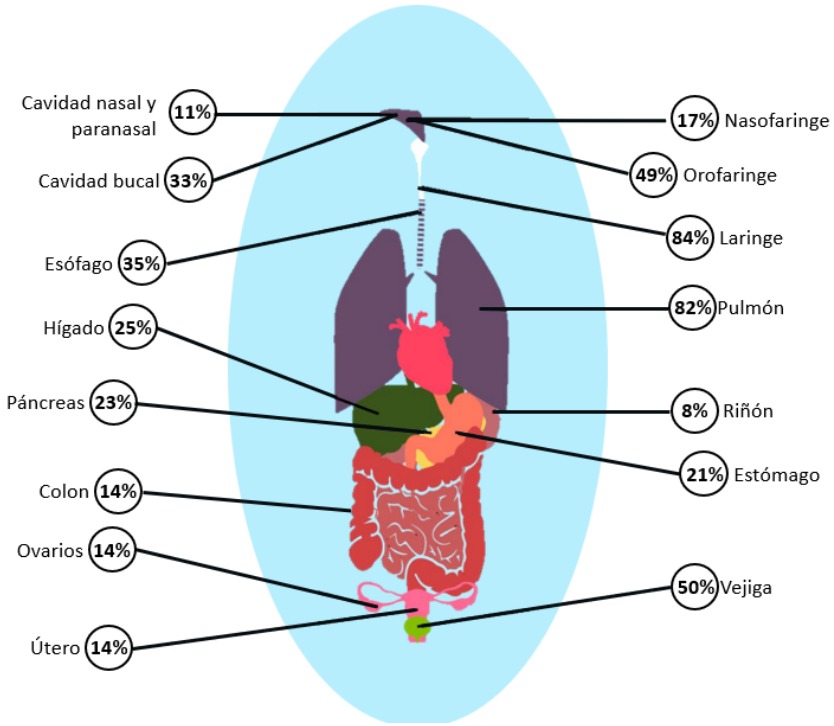

Figura 1.2: Tipos de cáncer relacionados con el tabaco Fuente de datos:AECC (AECC, 2022). Elaboración propia

marcha políticas antitabaco decisivas. Estas medidas incluyen la prohibición de fumar en espacios públicos cerrados, restricciones publicitarias para productos de tabaco, y el incremento de impuestos para disuadir el consumo. Sin embargo, y a pesar de la rigurosidad de estas políticas, la industria tabacalera mantiene su importancia económica en el país. El INE reportó que en 2022 el valor de producción de la industria tabacalera española se situó en 11,3 mil millones de euros, lo cual evidencia la resistencia de este sector frente a las iniciativas de salud pública (de Hacienda y Función Pública, 2022).

Ante este panorama, se hace necesario profundizar en el estudio y comprensión de las nuevas tendencias de consumo de tabaco. La creciente popularidad de los productos de tabaco calentado y cigarrillos electrónicos, especialmente entre los jóvenes, y las estrategias de marketing que emplea la industria, plantean desafíos significativos en términos de salud pública y prevención del tabaquismo (H. Dai, 2020). Además, la persistente desinformación sobre la seguridad de estos productos emergentes destaca la urgencia de realizar investigaciones objetivas. Estos estudios no solo ayudarían a desmitificar las afirmaciones de la industria, sino que también contribuirían a la formulación de políticas públicas más efectivas para combatir los riesgos asociados al consumo de tabaco en sus diversas formas.

En este contexto, la creciente prevalencia de las redes sociales en la vida cotidiana se muestra como un recurso valioso para abordar estos desafíos. Estas plataformas ofrecen un flujo constante y dinámico de datos generados por los usuarios, reflejando opiniones, expe-

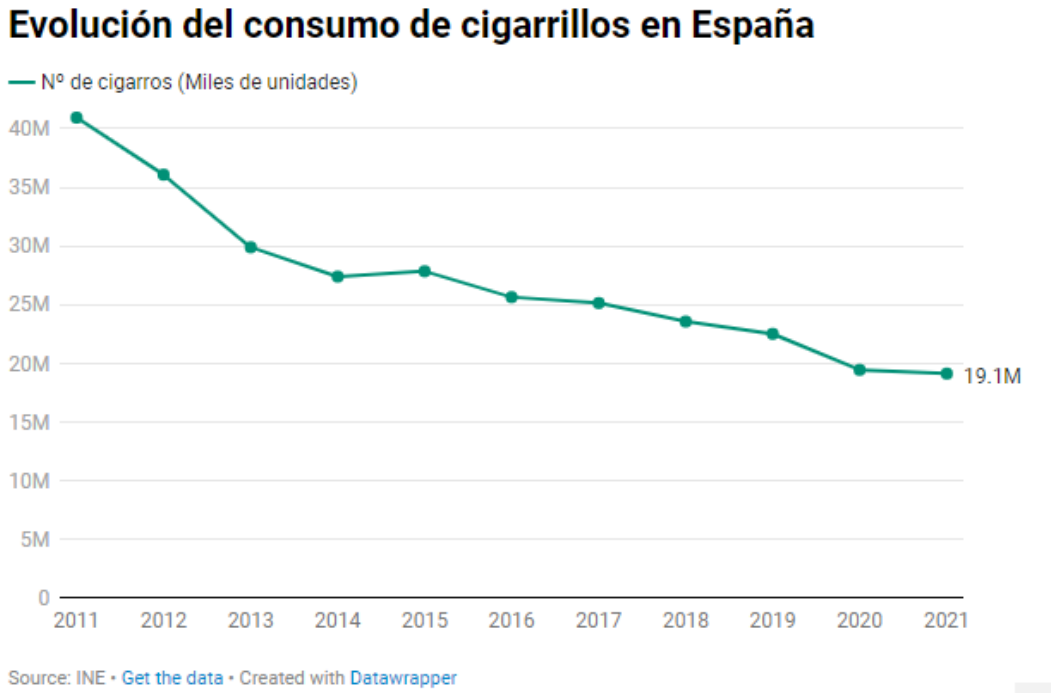

Figura 1.3: Evolución del consumo de cigarrillos en España Fuente de datos: INE (INE, 2021). Elaboración propia

riencias y debates en tiempo real. Esta riqueza de información convierte a las redes sociales en un recurso inestimable para investigadores y responsables de políticas públicas, permitiéndoles capturar el pulso de la sociedad y obtener perspectivas más profundas sobre los comportamientos relacionados con el consumo de tabaco, facilitando así el desarrollo de estrategias de intervención más informadas.

En particular, X (antes Twitter), se ha establecido como una herramienta clave en el ámbito de las redes sociales, jugando un papel esencial en la difusión de información y en la formación de la opinión pública. Con una base de 353,9 millones de usuarios mensuales activos en 2023, X se convierte en un recurso de gran relevancia para el análisis de tendencias sociales y opiniones (Statista, 2023). Esta plataforma no solo permite que los individuos compartan sus perspectivas y se relacionen con otros usuarios, sino que también funciona como un termómetro de los temas y discusiones predominantes en la sociedad. La habilidad de esta plataforma para capturar y reflejar las conversaciones en tiempo real sobre temas tan diversos proporciona un escenario de gran interés para el análisis de las percepciones y actitudes ciudadanas en un entorno digital en constante cambio (Bian et al., 2016).

En particular, en la discusión asociada al tabaco, un factor crítico que ha influido en los comentarios en las redes sociales es la polarización de opiniones. Esta tendencia a la polarización se manifiesta claramente en los debates sobre el tabaco: algunos usuarios promueven activamente su consumo o minimizan sus riesgos, mientras que otros defienden vehementemente la lucha contra el tabaquismo y sus peligros. Los estudios han demostrado que la

exposición a contenido relacionado con el tabaco en las redes sociales puede aumentar la probabilidad de consumo de tabaco, especialmente entre adolescentes y jóvenes adultos. En particular, los estudios propuestos por (Lee, 2016), (Hamdy y Gomaa, 2012) y (Kubin y von Sikorski, 2021) evidencian que las redes sociales pueden convertirse en escenarios de confrontación entre visiones opuestas, complicando así la formación de un juicio informado por parte de los usuarios y potencialmente obstaculizando las iniciativas de salud pública. En respuesta a esta situación, se han realizado investigaciones orientadas a analizar y a entender de manera más profunda la dinámica de las conversaciones sobre el tabaco en las redes sociales y su impacto en la percepción pública.

Dentro de esta línea investigativa se enmarca el presente Trabajo de Fin de Grado, el cual se enfoca en explorar la percepción y opinión de la población respecto al consumo de tabaco en la red social X. Para lograr este objetivo, se aplicarán técnicas avanzadas de minería de textos, permitiendo un análisis detallado de las publicaciones en esta plataforma. Este enfoque metodológico permitirá no solo identificar las tendencias predominantes y las actitudes hacia el tabaquismo, sino también comprender la dinámica de las discusiones en el área de análisis.

1.2. Objetivos

El objetivo principal de este Trabajo de Fin de Grado consiste en evaluar la percepción y opinión de la población en cuanto al consumo de tabaco, utilizando técnicas avanzadas de procesamiento del lenguaje natural y analítica de textos en publicaciones de la red social X (anteriormente conocido como Twitter). Este estudio tiene como propósito identificar y clasificar con precisión las temáticas de discusión más destacados sobre el tabaco en esta red social, además de valorar el tono emocional presente en dichas discusiones. Para alcanzar este fin, se han establecido los siguientes objetivos específicos:

- Identificar y describir la relevancia del análisis de redes sociales como herramienta efectiva para comprender la percepción ciudadana sobre fenómenos sociales, con un enfoque particular en la opinión y actitudes hacia el consumo de tabaco.
- Realizar una revisión de la literatura académica y científica para explorar las técnicas de minería de textos más pertinentes y efectivas utilizadas en el análisis de datos provenientes de redes sociales, especialmente enfocado en la percepción ciudadana sobre temas de salud como el tabaco.
- Identificar las diversas temáticas de debate que surgen en las redes sociales, específicamente en X, relacionadas con las percepciones en torno al tabaco.
- Evaluar la polaridad y tono emocional de las publicaciones en X sobre el tabaco, permitiendo entender la actitud general de la población hacia este fenómeno social.

1.3. Estructura del documento

La estructura de este trabajo se compone de seis secciones clave. En primer lugar, el presente capítulo abarca la justificación de la investigación, los objetivos establecidos y proporciona una visión general de cómo se organizará el contenido de la memoria. A continuación, en el Capítulo 2 se realiza un análisis detallado de la investigación previa relacionada con el tema. El capítulo 3, por su parte, se desglosa en subsecciones que detallan el proceso metodológico de investigación, incluyendo la pre-selección y extracción de datos, el pre-procesamiento de la información, el modelado de tópicos y el análisis de sentimientos. Posteriormente, en los Capítulos 4 y 5, se resumen los principales resultados y se discuten las implicaciones más relevantes de este estudio.

Capítulo 2

Analizando las percepciones y comportamientos sociales hacia el tabaco en X: Revisión de la Literatura

El análisis del consumo de tabaco adquiere una relevancia crítica en el contexto de la salud pública global. Este producto, profundamente arraigado en las sociedades modernas, no solo representa una preocupación importante por sus efectos adversos en la salud, sino también por su impacto en la economía y la cultura. En este contexto, comprender las percepciones y actitudes hacia el tabaco es fundamental para formular políticas efectivas de salud pública y estrategias de intervención. Teniendo en cuenta este objetivo, las redes sociales emergen como herramientas eficaces para la extracción de información valiosa. Plataformas como X (anteriormente conocida como Twitter) ofrecen una ventana única a las opiniones y comportamientos de millones de usuarios en tiempo real, reflejando un amplio espectro de perspectivas y discusiones sobre el tabaquismo. Este entorno digital proporciona una oportunidad para analizar y comprender la dinámica social en torno al tabaco, permitiendo a investigadores y responsables de políticas públicas acceder a datos ricos y contextualizados que anteriormente eran inaccesibles.

De esta manera, numerosos estudios han avanzado en esta dirección de investigación, concentrándose en la importancia de las plataformas digitales para el análisis de las percepciones públicas respecto al tabaquismo. Por ejemplo, una investigación preliminar propuesta por (Myslín, Zhu, Chapman, Conway, et al., 2013) implementó algoritmos de Machine Learning (ML), incluidos Naive Bayes, k-nearest neighbors y máquinas de soporte vectorial SVM (Support Vector Machines), para examinar 7,362 publicaciones en redes sociales vinculadas al tabaco. Este enfoque permitió no solo la identificación de publicaciones relevantes sino también la evaluación de los sentimientos subyacentes. La metodología se distinguió por su aplicación de clasificadores automatizados, alcanzando una precisión del 77,5 % en la identificación de publicaciones relevantes y un 70,5 % en la determinación de sentimientos. La

clasificación del análisis de sentimiento evaluó el tono de los textos como positivo, negativo o neutro, revelando que el 46 % de los mensajes eran positivos, el 32 % negativos y el 22 % neutros. Estos resultados demostraron la viabilidad de supervisar y analizar las percepciones del tabaco en tiempo real, abriendo camino para futuras investigaciones en este campo.

Siguiendo esta orientación investigativa, investigadores de la Universidad de Oxford (Cobb, Mays, y Graham, 2013) realizaron un estudio para analizar el impacto de los mensajes en foros en línea relacionados con el abandono del tabaco, tratamientos asociados y prevención de recaídas, poniendo especial énfasis en la vareniclina. Este estudio se enfocó en la elección inicial de medicamentos de los participantes, utilizando el software Salience Engine 4.1 para analizar el tono de los comentarios sobre la vareniclina.

A través de un análisis de regresión logística, los investigadores evaluaron el impacto de los mensajes positivos en foros, en la elección de la vareniclina como tratamiento. Los resultados indicaron que la exposición a comentarios favorables duplicaba la probabilidad de optar por la vareniclina y aumentaba en 2.5 veces la tendencia a continuar su uso. Este vínculo entre la percepción positiva y la decisión de tratamiento resalta la gran influencia de las conversaciones en línea en las decisiones asociadas a la salud.

En un enfoque complementario, un estudio realizado en la Universidad de Vermont por (Clark et al., 2014) ha expandido el campo de estudio en salud pública mediante la aplicación de análisis espacial y temporal. Esta investigación se centró en relacionar la actividad en X con el consumo de tabaco y cigarrillos electrónicos, adoptando un enfoque integral que no solo examinó la frecuencia de las menciones sobre el tabaco sino que también investigó las variaciones de estas discusiones a lo largo del tiempo y en distintas regiones geográficas. Utilizando datos geoespaciales, se mapearon las ubicaciones y momentos específicos de las conversaciones sobre el tabaco, y a través de la hedonometría, se evaluó el tono, ya sea positivo o negativo, de los mensajes publicados. Un resultado notable fue la identificación de una correlación positiva significativa entre la concentración de publicaciones sobre el tabaco y una percepción positiva del mismo en ciertas regiones, lo que permite comprender las variaciones en las actitudes regionales hacia el tabaco, influenciadas por el intercambio de opiniones en las redes sociales.

Estas investigaciones sentaron las bases para profundizar en el estudio de la percepción pública hacia los cigarrillos electrónicos (e-cigs), un tema de creciente interés en el ámbito de la salud pública y la regulación de productos de tabaco. En este contexto, un estudio propuesto por (Resende y Culotta, 2015) describe un enfoque metodológico para clasificar sentimientos hacia los e-cigs en X, analizando un amplio corpus de 105,605 publicaciones. Utilizando palabras clave específicas como ‘e-cigs”, “cigarrillo electrónico”, “vapor”, “vapeo”, “e-líquido” y “vaporizador personal”, esta investigación combinó técnicas de análisis de texto tradicionales y avanzadas, como *bag of words* y *SentiStrength*, junto con algoritmos de aprendizaje automático como SVM, Naïve Bayes, y *Random Forest*, para una evaluación detallada de las percepciones sobre el vapeo. El clasificador SVM demostró ser particular-

mente eficaz en este trabajo, identificando publicaciones relacionadas con publicidad e información sobre e-cigs con una precisión cercana al 80 %. Los resultados indicaron una clara inclinación positiva hacia los e-cigs entre los usuarios de X, destacando que un 11 % de las opiniones eran favorables frente a un reducido 3 % de opiniones negativas, evidenciando así la percepción mayoritariamente positiva de estos dispositivos entre la comunidad online.

Por otro lado, la investigación realizada por (Godea, Caragea, Bulgarov, y Ramisetty-Mikler, 2015) sexploró las actitudes y percepciones hacia los cigarrillos electrónicos a través de un análisis demográfico y de sentimientos de publicaciones en X a lo largo de un año. Este estudio se enfocó en categorizar los sentimientos de los mensajes como positivos, negativos o neutros, y en analizar la distribución de las opiniones según la edad y el género de los usuarios, con el objetivo de extraer información relevante de las actitudes hacia estos dispositivos entre diferentes grupos demográficos. Para abordar su objetivo, el estudio comenzó identificando publicaciones relevantes utilizando palabras clave específicas como: "e-cigarette", "e-cig", "ecig". Después del filtrado inicial, se seleccionaron 2,000 publicaciones de un total de 455,648 para una clasificación detallada de sentimientos, etiquetándolas manualmente como positivas, negativas o neutras. Utilizando un clasificador supervisado basado en regresión logística para el entrenamiento del modelo, se identificaron 103,103 publicaciones (22.6 %) como positivas y 56,652 (12.4 %) como negativas. Se destacó que, de las publicaciones no neutrales, aproximadamente el 65 % fueron catalogadas como positivas, demostrando una tendencia favorable hacia los cigarrillos electrónicos.

A su vez, la emergencia de la pandemia de Covid-19 ha reforzado significativamente la importancia de las redes sociales como reflejos de las percepciones ciudadanas, particularmente en asuntos críticos de salud pública como el consumo de tabaco y el vapeo. Este fenómeno ha evidenciado la manera en que eventos globales pueden influir y modificar las actitudes relacionadas con la salud. En este contexto, el estudio desarrollado por (Kamiński, Muth, y Bogdański, 2020) se propuso aprovechar esta circunstancia excepcional para investigar el impacto de la pandemia en las posturas hacia el tabaco y el vapeo. El estudio utilizó datos de X recopilados a través de la API (Interfaz de Programación de Aplicaciones) de X y el paquete rtweet en R. Dentro del análisis de 22,644,944 publicaciones, se encontró que 33,890 (0.15 %) estaban relacionadas tanto con el tabaco como con el Covid-19. Este subconjunto se examinó detalladamente, enfocándose en los sentimientos expresados, así como en las interacciones, como "likes" o "retweets", y su relación con el número de seguidores. Utilizando el paquete *tidytext* en R para el análisis de sentimientos, se llevó a cabo la tokenización de texto y la eliminación de palabras comunes, aplicando luego el diccionario de sentimientos "Bing" para calcular la diferencia entre sentimientos positivos y negativos en las publicaciones analizadas. De manera inesperada, los resultados inclinaron la balanza hacia sentimientos negativos, lo que sugiere un cambio en la percepción pública posiblemente motivado por las preocupaciones de salud desencadenadas por la Covid-19.

En el marco de un estudio complementario, (Yanamandra, Pant, y Mamidi, 2020) propu-

sieron un método innovador para investigar las discusiones sobre el tabaco en redes sociales. A través del conjunto de datos *SmokPro*, compuesto por más de 10,000 publicaciones en X sobre el tabaco, categorizadas manualmente en cinco grupos: menciones ambiguas, experiencias personales con el tabaco, advertencias sobre sus productos, anuncios comerciales y referencias a drogas no relacionadas con el tabaco, este estudio aporta una perspectiva novedosa al análisis de contenido en este ámbito. Se realizó un pre-procesamiento de las publicaciones para limpiar y preparar el texto antes de su clasificación, empleando posteriormente modelos avanzados de clasificación de texto como FastText, BERT (Bidirectional Encoder Representations from Transformers) y RoBERTa. El modelo BERT demostró ser el más eficaz, alcanzando una precisión del 97.10 % en la detección de productos de tabaco.

Otra aplicación interesante en esta línea es la desarrollada por (Sidani et al., 2020), quienes desarrollaron un estudio sobre la percepción pública de JUUL, un conocido sistema de administración de nicotina. El objetivo de la investigación incluía recopilar datos en tiempo real usando la API de X's Filtered Streams, con el fin de analizar publicaciones que incluyeran términos como "juul", "juuls" y "juuling". De las 67,934 publicaciones identificadas sobre JUUL, se eligió aleatoriamente un 2 % (1,209 publicaciones) para su codificación manual y posterior análisis de sentimientos. Este proceso incluyó una revisión iterativa para categorizar y destacar tendencias y temas predominantes, con especial atención en el uso de JUUL en lugares donde fumar cigarrillos tradicionales es limitado. Notablemente, el análisis mostró que el 71.5 % de las publicaciones reflejaban una percepción positiva de JUUL, frente a un 14.1 % que expresaban opiniones negativas. Entre los temas relevantes identificados se encontraba la visión de JUUL como una alternativa menos dañina y su elección como acto de rebeldía por algunos usuarios. Específicamente, el 55 % de las publicaciones relacionadas con el uso destacaban el consumo de JUUL en lugares donde fumar está prohibido, mientras que el 25 % comentaban sobre la permisividad de los adultos hacia el uso de estos dispositivos por parte de los jóvenes.

Por su parte, (Dobbs et al., 2023) emplearon un enfoque híbrido, combinando análisis cualitativo y cuantitativo para explorar las percepciones públicas sobre la Ley Federal de Tabaco 21 en X, que eleva la edad mínima para comprar productos de tabaco a 21 años en Estados Unidos. Del análisis de 231,447 publicaciones, se seleccionó una muestra del 2 % (4,628 publicaciones) para anotación manual, siguiendo un libro de códigos desarrollado desde una perspectiva inductiva para clasificar los sentimientos (a favor, en contra y neutral). La consistencia entre codificadores se comprobó mediante el coeficiente Cohen κ, asegurando la fiabilidad de los resultados, que mostraron un 46.2 % de publicaciones neutrales, 38.8 % en contra y 15 % a favor de la ley.

Recientemente, (Elmitwalli, Mehegan, Wellock, Gallagher, y Gilmore, 2024) llevó a cabo un estudio para analizar el sentimiento de las publicaciones sobre tabaco en X, con un enfoque particular en las discusiones en línea relacionadas con el control del tabaco durante la conferencia COP9. El objetivo principal fue identificar las tendencias y perspectivas pre-

dominantes entre los participantes de la plataforma. Inicialmente, se desarrolló una etapa de pre-procesado de las publicaciones antes de aplicar el modelo de Latent Dirichlet Allocation (LDA) para el modelado de tópicos, empleando técnicas como *word embedding* y *TF-IDF* (Term Frequency-Inverse Document Frequency) para preparar los datos de entrada. El modelado de tópicos permitió identificar categorías de discusión relevantes como la mitigación de riesgos, la regulación de cigarrillos electrónicos y las opiniones y percepciones públicas. En la etapa posterior de análisis de sentimientos, las publicaciones se categorizaron según su polaridad y se analizó la relación entre el sentimiento y la difusión de las mismas. Utilizando métodos de ML como la Regresión Logística, Naïve Bayes, SVM, XGBoost, *Extra Trees*, y *Random Forest*, este último se destacó por su alta precisión del 91.87 % en la clasificación de los temas identificados previamente. Como resultado, se observó una correlación significativa entre el sentimiento de las publicaciones y la frecuencia de *retweets*, con una tendencia a compartir mensajes de sentimientos intensos. Además, el análisis de toxicidad mostró que las discusiones mantenían bajos niveles de toxicidad, indicando un diálogo respetuoso sobre la reducción de daños en el control del tabaco.

El análisis previo de la literatura destaca la importancia de tareas de PLN como el modelado de tópicos y el análisis de sentimientos para extraer información valiosa sobre las percepciones ciudadanas en temas relacionados con el uso del tabaco. Aunque el modelado de tópicos no ha sido extensivamente explorado en este ámbito específico, ha proporcionado información relevante sobre las dinámicas y tendencias en las discusiones públicas. Entre las técnicas utilizadas, LDA se destaca por su eficacia en el análisis de textos en redes sociales, debido a su capacidad para descubrir temas subyacentes de manera no supervisada. Ese enfoque ofrece una metodología eficaz por su simplicidad, capacidad para procesar extensos conjuntos de datos, y su eficiencia comparativa en costos computacionales frente a arquitecturas más complejas. Estas ventajas hacen de LDA la técnica elegida para el análisis en este Trabajo de Fin de Grado, permitiendo el análisis e identificación de categorías temáticas presentes en las discusiones sobre tabaco en redes sociales.

En el análisis de sentimientos, los métodos varían desde enfoques basados en diccionarios hasta técnicas avanzadas de ML, incluidas las Redes Neuronales Profundas. Herramientas como Naive Bayes, *k-nearest neighbors* y SVM, así como software específico como *Salience Engine* 4.1 y técnicas de hedonometría, se utilizan para determinar la polaridad de los sentimientos en las publicaciones. Específicamente, Los métodos lexicográficos, basados en diccionarios, ofrecen ventajas como la simplicidad y la interpretabilidad directa en comparación con métodos más complejos, facilitando un análisis rápido y accesible de grandes volúmenes de texto.

Considerando estas ventajas, en el presente trabajo de Fin de Grado se optará por un método basado en diccionarios para el análisis de sentimientos. Específicamente, se implementará VADER (Valence Aware Dictionary for Sentiment Reasoning), una técnica diseñada para capturar la singularidad de los textos en redes sociales. VADER sobresale por su habi-

lidad para entender la complejidad lingüística y los matices emocionales propios de estas plataformas, incluyendo el uso de emoticonos, intensificadores y jerga informal. Su adaptabilidad y precisión en entornos de comunicación digital lo convierten en una herramienta idónea para analizar las percepciones sobre el tabaco en redes sociales.

Finalmente, la Tabla 2.1 presenta un resumen de los estudios más destacados analizados previamente, incluyendo tanto las palabras clave utilizadas como los modelos de análisis predominantes de cada investigación. Esta tabla proporciona una panorámica comparativa de las tendencias predominantes, metodologías y resultados clave en la investigación sobre las percepciones del tabaco en X.

Tabla 2.1: Resumen de los estudios sobre percepción pública llevados a cabo mediante el análisis de las publicaciones.

Referencia	Tamaño del Dataset	Método de Adquisición	Objetivo	Algoritmo	Resultados
(Myslín et al., 2013)	7.362 publicaciones	Palabras Clave: "cig*", "nicotina", "smok*", "tabaco", "hooka", "shisha","vapear","vapeo", "vaping" en un periodo de 15 días, desde diciembre de 2011 hasta julio de 2012	Realizar un análisis del contenido y el sentimiento de las publicaciones relacionadas con el tabaco en X, centrándose en los nuevos y emergentes productos como la cachimba y los cigarrillos electrónicos.	Análisis de Contenido y Sentimientos: Naïve Bayes, k-Nearest Neighbors y SVM.	46 % de las publicaciones expresaban sentimientos positivos hacia el tabaco, 32 % expresaban sentimientos negativos.
(Cobb et al., 2013)	13.200 publicaciones aprox.	Palabras Clave: "vareniclina", "bupropión","Chantix","Champix", "Zyban"	Analizar la exposición a mensajes en línea sobre el medicamento para dejar de fumar "vareniclina" y su efecto en la toma de decisiones de los fumadores en cuanto a su uso.	Análisis de sentimientos: Salience Engine 4.1.	Odds ratio de 2.05 (IC del 95 %: 1.66 - 2.54) para el cambio a vareniclina: aquellos expuestos a mensajes positivos tenían el doble de probabilidades de adoptar este medicamento en comparación con los no expuestos. Odds ratio para continuar usando vareniclina de 2.46 (IC del 95 %: 1.66 - 2.54): la probabilidad de mantenerse en el tratamiento era aproximadamente 2.5 veces mayor en aquellos expuestos a mensajes positivos.
(Clark et al., 2014)	20.000 publicaciones	Palabras Clave: "tabaco", "cigarrillo", "fumar", "cigarrillo electrónico" y "vapear"	Evaluar la densidad y el sentimiento de las publicaciones relacionadas con el tabaco y los cigarrillos electrónicos en X	Análisis de sentimientos: hedonometría	Correlación positiva significativa (Pearson's $r = 0.54$, $p < 0.01$) entre la densidad relativa de publicaciones relacionadas con el tabaco por estado y la positividad promedio de las publicaciones sobre tabaco.
(Godea et al., 2015)	955.368 publicaciones	Palabras clave:" e-cigarette", "ecigarette", "e-cig" y "ecig" en un periodo de un año	Analizar las actitudes y percepciones hacia los e-cig expresadas en X	Clasificación supervisada: Regresión logística	22,6 % publicaciones positivas; 12,4 % publicaciones negativas; 65 % de las publicaciones no neutrales fueron etiquetadas como positivas. Restringiendo los resultados en el umbral de mayor confianza para el clasificador, la precisión para las clases positiva y negativa fue del 96 % y 70 %, respectivamente
(Resende y Culotta, 2015)	105.605 publicaciones	Palabras clave:"cigarrillo electrico","e-cig" y "vapor", "vapeo", "e-liquido" y "vaporizador" en un periodo de marzo a abril de 2014	Identificar los sentimientos y la difusión de información relacionados con los e-cigs en X	Análisis de Sentimientos: SVM, Naïve Bayes y *Random Forest*	Los usuarios tienden a compartir opiniones/experiencias positivas (11 %) en mayor medida que opiniones negativas (3 %)
(Kamiński et al., 2020)	1.868.755 publicaciones	Palabras clave:"COVID", "coronavirus", "pandemia", "SARS" y "CoV"; "fumar", "cigarrillo", "humo" y "nicotina" durante el periodo del 1 de enero de 2020 al 1 de mayo de 2020.	Analizar la dinámica de la discusión en X sobre el tabaco durante la pandemia de COVID-19.	Análisis de sentimientos:paquete *tidytext* de R	33,890 (0.15 %) publicaciones estaban relacionadas tanto con el tabaco como con COVID-19. El sentimiento de las publicaciones fue generalmente negativo, siendo el más bajo a mediados de marzo. El tono se volvió menos negativo en abril.

Referencia	Tamaño del Dataset	Criterio de Selección	Objetivo	Algoritmos	Resultados
(Yanamandra et al., 2020)	10.000 publicaciones	Palabras clave: “tabaco”, “cigarrillo”, “fumar”, “cachimba”, “cigarro”, “vapear” y “nicotina” durante el mes de octubre de 2018	Clasificar publicaciones relacionadas con productos de tabaco, centrándose en la identificación de diferentes tipos de productos de tabaco	Modelos de Clasificación : BERT, RoBERTa y FastText	El modelo Cased BERT Large fue el más efectivo en la tarea de identificación de productos de tabaco en publicaciones, con una precisión y un F1 del 97.10 %.
(Sidani et al., 2020)	67.934 publicaciones	Palabras clave: “juul”, “juuls”, “juuling”,“vape”, “vapeo”, “e-cigarette”, “e-cig”,“nicotina”, “fumar” y “tabaco” .	Analizar la discusión y el sentimiento público en X en relación con el uso de JUUL, un popular dispositivo de vapeo.	Análisis de Sentimientos: codificación manual	71,5 % de las publicaciones expresaban un sentimiento positivo hacia JUUL, mientras que el 28,5 % restante expresaba un sentimiento negativo. 9,2 % de las publicaciones mencionaban el uso de JUUL en lugares donde fumar cigarrillos está prohibido. 20 % de las publicaciones mencionaban el uso de JUUL en el hogar o frente a adultos responsables
(Dobbs et al., 2023)	2.758 publicaciones	Palabras clave:“Tabaco 21”, “MLSA” (Edad Mínima de Venta Legal), “Edad para Fumar”, “Ley de Control del Tabaco”, “FDA” (Administración de Alimentos y Medicamentos), “Regulación del Tabaco”, “Prevención del Nicotina en Jóvenes”	explorar el sentimiento y los temas subyacentes en las publicaciones relacionadas con la Ley Federal de Tabaco 21 en los Estados Unidos.	Análisis de sentimientos: libro de códigos y análisis temático	La política neutral fue predominante (46,2 %), seguida de antipolítica (38,8 %). Los debates se centraron en noticias políticas, con argumentos a favor enfocados en la protección de jóvenes, mientras que los argumento en contra criticaban la ley por ser injusta hacia los jóvenes adictos a la nicotina.
(Elmitwalli et al., 2024)	7.376 publicaciones	Palabras clave:“cigarrillo”, “fumar”, “nicotina”, “vapeo” y “fumador” durante el mes de noviembre de 2021	Analizar la percepción pública centrándose en la “reducción de daños” dentro del contexto más amplio del control del tabaco	LDA, Análisis de sentimientos, análisis de toxicidad y técnicas de aprendizaje supertvisado (árboles de decisión,SVM y regresión logística)	Precisión de la Clasificación de Tópicos: 91.87 % con *Random Forest*,lo que permitió identificar categorías relevantes como la mitigación de riesgos, la regulación de productos electrónicos y las opiniones y percepciones públicas. Correlación significativa entre el puntaje de sentimiento de las publicaciones y la cantidad de *re-tweets*.

Capítulo 3

Metodología de Análisis de Datos

Este capítulo detalla la metodología aplicada para analizar las percepciones de la ciudadanía hacia el tabaco a través de la plataforma X. Se estructura en cinco fases distintas, describiendo las actividades correspondientes a cada una y su secuencia lógica, las cuales se visualizan en la Figura 3.1, facilitando así una comprensión clara del flujo de trabajo adoptado en esta investigación

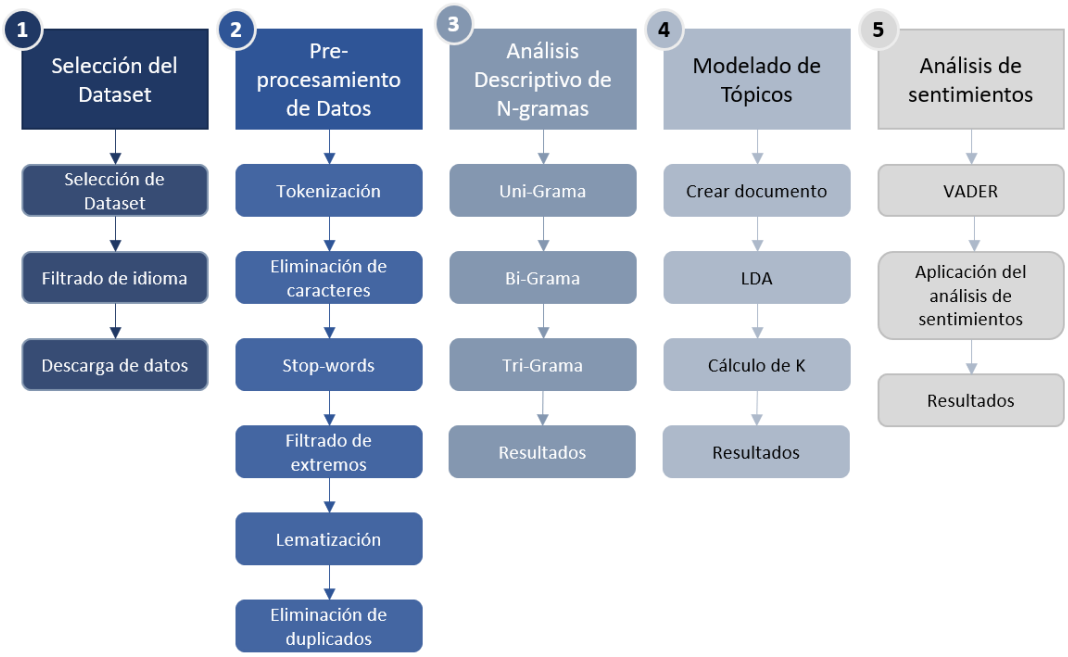

Figura 3.1: Flujo de actividades de la metodología implementada. Elaboración propia

Las cinco etapas representadas en la Figura 3.1 comprenden desde la selección del conjunto de datos hasta el análisis de sentimientos, incluyendo el preprocesamiento de datos, el análisis descriptivo de n-gramas y el modelado de tópicos

La primera etapa implica la selección del conjunto de datos, definiendo criterios claros

para su elección y realizando un análisis preliminar de los datos y variables incluidos. La segunda etapa, el pre-procesamiento de datos, se enfoca en preparar el conjunto para análisis futuros, eliminando ruidos e inconsistencias mediante la limpieza, normalización y estructuración de los datos. En la tercera fase, el análisis descriptivo de n-gramas, se emplean técnicas como TF-IDF para destacar los conceptos clave en las discusiones sobre tabaco, revelando así los temas generales y preocupaciones predominantes de los usuarios.

Posteriormente, el modelado de tópicos mediante el algoritmo LDA permite clasificar los temas en categorías definidas, proporcionando un análisis más detallado de los principales puntos de interés y discusiones en la comunidad en línea. Esta técnica de ML no supervisado es fundamental para comprender las diversas perspectivas abordadas por los usuarios sobre el tabaco. Finalmente, el análisis de sentimientos se realiza con VADER, brindando una evaluación cuantitativa de las emociones en los textos, categorizándolos en positivos, negativos o neutros. A continuación, se procede a detallar las tareas específicas en cada fase de la metodología.

3.1. Selección de Dataset

La adquisición de X por Elon Musk en octubre de 2022 ha provocado cambios significativos en el uso de su API, afectando a quienes utilizan datos de la plataforma para investigación o desarrollo de aplicaciones. Estas modificaciones han restringido el acceso a la información, imponiendo a investigadores y desarrolladores la necesidad de adaptarse a nuevas normativas. La implementación de una política de precios ha limitado aún más este acceso, diferenciando niveles de acceso con variaciones en costos y funcionalidades. Las opciones sin costo presentan restricciones considerables, mientras que las alternativas de pago, aunque brindan mayor acceso a los datos, implican costos elevados, afectando la capacidad de recopilación y manejo de información (X, 2023).

Dadas las limitaciones impuestas por cambios en la API de X, este estudio considera un conjunto de datos público que facilita el análisis requerido: el dataset SmokPro (Smokpro, 2021). Tras revisar varios repositorios y evaluar la cantidad y tipo de datos disponibles, se seleccionó SmokPro, que incluye 2,117 publicaciones sobre productos de tabaco en X. Este conjunto de datos se ha tratado específicamente para incluir solo publicaciones en inglés, excluyendo menciones directas para proteger la privacidad de los usuarios. Del conjunto de datos resultante, se analizará exclusivamente la variable de texto, que contiene la información textual de cada publicación. Aunque el dataset proporciona una clasificación inicial basada en tipos de tabaco, este estudio no la incorporará al análisis. En su lugar, se adoptará un enfoque de categorización no supervisada para explorar las diversas categorías de discusión, permitiendo una comprensión más detallada y matizada de las conversaciones sobre el tabaco en la plataforma.

3.2. Pre-procesamiento de los Datos

El preprocesamiento de datos es fundamental en el análisis de texto, ya que prepara los datos para su posterior tratamiento. Esta fase se centra en la limpieza de datos, eliminando elementos innecesarios como valores atípicos o duplicados que podrían comprometer el análisis. Esta etapa incluye diversas actividades necesarias para garantizar la calidad de los datos, tales como la normalización de texto, la eliminación de *stopwords* y la lematización del texto, facilitando así un análisis más eficaz del corpus textual. A continuación, se describe en detalle cada una de ellas.

El primer paso en el preprocesamiento es la tokenización, que consiste en dividir el texto en unidades más manejables, conocidas como *tokens*, que pueden ser palabras, símbolos o caracteres. Por ejemplo, la frase "El consumo del tabaco es un problema de salud pública" se dividiría en los siguientes tokens (en este estudio definidos como palabras): "El", "consumo", "del", "tabaco", "es", "un", "problema", "de", "salud", "pública". Este tratamiento facilita su análisis al permitir la separación y normalización de cada elemento de la oración, simplificando la identificación de patrones dentro del texto. A continuación, se normaliza el texto en minúsculas y se elimina cualquier carácter no deseado como enlaces, URLs, números, espacios excesivos, puntuación y hashtags. Estos elementos pueden distorsionar el análisis, introduciendo ruido y desviando la atención de los patrones lingüísticos significativos. Su eliminación es requerida para asegurar que el análisis se centre en el contenido textual relevante, mejorando así la calidad de los resultados al evitar interpretaciones erróneas o análisis sesgados debido a información irrelevante.

La etapa siguiente se enfoca en una limpieza exhaustiva mediante la eliminación de las *stopwords*, palabras comunes con limitada carga semántica, como "la", "en","y", "el", "de", "que" y "a", que suelen ser omitidas en el análisis textual por su baja contribución al significado. Para esta tarea, se utilizan diccionarios especializados para filtrar estas palabras y, según el contexto del estudio, se pueden agregar términos específicos que, aunque frecuentes, no aporten valor significativo al análisis. Posteriormente, se efectúa un filtrado para descartar palabras extremas, es decir, aquellas con una presencia muy baja o excesivamente alta en el conjunto de datos, debido a su escasa contribución al análisis. También, se eliminan las publicaciones de menos de tres caracteres para garantizar el enfoque en contenido significativo para el modelado de tópicos. Este proceso es de gran importancia para depurar el texto y concentrarse en el contenido más relevante.

El proceso de preprocesamiento continúa con la lematización, transformando palabras a su forma base, por ejemplo, convirtiendo "consumió", "consumirá" y "consumiste" en "consumir". Este paso es fundamental para el modelado de tópicos, ya que agrupa variaciones de la misma palabra, mejorando la consistencia y relevancia del análisis. Finalmente, se eliminan las publicaciones duplicadas para asegurar la unicidad de la información analizada, previniendo redundancias que puedan distorsionar los resultados del análisis. Es importante

destacar que estas fases son requeridas tanto para el análisis descriptivo de n-gramas como para el modelado de tópicos. Sin embargo, la estrategia de preprocesamiento se modifica en el análisis de sentimientos. Para este último, se evita la normalización a minúsculas, se conservan las *stopwords* y los signos de puntuación, ya que VADER utiliza estos elementos para determinar la polaridad de los textos y garantizar una alta precisión en la cuantificación del sentimiento.

3.3. Exploración de N-Gramas

Tras completar el preprocesamiento de los datos textuales, se lleva a cabo un análisis descriptivo mediante N-gramas para detectar las secuencias de palabras más significativas en los datos. Los N-gramas, que son secuencias contiguas de N elementos lingüísticos, facilitan la identificación de temas clave al evaluar combinaciones relevantes de palabras. Este análisis, desde unigramas (palabras individuales) hasta bigramas y trigramas (secuencias de dos o tres palabras), proporciona una visión general de los conceptos predominantes y los conceptos recurrentes en el conjunto de datos. Para determinar la relevancia de secuencias de palabras y su importancia individual, se utiliza la técnica Frecuencia de Término-Frecuencia Inversa de Documento (TF-IDF), que mide la significancia de un término dentro de un documento en relación a su frecuencia en el corpus completo. Esta metodología combina la Frecuencia de Término (TF), que indica la frecuencia de aparición de un término en el documento, con la Frecuencia Inversa de Documento (IDF), evaluando la exclusividad del término en el corpus, como se describe en la ecuación 3.1. Este enfoque permite resaltar términos que son especialmente pertinentes para el tema analizado.

$$TF{-}IDF(i) = TF(i) \times IDF(i) \tag{3.1}$$

En la ecuación 3.1, "i" denota un término concreto. En pocas palabras, la metodología TF-IDF ajusta la frecuencia de un término en un documento específico (TF, ver ecuación 3.2) frente a su relevancia en el corpus completo (IDF, ver ecuación 3.3), realzando la significancia de términos poco comunes. Al asignar mayor valor a estos términos raros, se logra una estimación más precisa de su importancia, enriqueciendo el análisis al destacar su contribución única al contenido del texto.

$$TF(i) = \frac{\text{Frecuencia absoluta de la palabra i en el documento}}{\text{número total de palabras en el conjunto de datos}} \tag{3.2}$$

$$IDF(i) = \log_e \frac{\text{número total de documentos en el corpus}}{\text{número de documentos que incluyen la palabra i}} \tag{3.3}$$

La combinación de TF y IDF en el cálculo TF-IDF produce una métrica que destaca la relevancia de términos tanto individualmente como en el conjunto del corpus, permitiendo

distinguir la importancia de cada término y resaltar aquellos más críticos y únicos en el análisis. Una alta puntuación en TF-IDF indica la relevancia especial de un término. Finalmente, al organizar los n-gramas según su longitud (unigramas, bigramas y trigramas), se identifican patrones de conversación generales y terminología específica, facilitando la identificación de relaciones clave entre términos, determinante para el modelado de tópicos y el análisis posterior de sentimientos.

3.4. Modelado de Tópicos

El modelado de tópicos es una técnica avanzada de análisis automático de texto muy valorada en el campo del PLN. Su propósito consiste en identificar los temas o "tópicos" subyacentes en un conjunto de documentos. Esta herramienta es fundamental para estructurar y analizar grandes volúmenes de datos textuales, facilitando una interpretación detallada del contenido del corpus textual. Dentro de las técnicas propuestas para el modelado de tópicos, Latent Dirichlet Allocation (LDA) destaca por su robustez, flexibilidad y la alta interpretabilidad de sus resultados, lo cual la convierte en una metodología altamente valorada en el campo Blei, Ng, y Jordan (2003).

En términos generales, LDA modela cada documento como una mezcla de diversos tópicos, asignando probabilísticamente cada palabra a un tema. Esta asignación se realiza con base tanto en la distribución de los tópicos dentro del documento como en la capacidad del modelo para asociar palabras con temas específicos. Este proceso iterativo se perfecciona continuamente para descubrir la estructura temática latente entre los documentos, denominada "latencia" porque estos temas, aunque no son explícitamente visibles en los datos, se infieren a través del análisis. En la figura 3.2 se muestra una representación visual del modelado de tópicos mediante LDA, presentando los parámetros clave que han sido tomados en consideración:

- M: Este parámetro señala el total de documentos presentes en el corpus, estableciendo así el alcance del análisis y el volumen de información que será analizada mediante el modelado.
- N: El número de palabras por documento, el cual varía en cada documento influyendo en la distribución temática dentro de cada uno.
- α (Alfa) y η (Eta): Los hiperparámetros de Dirichlet α y η son fundamentales en LDA, definiendo la distribución de tópicos por documento y de palabras por tópico, respectivamente. α regula la variedad de tópicos en cada documento, mientras que η controla la diversidad de palabras asociadas a cada tópico. Ambos parámetros son requeridos para ajustar la granularidad del modelo, influenciando directamente la especificidad y generalidad en la identificación de tópicos.

- θ (Theta): Distribución de tópicos en un documento, extraída de una muestra a priori de Dirichlet, basada en α.
- β (Beta): Distribución de palabras en un tópico, también derivada de una muestra a priori de Dirichlet, utilizando η.
- Z: Este parámetro establece la relación entre cada palabra en el documento y los tópicos inferidos, asignando tópicos específicos a cada palabra.
- W: Cada palabra en un documento, este parámetro actúa como la unidad fundamental para el análisis en el modelado de tópicos.

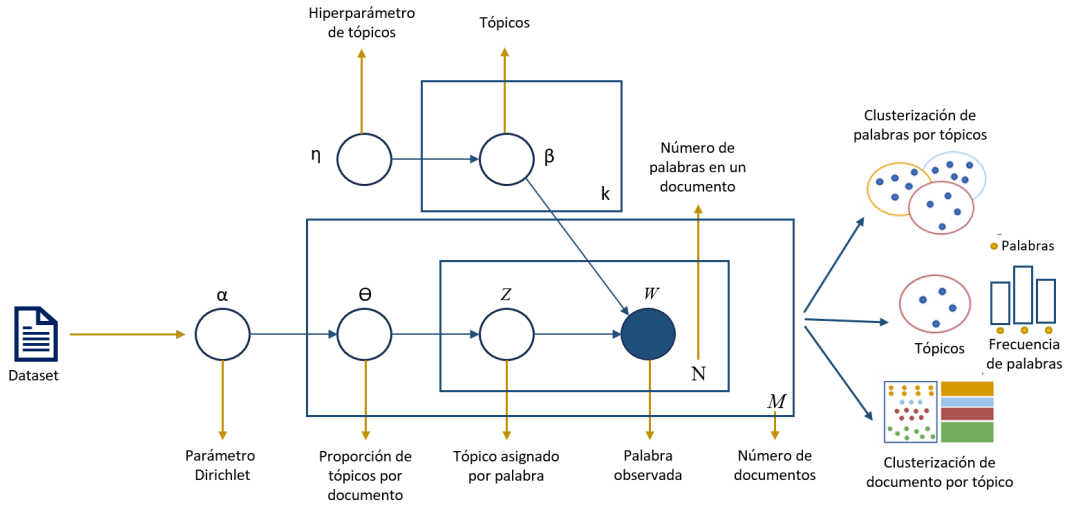

Figura 3.2: Descripción del modelo LDA. Elaboración propia

Para el óptimo funcionamiento de LDA, se requiere determinar de antemano el número adecuado de tópicos, denominado "k". Este paso es fundamental ya que influye en la precisión e interpretabilidad de los temas identificados por el modelo. Dado que no existe una norma establecida, es necesario equilibrar la especificidad y generalidad de los temas para evitar la fragmentación excesiva o la mezcla inadecuada de términos. Un indicador clave para ajustar este parámetro es la coherencia temática, reflejando información relevante sobre las relaciones entre las palabras clave de cada tema. Por tanto, en este estudio se evaluará la coherencia de los temas para distintos valores de número de tópicos, y se determinará el conjunto óptimo que resulte en el conjunto de tópicos con el mayor índice de coherencia. Este enfoque se tiene en cuenta con el objetivo de mejorar la interpretación de los temas identificados.

Herramientas de visualización como pyLDAvis resultan de gran utilidad en el análisis, ofreciendo representaciones gráficas de la distancia entre tópicos y facilitando la comprensión de su distribución y solapamiento mediante diversas métricas de coherencia (Sievert

y Shirley, 2014). A través de la reducción de dimensionalidad, como el uso de t-SNE (t-distributed Stochastic Neighbor Embedding) , pyLDAvis permite visualizar los tópicos en un espacio bidimensional, y de esta manera, evaluar su diferenciación y solapamiento. Este análisis es necesario para confirmar que los tópicos modelados sean distintivos y coherentes, evitando intersecciones significativas que pudieran afectar su claridad interpretativa.

3.5. Análisis de Sentimientos

El análisis de sentimientos es una técnica avanzada en el campo del PLN, que juega un papel muy importante en la comprensión de las emociones humanas expresadas a través del lenguaje escrito. Esta metodología permite identificar y clasificar las emociones contenidas en un texto, asignándolas a categorías como positivas, negativas o neutras. Es especialmente útil en redes sociales, donde permite monitorear percepciones y opiniones de los usuarios sobre diversos temas. Al analizar grandes volúmenes de publicaciones, permite extraer información relevante de la actitud general hacia productos, servicios o temas específicos, facilitando la toma de decisiones basada en datos y mejorando la comprensión del comportamiento de los usuarios en el entorno digital.

Para este Trabajo de Fin de Grado, que busca comprender la percepción pública sobre el consumo de tabaco, se ha seleccionado el análisis de sentimientos basado en diccionarios, utilizando específicamente el Lexicón VADER. Esta elección se debe a la robustez y la validación de VADER en estudios previos, destacados en la investigación de (Hutto y Gilbert, 2014). Esta investigación no solo detalla el algoritmo que fundamenta VADER, sino que también valida su capacidad para analizar el lenguaje informal y las expresiones características de las redes sociales. En este contexto, VADER se presenta como una herramienta especialmente adecuada para el análisis de sentimientos en estos entornos digitales. Este diccionario destaca por combinar elementos léxicos y sentimentales, asignando una polaridad emocional específica a cada palabra, lo cual permite medir su intensidad con gran precisión.

La implementación de VADER en este estudio ofrece una evaluación del sentimiento, asignando puntuaciones entre -4 y +4. Esta escala potencia el análisis al distinguir entre sentimientos positivos y negativos y permite determinar su grado de intensidad. Por ejemplo, expresiones altamente positivas como “excelente” se aproximan a +4, mientras que comentarios negativos intensos como “terrible” se acercan a -4.

Específicamente, VADER implementa el siguiente conjunto de reglas (Hutto y Gilbert, 2014):

- **Puntuación y mayúsculas**: Esta regla se refiere al modo en como se usan los signos de puntuación y letras mayúsculas para enfatizar o expresar emociones en un texto. Por ejemplo, “¡NO puedo creerlo!” expresa una emoción más intensa que simplemente decir “No puedo creerlo”.

- **Modificadores de grado**: Términos, como “muy”, “bastante”, “extremadamente”, se utilizan para modificar la intensidad de las expresiones. Al decir “Está muy feliz”, el “muy” intensifica la felicidad expresada. En cambio, “Está algo cansado” minimiza la intensidad del cansancio.

- **Consideraciones sintácticas**: La estructura de las frases puede alterar significativamente el significado o la intensidad de la expresión. Por ejemplo, incluir una negación puede cambiar completamente el sentimiento, como en “No estoy feliz” comparado con “Estoy feliz”. Asimismo, el uso de conjunciones como “pero” o “sin embargo” puede introducir un matiz de contraste, como en “Estoy feliz, pero cansado”.

- **La reglas de gramática y sintaxis**: Se refieren a reglas generales que guían la construcción de oraciones, empleando distintos elementos lingüísticos para expresar emociones. El uso de adverbios como “rápidamente” en “Se calmó rápidamente” no solo indica la acción de calmarse, sino también la rapidez con que sucede, añadiendo un significado adicional a la frase. La selección de adjetivos, la estructura de la oración y la relación entre las partes del discurso son fundamentales para interpretar la intensidad de los sentimientos expresados.

En la implementación práctica, el análisis se centra en el cálculo de una puntuación compuesta para cada texto, integrando las puntuaciones individuales de las palabras y normalizando el resultado en un rango de -1 a 1. Es importante señalar que el pre-procesado de datos en el análisis de sentimientos cambia en relación al modelado de tópicos. En este contexto, se evita la normalización para conservar tanto las minúsculas como las mayúsculas, y se mantienen las stopwords y los signos de puntuación. Esta diferencia en el enfoque de preprocesamiento garantiza una representación más auténtica del texto original en el análisis de sentimientos, lo que permite capturar matices lingüísticos importantes para una interpretación precisa de las emociones expresadas en las publicaciones.

La incorporación de VADER amplía el espectro de análisis, aplicándose desde el monitoreo de percepciones de marca hasta la evaluación de respuestas ante eventos significativos, demostrando su valor en diversos ámbitos. Los estudios propuestos por (Shah, Shah, Rand, y Champon, 2024), y por (Kumari et al., 2024), describen la aplicabilidad de las técnicas de análisis de sentimientos en redes sociales, al captar matices emocionales y tendencias en la opinión pública que otros métodos podrían no detectar. En conclusión, la selección de VADER en este estudio resalta la necesidad de herramientas precisas para el análisis de sentimientos, particularmente en entornos de lenguaje informal como las redes sociales. Su habilidad para discernir y cuantificar emociones en textos, junto con su implementación sencilla y eficiencia computacional, facilita la obtención de información valiosa para entidades comerciales y análisis socio-económicos.

Capítulo 4

Resultados

Este capítulo muestra los resultados alcanzados mediante la metodología expuesta en el Capítulo 3, aplicada al conjunto de datos seleccionado para esta investigación. Es importante destacar que el foco del presente análisis radica en extraer información relevante sobre las percepciones ciudadanas acerca del tabaco. El código empleado ha sido subido en un repositorio en GitHub, estando disponible gratuitamente para su consulta y revisión (Mazarío.S, 2024). El pre-procesamiento de texto fue realizado utilizando el lenguaje R, y la experimentación con los diferentes modelos se efectuó en Python.

4.1. Selección de Dataset y Preprocesamiento de los datos

En el contexto de este estudio, tal como se explica en la sección 3.1 dedicada a la preparación de los datos, se eligió trabajar con un conjunto de datos preexistente. En este caso, este conjunto fue inicialmente obtenido aplicando un filtro de palabras clave relacionadas con productos de tabaco. Este proceso implicó la elección exclusiva de publicaciones en inglés, y la eliminación de menciones o referencias directas no pertinentes en las publicaciones. Como resultado de este proceso de filtrado, el conjunto de datos que sirve de base para este estudio quedó compuesto por 2,117 publicaciones enfocadas en el tema del tabaco.

En la etapa de limpieza y pre-procesamiento, se seleccionaron las variables importantes y se les dio el formato necesario para su posterior análisis, como se menciona en la sección 3.2. Este proceso incluyó la eliminación de datos incompletos y la estandarización del contenido del texto, eliminando signos de puntuación y números, además de ajustar todo el texto a minúsculas. Con estos pasos, se obtuvo un conjunto de datos limpio y uniforme. Al concluir esta etapa, el conjunto de datos final para el análisis esta compuesto de 1.814 publicaciones, seleccionadas para investigar y entender los patrones y tendencias en las conversaciones sobre productos de tabaco en X.

En primer lugar, para realizar un análisis descriptivo inicial del contenido, se evalua la distribución de ##*hashtags* y emoticonos en el conjunto específico de publicaciones. De esta

manera, en la Figura 4.1 se analiza la distribución de los 20 *##hashtags* más frecuentes en el conjunto de datos, y se observa una tendencia predominante en la cual las palabras “smoke”, “vape”, “tobacco” y “cigar” forman la base de la mayoría de las etiquetas más utilizadas. De este núcleo se destacan *##hashtags* complementarios que profundizan en aspectos específicos de la cultura y estilo de vida asociados al consumo de tabaco. Por ejemplo, #cigarlife y #smokingfetish podrían referirse a una faceta más subcultural o de nicho relacionada con la actividad de fumar. Esta agrupación de *##hashtags* sugiere un diálogo en línea enfocado no solo en el acto de fumar sino también en las identidades y comunidades que se forman alrededor de este hábito

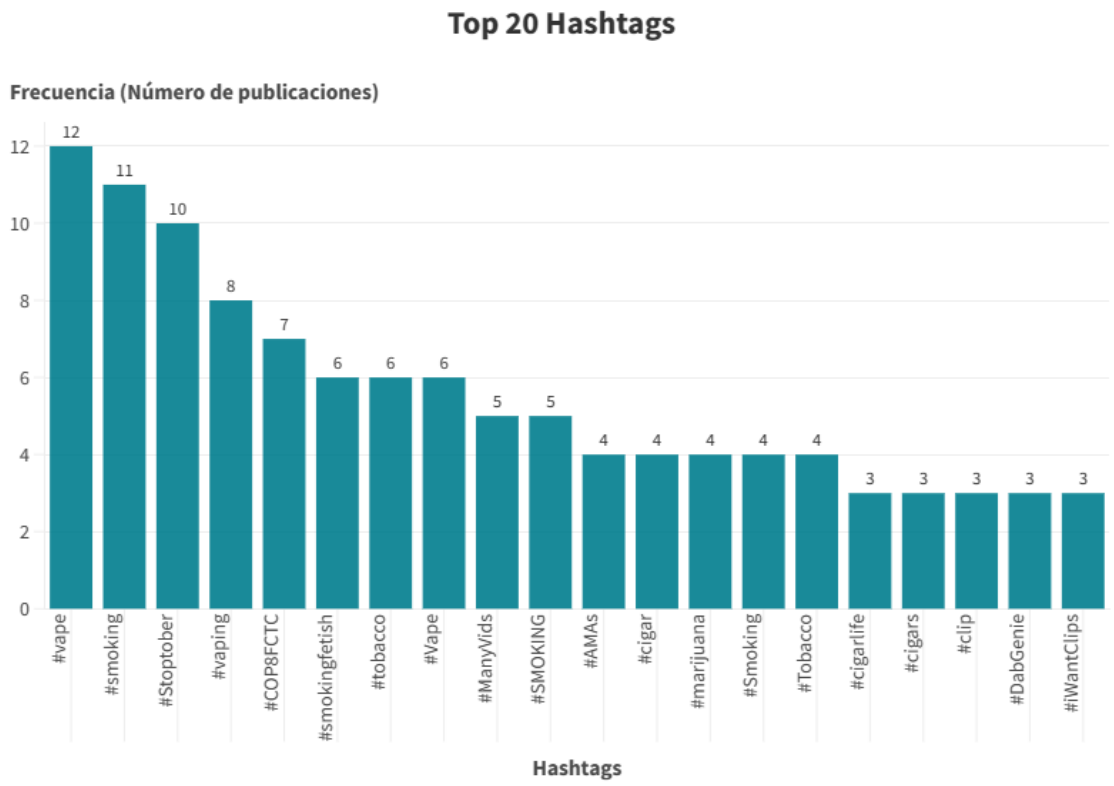

Figura 4.1: Cantidad de publicaciones por los Principales Hashtags
Elaboración propia

En cuanto a la distribución de frecuencia para los emoticonos,los cuales se muestran en la Figura 4.2, aproximadamente el 18 % de estos corresponden a caras sonrientes con lágrimas, lo que podría interpretarse como una actitud de diversión o sátira en las interacciones. Significativamente, los emoticonos relacionados con el tema del tabaco como la llama, el cigarrillo y el humo también figuran entre los más usados, lo que indica que los elementos visuales utilizados están estrechamente vinculados al tema central del tabaco en el corpus analizado. Además, los emoticonos de caras enamoradas y corazones rojos sugieren emociones positivas como afecto y aprecio entre los usuarios, mientras que los símbolos de llanto o tristeza, así como aquellos que podrían interpretarse como representaciones de la cesación

del tabaco, resaltan las emociones negativas o los esfuerzos por dejar de fumar.

Estos resultados muestran la importancia de los emoticonos como instrumentos de comunicación emocional en x, permitiendo una interpretación inicial del sentimiento subyacente en el corpus. La presencia frecuente de emoticonos sonrientes puede ser indicativo de un ambiente de cordialidad y humor entre los usuarios, lo que podría fomentar interacciones más amenas en la plataforma. Al mismo tiempo, la variedad de emoticonos usados, desde corazones que sugieren afecto hasta aquellos que expresan tristeza o preocupación, refleja el amplio rango de emociones que el tema del tabaco provoca en el conversaciones en línea.

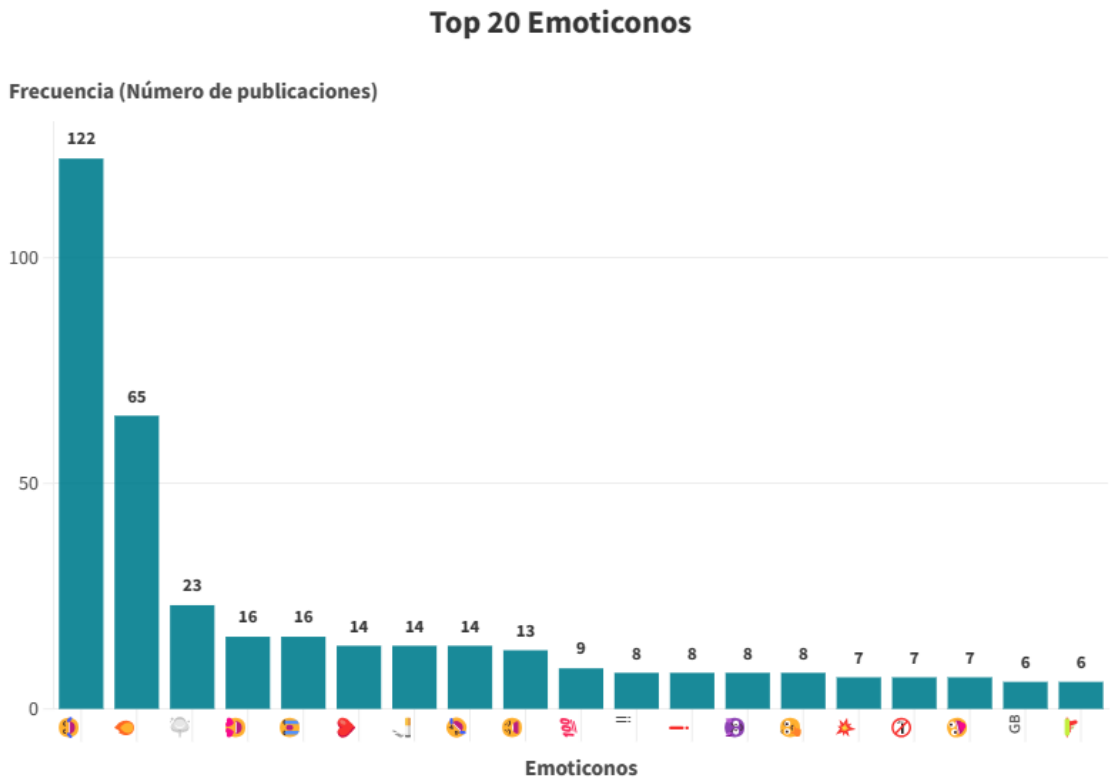

Figura 4.2: Cantidad de publicaciones en X por emoticono
Elaboración propia

4.2. Análisis exploratorio de N-Grama

A partir de los datos preprocesados, se lleva a cabo un análisis de las combinaciones de palabras de longitud N, denominadas N-gramas. Para explorar estos datos, inicialmente se genera una nube de palabras como la mostrada en la Figura 4.3. Esta representación proporciona una visión general de las 200 palabras más frecuentes presentes en las discusiones sobre los productos de tabaco en línea. En este tipo de visualización, la frecuencia de cada término está directamente correlacionada con su tamaño, lo que significa que las palabras más men-

cionadas aparecen con un tamaño mayor. Esto facilita la identificación rápida de los términos dominantes en el discurso, mostrando de manera visual los conceptos más discutidos en el corpus.

Figura 4.3: Nube de palabras
Elaboración propia

Al examinar la nube de palabras detalladamente, se identifican los términos más notables como “cigarette”, “tobacco”, “vape”, “hookah”, “crack”, y “weed”, “pot”, los cuales sugieren un debate en curso sobre las diferentes opciones de productos de tabaco, incluyendo tanto el uso tradicional como nuevos dispositivos vaporizadores. Estos términos destacan también la variedad en el consumo de tabaco, mostrando tanto las alternativas como las preferencias dentro de la comunidad. Al mismo tiempo, términos como “quit” y “stop” reflejan la intención de dejar el hábito de fumar, demostrando además una preocupación por la salud y un esfuerzo por superar la dependencia. Por otro lado, palabras como “drug”, “drink”, y “nicotine” muestran un enfoque hacia las adicciones y la salud pública, sugiriendo que las discusiones no se limitan solo al tabaco sino que abarcan un espectro más amplio de sustancias y comportamientos adictivos. Además, la inclusión de términos como “love”, “bad”, “shit”, y “good” en la nube de palabras podrían estar asociados a sentimientos y percepciones hacia el consumo de tabaco, sus productos y las adicciones. Estos términos proveen un análisis preliminar sobre las categorías de discusión y el análisis de sentimientos, mostrando los posibles desafíos individuales y las opiniones variadas sobre la cultura del consumo, además de las emociones y las valoraciones asociadas a esta temática.

Siguiendo con el análisis inicial, se lleva a cabo una evaluación detallada de los N-gramas más destacados en el conjunto de datos, empleando la métrica TF-IDF para cada término, tal como se explica en la sección 3.3. El propósito es identificar tanto los unigramas, bigramas y trigramas más significativos, para así adquirir una comprensión completa no solo de las palabras más usadas individualmente, sino también de las combinaciones de palabras que desempeñan un papel dominante en el corpus. La primera fase del análisis se centró en identificar los términos individuales de mayor relevancia que aparecen en el corpus. De esta manera, la distribución de la puntuación TF-IDF para cada palabra del conjunto de datos ha sido calculada. Los resultados más importantes de este estudio se observan en la Figura 4.4, mostrando las palabras dominantes dentro del corpus. Términos como “cigarette”, “vape”, “weed” ,“tobacco”, “nicotine”,“stop”, “quit”,“hookah”, y “crack”, sobresalen como las más relevantes, confirmando los patrones identificados en la nube de palabras de la Figura 4.3. Estas palabras clave revelan la manera en que los usuarios abordan temas de actualidad, como el consumo de tabaco y sus alternativas, reflejando intereses que abarcan desde el uso recreativo hasta la abstinencia y la salud pública.

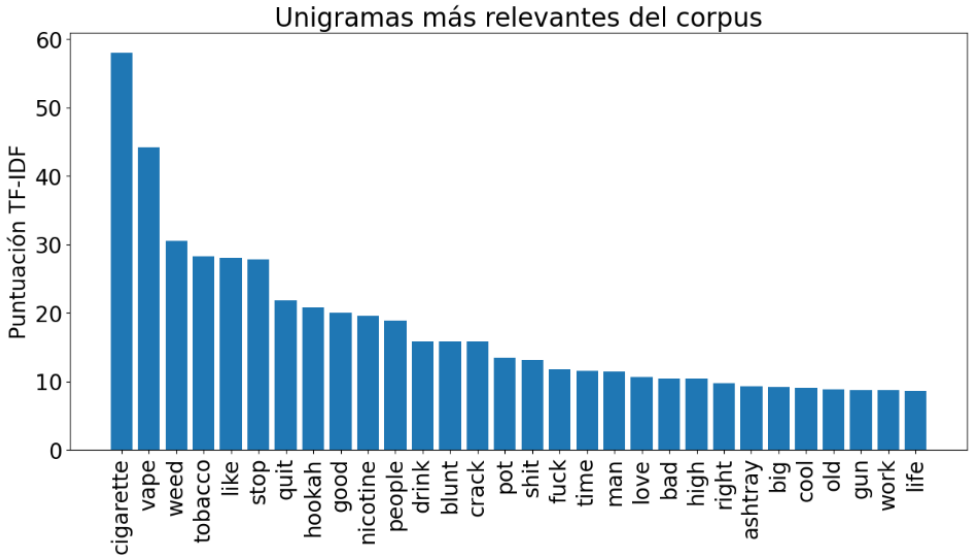

Figura 4.4: Unigramas más relevantes
Elaboración propia

Continuando con el análisis metodológico, se procedió a la aplicación de la métrica TF-IDF a cada bigrama del corpus, entendiendo por bigramas las secuencias de dos palabras consecutivas. Los resultados obtenidos, que se presentan en la Figura 4.5, destacan los 30 bigramas más relevantesn en el análisis. Estos bigramas se dividen principalmente en cuatro categorías distintas, reflejando preocupaciones sociales y la necesidad de políticas de salud pública efectivas. En la primera categoría, centrada en la acción de cesar el consumo, se identificaron bigramas como “stop weed”, “stop cigarette”, “stop pot”, “stop shit”, “ban tobacco”, “help stop” y “quit cigarette”. Estos términos destacan un esfuerzo consciente por parte de la

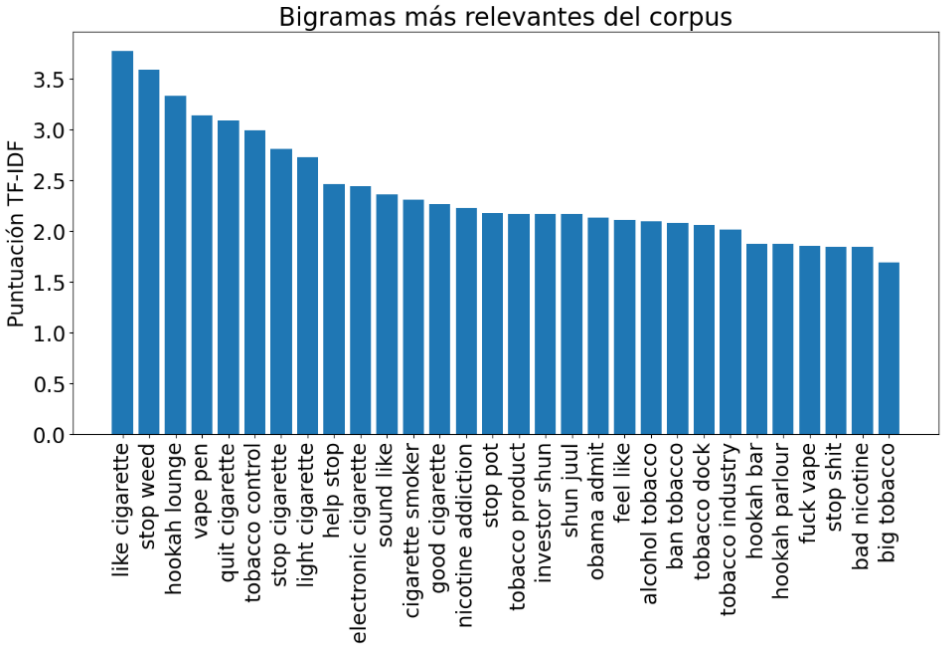

Figura 4.5: Bigramas más relevantes
Elaboración propia

población para terminar con el consumo de tabaco y otras sustancias, evidenciando la conciencia sobre los daños asociados al tabaquismo y el uso de drogas. La segunda categoría, enfocada en la adicción y el control del tabaco, resalta bigramas como “tobacco control”, “nicotine addiction”, y “bad nicotine”. Estos términos reflejan una comprensión de la dependencia al tabaco como una adicción que requiere intervención y políticas específicas para su control. Estos conceptos podrían estar asociados a un reconocimiento del tabaquismo como un problema de salud pública significativo que necesita una respuesta coordinada, apoyando los esfuerzos de la primera categoría.

En la tercera categoría, relacionada con el contexto social del consumo, se encuentran los bigramas “hookah bar” y “hookah parlour”. Estos reflejan la existencia de espacios sociales que permiten y fomentan el consumo de tabaco, presentándolo como una actividad socialmente aceptable o incluso deseable. La presencia de estos términos en el análisis podría indicar una dimensión social del consumo de tabaco que contribuye a su normalización y presenta desafíos adicionales para las políticas de salud pública destinadas a reducir su prevalencia. Finalmente, se introduce una cuarta categoría que, aunque compuesta por bigramas con connotaciones aparentemente positivas como “like cigarette”, “good cigarette”, y “wish cigarette”, destaca el hecho de que el consumo de tabaco sigue estando muy presente en la sociedad. Finalmente, se introduce una cuarta categoría que, aunque compuesta por bigramas con connotaciones aparentemente positivas, como “like cigarette”, “good cigarette”, y “wish cigarette”, podría indicar una persistente normalización del consumo de tabaco en la sociedad. Estos términos sugieren una complicada relación cultural con el tabaquismo, don-

de, a pesar de los conocidos riesgos para la salud, existen percepciones positivas que pueden dificultar los esfuerzos por reducir su prevalencia.

Realizando un análisis adicional sobre el corpus, centrado en la identificación de trigramas, se destacan los conjuntos de 3 palabras consecutivas más relevantes del conjunto de datos. Estos resultados, presentados en la Figura 4.6, revelan trigramas clave como “weed protest push”, “nightmare vivid stop”, “nicotine stan brave”, “bar hookah parlour”, y “aspire vape save”. Cada uno de estos trigramas profundiza en los temas identificados previamente a través de los bigramas, enfocándose en diferentes aspectos de las conversaciones públicas sobre el consumo de tabaco.

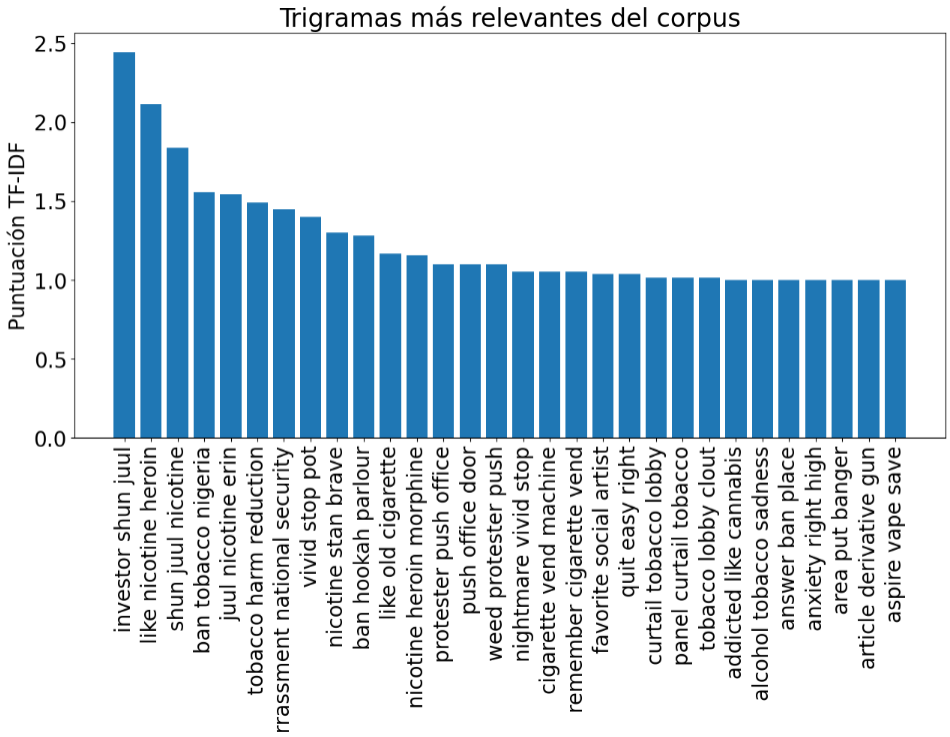

Figura 4.6: Trigramas más relevantes
Elaboración propia

Respecto a la primera categoría derivada del análisis de bigramas, que aborda la cesación del consumo, los trigramas “nightmare vivid stop” y “alcohol tobacco sadness” resaltan la complejidad emocional y de salud que implica la lucha contra estos productos. Estos trigramas muestran el impacto negativo que el consumo de tabaco y alcohol tiene en las personas, resaltando la urgencia de estrategias de apoyo eficaces para facilitar su abandono. De acuerdo a la segunda categoría de los bigramas, enfocada en la adicción y la necesidad de control regulatorio, el análisis de trigramas incluye conceptos como “addicted like cannabis” resaltando la naturaleza adictiva de estas sustancias. A continuación, “weed protest push” evidencia la demanda social por un control más riguroso sobre estos productos, destacando el paso de reconocer el problema a participar activamente en el debate y la acción pública.

La tercera categoría, relacionada con el contexto social del consumo, se ve caracterizada por el trigrama “ban hookah parlour”. Este término resalta, como se mencionó en la tercera categoría, la problemática de la normalización y promoción social del consumo de tabaco en espacios específicos y refuerza la necesidad de políticas públicas que limiten o prohíban estos entornos propicios para el consumo de tabaco. Finalmente, se discute la percepción y defensa del consumo de nicotina con trigramas como “nicotine stan brave” y “aspire vape save”, los cuales reflejan la persistencia de actitudes positivas hacia el consumo de tabaco y el uso de vaporizadores.

4.3. Modelado de las categorías de discusión

Tras completar la fase inicial de análisis descriptivo, centrada en identificar los conceptos clave desde la perspectiva de los usuarios, el estudio continúa con el modelado de tópicos, detallado en la sección 3.4 de este documento. Con el fin de implementar el modelo LDA, es necesario determinar el número óptimo de tópicos, denominado “k”, dentro del corpus para asegurar que el modelo sea tanto descriptivo como interpretable. Para lograr este objetivo, se realiza el cálculo del índice de coherencia evaluando diferentes cantidades de tópicos, desde dos hasta diez, como se visualiza en la Figura 4.7. Con el fin de garantizar que el modelo resultante sea interpretable y de alta calidad en cuanto a la representación de los temas intrínsecos del corpus analizado, se realiza un análisis para determinar el número de categorías que maximiza el índice de coherencia.

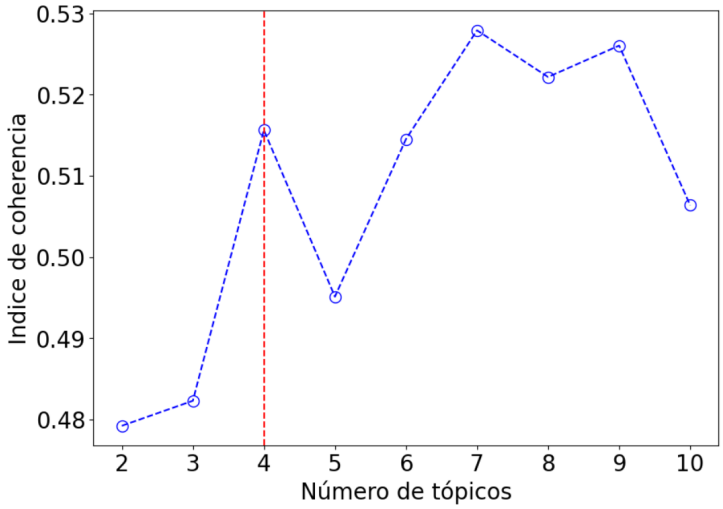

Figura 4.7: Relación entre el índice de coherencia y el número de tópicos
Elaboración propia

A pesar de que el máximo valor de coherencia se registra para $k = 7$, este indicador no confirma por sí solo que siete sea la cifra ideal de tópicos. La decisión sobre el número apropiado de tópicos requiere un balance entre la facilidad de interpretación del modelo y la

obtención de un índice de coherencia elevado, indicativo de la calidad y la relación significativa entre las palabras que conforman cada tópico. Un número reducido de tópicos podría simplificar demasiado el modelo, omitiendo información importante, mientras que un exceso en el número de tópicos podría generar solapamientos o redundancias, entorpeciendo la claridad del análisis. Se nota un incremento notable en el índice de coherencia al moverse de $k = 3$ a $k = 4$, señalando una mejora considerable en la organización y conexión entre las palabras dentro de los tópicos generados. Teniendo en cuenta este análisis y buscando un equilibrio entre menor número de tópicos y mayor índice de coherencia, se decide optar por $k = 4$ como el número óptimo de tópicos para este estudio. Con la determinación de este número óptimo, este estudio procede a explorar la estructura y la facilidad de interpretación de los tópicos desarrollados, enfocándose en analizar la distancia entre ellos para medir su diferenciación. Como se menciona en la Sección 3.4, una mayor distancia entre tópicos sugiere una mayor diferenciación y definición entre los mismos, lo cual facilita su comprensión y análisis. La presentación del mapa de distancias intertópicas para un modelo con $k = 4$, representado en la Figura 4.8, demuestra que los círculos representativos de cada tópico mantienen su individualidad. No Se observa ningún solapamiento entre los tópicos lo que significa que hay diferencias entre ellos, facilitando un análisis distintivo.

De esta manera, se ha ajustado un modelo con 4 tópicos representativos. Como resultado, la Tabla 4.1 ofrece un análisis detallado de cada tópico, identificando palabras clave, bigramas y trigramas específicos que contribuyen a comprender sus contenidos. La presencia reiterada de términos como “weed”, “vape”, “cigarette”,“tobacco” y “hookah” en todos los tópicos defiende la importancia de los productos relacionados con el tabaco dentro de las discusiones analizadas. Aunque estos términos aparezcan en todos los temas, cada uno de ellos aborda aspectos distintos y específicos relacionados con el consumo de tabaco. Esta repetición de palabras clave no implica directamente una homogeneidad temática entre los tópicos. Por el contrario, enfatiza cómo distintas dimensiones de un tema común pueden entrelazarse a través de las discusiones en línea.

En los resultados presentados en la Tabla 4.1 , el Tópico 1 aborda el **abandono de los productos de nicotina**, destacando en palabras clave como “cigarette”, “stop”, “quit” y “nicotine”. Los bigramas como “stop weed”, “help stop” y “quit cigarette” reflejan un esfuerzo por terminar con el hábito de fumar, subrayando los riesgos de salud asociados como lo evidencia el términos “lung cancer”. Además, los trigramas “easy quit cigarette” y “cause mouth cancer” resaltan, respectivamente, la percepción de facilidad para dejar de fumar y las graves consecuencias de salud asociadas al consumo. Por su parte, el Tópico 2 analiza las **posturas sobre el consumo de nicotina y alcohol**, con palabras clave como “drink”, “like”, “fuck” y “nicotine”. Los bigramas “fuck vape” y “ban hookah” reflejan una actitud crítica hacia estos productos y piden la prohibición de estos productos; mientras expresiones como “like cigarette” y “cigarette right” presentan un punto de vista más favorable. Los trigramas “ban hookah parlour” y “nicotine stan brave” amplian el significado de los bigramas, mostrando

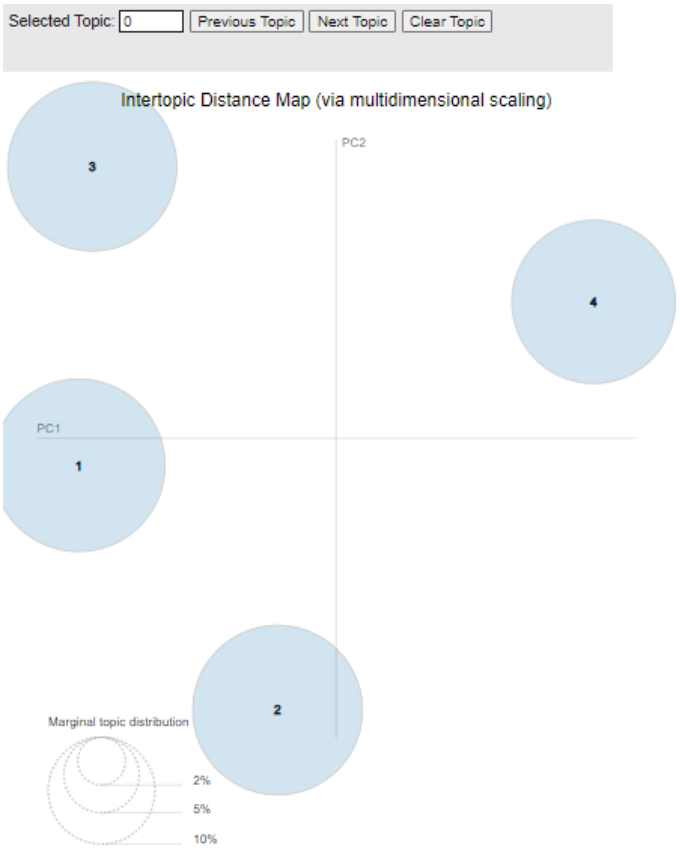

Figura 4.8: Distancia Intertópica para un modelo de 4 tópicos
Elaboración propia

también opiniones a favor y en contra: el primero pide mayores restricciones, mientras que el segundo apoya este consumo.

Respecto al Tópico 3, esta categoría se centra en **las dinámicas sociales relacionadas con el consumo del tabaco y la marihuana**. Las palabras "tobacco", "blunt", "weed", y "vape" hacen referencia a los productos más frecuentes de este ámbito. Como complemento, los bigramas "good company', "blunt sessions" y "bros match" muestran que el consumo de estos productos sucede en diversos entornos sociales. Sin embargo, bigramas como "asthma overexercising" resaltan también las posibles consecuencias negativas para la salud derivadas de fumar. Finalmente, los trigramas "blunt watch sunrise" y "finally vape happy" reflejan experiencias positivas, mientras que "nightmare vivid stop" habla de la superación de una mala experiencia previa. Por último, el Tópico 4 aborda la **regulación y el control del tabaco y la marihuana**, centrando la atención en las políticas gubernamentales y las reacciones públicas. Las palabras clave "gun", "shit", "nicotine", y "control" sugieren un enfoque crítico hacia estas sustancias. El bigrama "tobacco control" muestra la urgencia de limitar el tabaco, mientras que "nicotine addiction" destaca los desafíos de la dependencia. Los términos "weed protester" y "protester push" capturan el activismo y la resistencia pública.

Tabla 4.1: Tópicos principales junto con sus respectivos n-gramas más relevantes .

Tópico	Categoría	Palabras clave	Bigramas más relevantes	Trigramas más relevantes
1	Abandono de Productos de Nicotina	“cigarette”, “stop”, “quit”, y “nicotine”	“stop weed”, “help stop”, “quit cigarette”, “like cigarette”, “like nicotine” y bad nicotine	“easy quit cigarette”, “cause mouth cancer”,“burn bag cigarette” y “anxiety right high”
2	Opiniones sobre el Consumo de Nicotina y Alcohol	“drink”, “like”, “fuck” y “nicotine”	“fuck vape”, “ban hookah”,“like cigarette” y “cigarette right”	‘ban hookah parlour”, “nicotine stan brave”, “bad cigarette life” y “like revolutionary vape”
3	Dinámicas sociales relacionadas con el Consumo del Tabaco	“tobacco”, “blunt”, “weed”, y “vape”	“good company’, “blunt sessions” “bros match” y “asthma overexercising”	“blunt watch sunrise”,“finally vape happy”,“nightmare vivid stop”
4	Control y Regulación del Uso del Tabaco y la Marihuana	“gun”, “shit”, “nicotine”, y “control”	“tobacco control”, “nicotine addiction”, ’Weed protester’ y “protester push”	’like nicotine heroin”, “nicotine heroin morphine” ,“Handle problem weed” y “alcohol tobacco sadness”

Los trigramas “like nicotine heroin” y “nicotine heroin morphine” profundizan esta visión, comparando la nicotina con drogas duras, y “handle problem weed” enfatiza la necesidad de confrontar los problemas asociados al consumo de estos productos.

La Figura 4.9 muestra la distribución de los tópicos en el corpus, destacando el Tópico 1, “Abandono de Productos de Nicotina”, como el más prevalente con 486 menciones, centrado en las discusiones sobre el cese del consumo de nicotina. Los Tópicos 2 y 3, con un volumen igual de publicaciones, exploran respectivamente las “Opiniones sobre el Consumo de Nicotina y Alcohol”, abarcando desde críticas hasta percepciones favorables, y “Dinámicas sociales relacionadas con el Consumo del Tabaco”, que analiza la integración del tabaco y la marihuana en diversos estilos de vida, destacando tanto las percepciones negativas como los aspectos aceptados socialmente. El Tópico 4,“ Control y Regulación del Uso del Tabaco y la Marihuana”, aunque menos frecuente, destaca la importancia de las políticas de salud pública en la regulación de estos productos. Estos resultados muestran que, aunque predominan las críticas hacia el tabaquismo y sus alternativas, también hay discursos que valoran sus aspectos positivos, como su aceptación y uso en entornos sociales. Asimismo, se destaca una cantidad considerable de publicaciones enfocadas en el cese del consumo y en esfuerzos por controlar y regular el uso de estos productos. Este resultado destaca la existencia de un diálogo activo y multifacético sobre el tabaco y la marihuana, abarcando desde la preocupación por la salud hasta el reconocimiento de sus roles en ciertos contextos sociales.

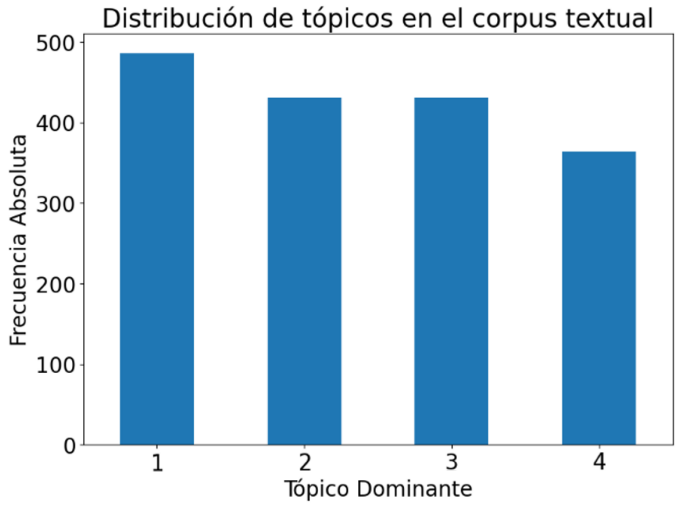

Figura 4.9: Frecuencia distribución de tópicos
Elaboración propia

4.4. Análisis de Sentimientos

El análisis de sentimiento es una herramienta analítica diseñada para identificar y cuantificar el tono emocional contenido en los textos. Este enfoque asigna una puntuación a cada fragmento de texto: valores negativos indican una tendencia hacia emociones desfavorables, valores positivos hacia emociones favorables, y puntuaciones alrededor de cero reflejan una neutralidad emocional. Específicamente, las puntuaciones varían en una escala donde -1 a -0.5 denotan emociones muy negativas, -0.5 a 0 ligeramente negativas, 0 es neutral, 0 a 0.5 ligeramente positivas, y 1 muy positivas, permitiendo una interpretación precisa del tono emocional en el texto analizado. En este estudio se utiliza VADER, un diccionario especializado en análisis de sentimientos descrito en el Capítulo 3. VADER incorpora elementos lingüísticos específicos de las discusiones en redes sociales, como el uso de signos de puntuación, emoticonos, mayúsculas y stopwords, para ofrecer una evaluación precisa de los sentimientos expresados en las publicaciones.

El análisis cuantitativo de las puntuaciones de sentimiento, tras el pre-procesamiento descrito en la Sección 3.5, se visualiza mediante un diagrama de caja en la Figura 4.10, donde se excluyen valores atípicos para mantener la integridad de la interpretación. La mediana se sitúa ligeramente por encima del cero, lo que indica una neutralidad general con una inclinación mínima hacia sentimientos positivos. El primer cuartil ubicado aproximadamente en -0.25 sugiere que los sentimientos negativos no son extremadamente marcados, mientras que el tercer cuartil, alrededor de 0.35, muestra que los sentimientos positivos son algo más frecuentes pero sin ser considerablemente dominantes. Estos resultados sugieren una distribución balanceada de emociones en el discurso relacionado con el tabaco, con una ligera prevalencia de reacciones positivas en las interacciones en redes sociales.

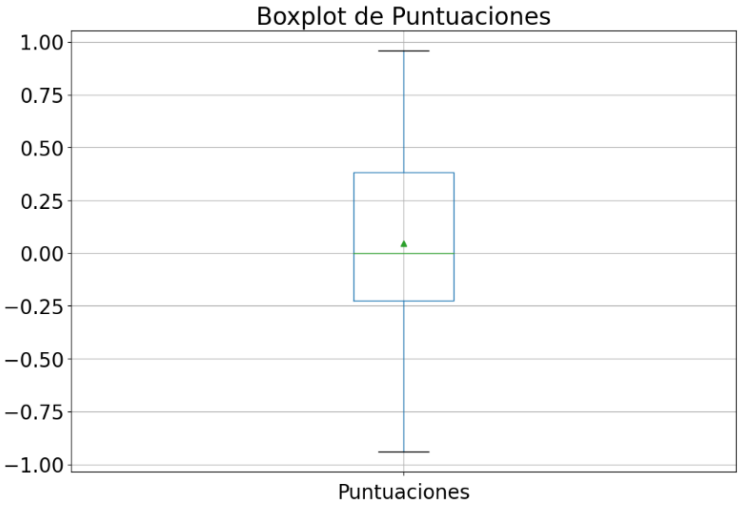

Figura 4.10: Distribución de la puntuación compuesta
Elaboración propia

Para complementar el estudio, se añade un histograma que se observa en la Figura 4.11. Este gráfico muestra una concentración significativa de publicaciones en torno al punto neutro, lo que corrobora una tendencia general a la neutralidad en la expresión de sentimientos. Aunque hay presencia tanto de sentimientos positivos como negativos, estos tienden a ser moderados en lugar de extremos, con un ligero predominio de los positivos. Esto indica que, mientras hay una variedad de emociones expresadas en el diálogo sobre el tabaco, la mayoría de las conversaciones mantienen un tono equilibrado.

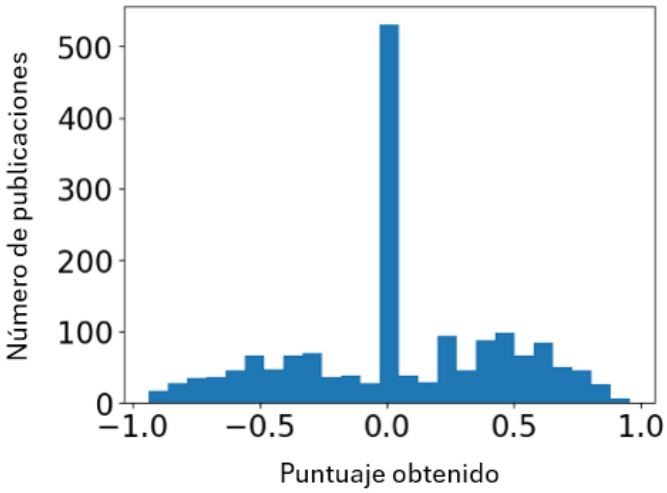

Figura 4.11: Número de publicaciones por puntuación
Elaboración propia

En resumen, el análisis de sentimientos revela una ligera preponderancia de sentimientos positivos, lo que podría indicar una disposición generalmente optimista entre los usuarios o una menor propensión a la negatividad en las discusiones sobre el tabaco. Cabe destacar que estos resultados reflejan las particularidades del conjunto de datos analizado y podrían diferir en otros contextos o con muestras diferentes.

Con el objetivo de comprender mejor las emociones expresadas en el conjunto de datos, se realiza un estudio donde se evalúa la puntuación compuesta lograda en los distintos tópicos identificados en la etapa de modelado de tópicos (Ver Figura 4.12). Sobre el análisis del tópico 1, que se enfoca en la idea de abandonar los productos de nicotina, el diagrama muestra que la mayoría de las opiniones se inclinan hacia una posición neutral o ligeramente positivas. Esta variación de opiniones podría deberse a que algunas personas consideran que dejar de fumar es posible y beneficioso, mientras que otras aún ven con buenos ojos continuar con el consumo. Además, se resalta que la preocupación por los daños a la salud que estos productos pueden causar juega un papel importante en cómo la gente aborda este tema. Sin embargo, que existan opiniones positivas o de aceptación muestra que todavía hay un debate abierto en la sociedad sobre si se debería o no dejar estos productos.

El análisis de sentimientos del tópico 2 aborda la dualidad de actitudes hacia el consumo de nicotina y alcohol, un tema que provoca tanto aceptación cultural como rechazo por su uso por motivos de salud. Los datos del diagrama revelan que la sociedad está notablemente dividida respecto a este tema, mostrando que, aunque existe una preocupación evidente por sus riesgos, la aceptación del consumo de tabaco sigue presente. La mediana, situada en el centro del diagrama, simboliza esta neutralidad colectiva y refleja como la sociedad pone en balance las consecuencias negativas del consumo y las tradiciones asociadas a su uso.

Al examinar el diagrama del Tópico 3,que borda el tema de las dinámicas sociales relacionadas con el consumo del tabaco, se observa que la mediana se desplaza ligeramente hacia el lado positivo del eje de puntuación,aunque predomina la presencia de emociones neutras. Esto indica una aceptación moderada en la cultura y el estilo de vida alrededor del consumo del tabaco y la marihuana, mostrando que, aunque estos productos no son buenos, su uso es ampliamente aceptado en ciertos grupos sociales. Resalta un interés de mantener el consumo a pesar de la existencia de malas experiencias por parte de algunos usuarios.

Finalmente, el Tópico 4 examina la distribución equilibrada de actitudes hacia el control y la regulación del consumo de tabaco y marihuana . La mediana, ubicada en el centro del diagrama, vuelve a mostrar una clara división en la opinión pública, que va desde el reconocimiento de la necesidad de control del tabaco hasta una aceptación de su uso en sociedad. Este equilibrio refleja tanto las acciones de grupos interesados como las protestas contra el consumo y sus adicciones. A pesar de la presión por una regulación más estricta, se percibe una corriente social que tolera y defiende la elección personal del consumo de estas sustancias. La neutralidad observada sugiere que la sociedad está aprendiendo a convivir con estos hábitos.

El análisis de sentimientos realizado revela una tendencia neutral hacia la opinión pública sobre el consumo de nicotina y tabaco, con un ligero sesgo hacia posturas positivas. Esta tendencia señala que las perspectivas moderadas prevalecen sobre los extremos en los debates relacionados con estos productos. El estudio subraya que la relación del consumo con los riesgos hacia la salud esta muy presente en la conversación, sin llegar a un rechazo absoluto del tabaco.Eso se respalda con las medias obtenidas en los diagramas de los cuatro tópicos, las cuales son 0.04695, 0.0728, 0.0300 y 0.0403. A pesar de que todas están cerca del de cero,valor neutral, el tópico 2 muestra una tendencia ligeramente más positiva, mientras que el tópico 3 es levemente más negativo.

En el caso del Tópico 1, las opiniones divididas sobre dejar de fumar indican que la sociedad está considerando cómo balancear el conocimiento de los riesgos para la salud con la elección personal de consumir tabaco. Por su parte, el Tópico 4 observa que, a pesar de la tendencia hacia una mayor regulación de tabaco y marihuana, no existe un consenso total a favor de estas políticas, lo que muestra un equilibrio entre el deseo de autonomía personal y la necesidad de control.

Respecto a los Tópicos 2 y 3, se percibe una división en cuanto a la aceptación social del

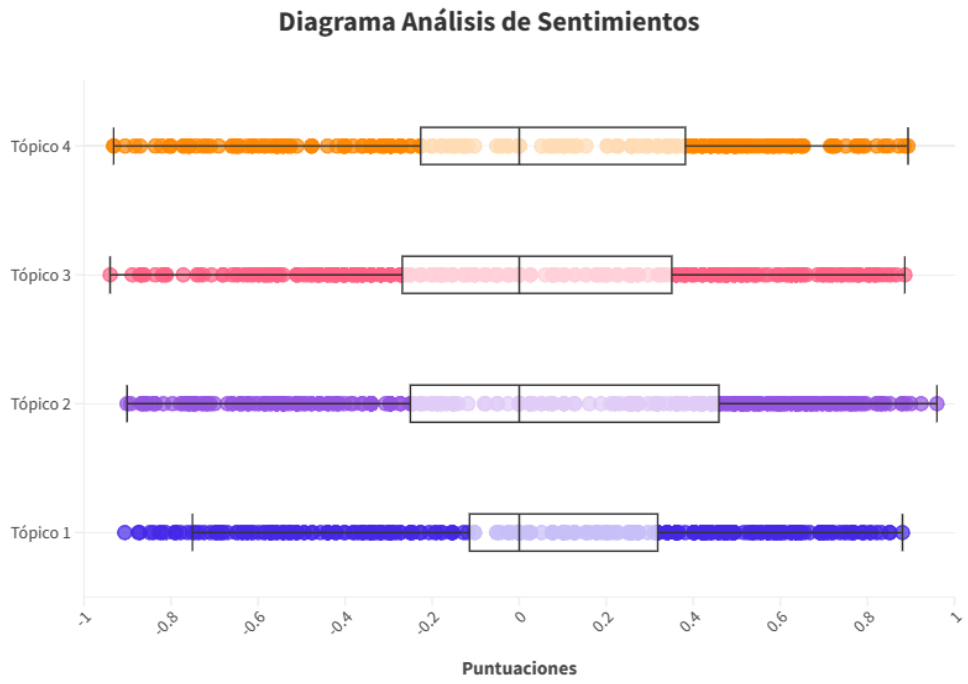

Figura 4.12: Puntuación del sentimiento para cada tópico
Elaboración propia

tabaco frente a los esfuerzos por promover la salud pública. Estos hallazgos revelan que, a pesar de la conciencia clara sobre los peligros para la salud, el hábito de fumar se mantiene como una costumbre culturalmente arraigada. En su totalidad, se aprecian opiniones heterogéneas, lo que demuestra que la sociedad no ha alcanzado un acuerdo unánime sobre la gestión del consumo de tabaco.

Capítulo 5

Conclusiones

La comprensión de las percepciones públicas sobre el tabaquismo es necesaria para el diseño de políticas de salud eficaces. En este sentido, las redes sociales surgen como fuentes de datos ricas y actualizadas, proporcionando información significativa sobre la opinión ciudadana. La plataforma X, en particular, ofrece acceso a discusiones y tendencias en tiempo real, presentándose como una herramienta de gran utilidad para autoridades sanitarias y formuladores de políticas. Este acceso a información diversa y continua posibilita el desarrollo de estrategias de intervención y educación alineadas con las necesidades y percepciones actuales de la sociedad.

En este contexto, el presente trabajo de fin de grado tiene como objetivo realizar un análisis de las publicaciones en X referentes al tabaco. Este análisis ha permitido identificar no solo las diversas perspectivas y posturas respecto al consumo de tabaco sino también los sentimientos y las conversaciones más extendidas en torno a esta sustancia.

Para abordar la relevancia de las percepciones públicas sobre el tabaquismo, se ha efectuado un proceso de revisión de literatura, poniendo especial atención en los estudios de análisis automático de opiniones. Se ha constatado que, si bien el modelado de tópicos usando LDA no ha sido ampliamente explorado en este ámbito, ha demostrado su utilidad al identificar de forma no supervisada las corrientes y tendencias subyacentes en las discusiones públicas. Este enfoque resalta por su sencillez, manejo eficiente de grandes volúmenes de datos y por ser menos demandante en recursos computacionales comparado con modelos más complejos. Respecto al análisis de sentimientos, se encontró que existen desde técnicas lexicográficas hasta métodos avanzados de ML, incluyendo redes neuronales. No obstante, los enfoques lexicográficos sobresalen por su capacidad para analizar rápidamente extensas cantidades de texto con una interpretación directa y clara. Por ello, en este trabajo, se ha preferido emplear el léxico VADER por su ajuste particular a las redes sociales, ofreciendo un análisis refinado de las emociones en las publicaciones.

Basándose en la revisión bibliográfica de este trabajo, se ha desarrollado una metodología compuesta por la preparación de datos, análisis descriptivo de N-Gramas, modelado de

tópicos y análisis de sentimientos. Este enfoque metodológico destaca por su aplicabilidad a otros temas de interés social y salud pública, lo que amplía la capacidad para comprender y responder a las dinámicas sociales en el contexto digital contemporáneo. El análisis descriptivo de N-Gramas muestra que términos como "cigarette", "tobacco", "vape" y "hookah" y "weed" son frecuentemente discutidos en el conjunto de datos, reflejando un interés activo en productos de nicotina y sustancias similares. Posteriormente, la obtención de los 30 bigramas más relevantes permitió clasificar los datos en cuatro categorías: la primera centrada en la acción de cesar el consumo, la segunda enfocada en la adicción y el control del tabaco, la tercera relacionada con el contexto social del consumo, y la cuarta abordando la normalización del consumo de tabaco en la sociedad. Este análisis proporciona una base sólida para la segmentación en el modelado de tópicos y ayuda a identificar las principales áreas de preocupación y discusión dentro de la comunidad en línea. En particular, el modelado de tópicos ha identificado cuatro tópicos principales sobre el consumo de tabaco y productos nocivos en la plataforma X. El primer tópico aborda el abandono de los productos de nicotina, donde se refleja el uso de la plataforma como medio para buscar apoyo y estrategias para dejar de fumar. El segundo tópico explora las posturas y opiniones sobre el consumo de nicotina y alcohol, destacando tanto las críticas hacia estas prácticas como las opiniones que valoran tanto el producto como la libertad individual. El tercer tópico se centra en las dinámicas sociales relacionadas con el consumo de tabaco y marihuana, resaltando que este consumo es parte de la interacción social y está influenciado por el entorno. Por último, el cuarto tópico se concentra en la regulación y control de estas sustancias, mostrando una amplia gama de reacciones públicas, desde apoyo hasta oposición, con un enfoque general hacia la mejora de la regulación y control. Por otra parte, el análisis de sentimientos muestra que, en las publicaciones sobre el consumo de tabaco y productos similares, la mayoría de los usuarios se posicionan en un rango emocional neutral con un leve sesgo positivo. Esto indica que, aunque hay una conciencia general sobre los riesgos para la salud, como las adicciones y enfermedades graves como el cáncer, existe una aceptación moderada de estos hábitos dentro de la cultura social. Así, los datos muestran una distribución equilibrada de sentimientos negativos y positivos. Este análisis se respalda con las medias obtenidas en los diagramas de los cuatro tópicos, las cuales son 0.04695, 0.0728, 0.0300 y 0.0403. Aunque todas están cerca del valor neutral de cero, el tópico 2 muestra una tendencia ligeramente más positiva, mientras que el tópico 3 es levemente más negativo.

Para futuros trabajos, se recomienda analizar las publicaciones sobre tabaco segmentándolas por género y ubicación geográfica. Esto permitiría una comprensión más profunda de las diferencias en percepciones y comportamientos, apoyando el desarrollo de políticas públicas más específicas y efectivas. Además, sería valioso investigar la posible utilización de perfiles ocultos o ficticios por parte de las tabacaleras para evadir la prohibición de publicidad en redes sociales, vigente en la Unión Europea desde 1987 (Europea, 2024). Comprender estos métodos podría fortalecer el control y asegurar el cumplimiento de las regulaciones sobre

publicidad de tabaco.

Declaración de Uso de Herramientas de Inteligencia Artificial Generativa en Trabajos Fin de Grado

ADVERTENCIA: Desde la Universidad consideramos que *ChatGPT* u otras herramientas similares son herramientas muy útiles en la vida académica, aunque su uso queda siempre bajo la responsabilidad del alumno, puesto que las respuestas que proporciona pueden no ser veraces. En este sentido, NO está permitido su uso en la elaboración del Trabajo fin de Grado para generar código porque estas herramientas no son fiables en esa tarea. Aunque el código funcione, no hay garantías de que metodológicamente sea correcto, y es altamente probable que no lo sea.

Por la presente, yo, Sofía Mazarío Olivé, estudiante de Doble Grado en ADE y Business Analytics (E-2 + Analytics) de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado "El Tabaco en la Era Digital: Explorando la Percepción Ciudadana a Través del Análisis de Publicaciones", declaro que he utilizado la herramienta de Inteligencia Artificial Generativa *ChatGPT* u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación:

1. Interpretador de código: Para realizar análisis de datos preliminares.
2. Corrector de estilo literario y de lenguaje: Para mejorar la calidad lingüística y estilística del texto.
3. Revisor: Para recibir sugerencias sobre cómo mejorar y perfeccionar el trabajo con diferentes niveles de exigencia.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado *ChatGPT* u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: Abril 2024

Firma: *S.M.O* (Sofía Mazarío Olivé)

Referencias

AECC. (2022). *El consumo de tabaco en españa y el mundo, en datos y gráficos.* (acceso Octubre 20, 2023) https://www.epdata.es/datos/consumo-tabaco-espana-datos-graficos/377#:~:text=Casi%20en%20las%20%C3%BAltimas%20dos,20%2C4%25%20en%202025..

AECC. (2023). *Relación entre tabaco y cáncer: entre el 80 % y 90 % de los cánceres de pulmón se dan en fumadores o ex fumadores.* (acceso Octubre 20, 2023) https://blog.contraelcancer.es/relacion-tabaco.

Bian, J., Yoshigoe, K., Hicks, A., Yuan, J., He, Z., Xie, M., … Modave, F. (2016). Mining twitter to assess the public perception of the "internet of things". *PloS one*, *11*(7), e0158450.

Blei, D. M., Ng, A. Y., y Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Clark, E. M., Jones, C., Gaalema, D., White, T., Redner, R., Everett, R., … Danforth, C. (2014). Social media meets population health: a sentiment and demographic analysis of tobacco and e-cigarette use across the "twittersphere". *Value in Health*, *17*(7), A603.

Cobb, N. K., Mays, D., y Graham, A. L. (2013). Sentiment analysis to determine the impact of online messages on smokers' choices to use varenicline. *Journal of the National Cancer Institute Monographs*, *2013*(47), 224–230.

Dai, H. (2020). Heated tobacco product use and associated factors among us youth, 2019. *Drug and alcohol dependence*, *214*, 108150.

Dai, X., Gakidou, E., y Lopez, A. D. (2022). *Evolution of the global smoking epidemic over the past half century: strengthening the evidence base for policy action* (Vol. 31). doi: 10.1136/tobaccocontrol-2021-056535

de Hacienda y Función Pública, M. (2022). *EstadÍsticas mercado de tabacos 2022.* (acceso Octubre 20, 2023) https://www.hacienda.gob.es/es-ES/Areas%20Tematicas/CMTabacos/Paginas/EstadisticasCMT2022.aspx.

del corazón, F. (2022). *Radiografía de la mortalidad por tabaquismo en españa.* (acceso Octubre 20, 2023) https://fundaciondelcorazon.com/prensa/notas-de-prensa/3827-el-tabaco-mata-a-mas-de-14-000-espanoles-al-ano-por-enfermedades-cardiovasculares.html#:~:text=Diferencias%

20por%20sexo%20y%20CC.&text=El%20estudio%20refleja%20que%2C%20del,Cantabria%20(24%2C3%25)..

Dobbs, P. D., Boykin, A. A., Ezike, N., Myers, A. J., Colditz, J. B., y Primack, B. A. (2023). Twitter sentiment about the us federal tobacco 21 law: Mixed methods analysis. *JMIR Formative Research*, *7*(1), e50346.

Elmitwalli, S., Mehegan, J., Wellock, G., Gallagher, A., y Gilmore, A. (2024). Topic prediction for tobacco control based on cop9 tweets using machine learning techniques. *Plos one*, *19*(2), e0298298.

Europea, C. (2024). *Prohibición de la publicidad y el patrocinio transfronterizos del tabaco.* (acceso Abril, 2024) https://health.ec.europa.eu/tobacco/ban-cross-border-tobacco-advertising-and-sponsorship_es#:~:text=La%20publicidad%20y%20el%20patrocinio%20del%20tabaco%20en%20televisi%C3%B3n%20se,89%2F552%2FCEE)%20.

Godea, A. K., Caragea, C., Bulgarov, F. A., y Ramisetty-Mikler, S. (2015). An analysis of twitter data on ecigarette sentiments and promotion. , 205–215.

Hamdy, N., y Gomaa, E. H. (2012). Framing the egyptian uprising in arabic language newspapers and social media. *Journal of Communication*, *62*(2), 195–211.

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).

INE. (2021). *Sin tabaco.* (acceso Octubre 20, 2023) https://www.ine.es/infografias/infografia_tabaco.pdf.

INE. (2022). *El impacto económico y social del tabaquismo.* (acceso Octubre 20, 2023) https://ine.es/ss/Satellite?c=INESeccion_C&cid=1259944493195&p=1254735110672&pagename=ProductosYServicios%2FPYSLayout&param1=PYSDetalleFichaIndicador&param3=1259937499084.

Kamiński, M., Muth, A., y Bogdański, P. (2020). Smoking, vaping, and tobacco industry during covid-19 pandemic: Twitter data analysis. *Cyberpsychology, Behavior, and Social Networking*, *23*(12), 811–817.

Kubin, E., y von Sikorski, C. (2021). The role of (social) media in political polarization: a systematic review. *Annals of the International Communication Association*, *45*(3), 188–206.

Kumari, S., Sharma, A., Chhabra, A., Gupta, A., Singh, S., y Verma, R. (2024). Analysing public sentiment towards robotic surgery: an x (formerly twitter) based study. *Social Network Analysis and Mining*, *14*(1), 1–18.

Lee, F. L. (2016). Impact of social media on opinion polarization in varying times. *Communication and the Public*, *1*(1), 56–71.

Mazarío.S. (2024). *Mazarío.s.* (acceso Abril, 2024) https://github.com/SofiaMazario/SofiaMazario.

Myslín, M., Zhu, S.-H., Chapman, W., Conway, M., y cols. (2013). Using twitter to examine smoking behavior and perceptions of emerging tobacco products. *Journal of medical Internet research*, *15*(8), e2534.

OMS. (2022). *El consumo de tabaco en españa y el mundo, en datos y gráficos.* (acceso Octubre 20, 2023) https://www.epdata.es/datos/consumo-tabaco-espana-datos-graficos/377#:~:text=Casi%20en%20las%20%C3%BAltimas%20dos,20%2C4%25%20en%202025..

OMS. (2023). *El impacto económico y social del tabaquismo.* (acceso Octubre 20, 2023) https://www.who.int/es/news-room/fact-sheets/detail/tobacco.

Resende, E. C., y Culotta, A. (2015). A demographic and sentiment analysis of e-cigarette messages on twitter. *Computer Science Department, Illinois Institute of Technology*, 2020–07.

SEPAR. (2023). *Tabaquismo.* (acceso Octubre 20, 2023) https://www.separ.es/node/882.

Shah, A., Shah, S., Rand, B., y Champon, X. (2024). The celebrity factor: Exploring the impact of influencers on covid-19 vaccine sentiment through bayesian modeling of time series. *The Journal of the Southern Association for Information Systems*, *11*(1), 31–52.

Sidani, J. E., Colditz, J. B., Barrett, E. L., Chu, K.-H., James, A. E., y Primack, B. A. (2020). Juul on twitter: analyzing tweets about use of a new nicotine delivery system. *Journal of School Health*, *90*(2), 135–142.

Sievert, C., y Shirley, K. (2014). Ldavis: A method for visualizing and interpreting topics. En *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63–70).

Smokpro. (2021). *Smokpro: Towards tobacco product identification in social media text.* (acceso Octubre 20, 2023) https://github.com/himakaryv/smokpro-tobacco-product-classification/blob/main/gold_corpus.csv.

Statista. (2023). *Previsión del número de usuarios mensuales activos (mau) de x (twitter) a nivel mundial desde 2021 hasta 2024.* (acceso Noviembre 18, 2023) https://es.statista.com/estadisticas/636174/numero-de-usuarios-mensuales-activos-de-twitter-en-el-mundo/.

X. (2023). *Acerca de x premium.* (acceso Octubre 20, 2023) https://help.twitter.com/es/using-x/x-premium#tbcost.

Yach, D., y Bettcher, D. (2000). Globalisation of tobacco industry influence and new global responses. *Tobacco control*, *9*(2), 206.

Yanamandra, V. H., Pant, K., y Mamidi, R. (2020). Smokpro: Towards tobacco product identification in social media text. *SIIRH@ ECIR*.