



Facultad de Ciencias Económicas y Empresariales

Análisis de Opiniones sobre el Aprendizaje Online en Comunidades Digitales

Autor: Pablo Soriano Fernández

Director: Jenny Alexandra Cifuentes Quintero

MADRID | Abril 2026

Resumen

El aprendizaje *online* se ha consolidado como una de las principales transformaciones del sistema educativo global, especialmente tras la pandemia de COVID-19, que forzó a más de 1.600 millones de estudiantes en 190 países a migrar hacia entornos virtuales de enseñanza. Este fenómeno ha venido acompañado de un crecimiento sostenido del sector del *e-learning*, cuyo valor de mercado alcanzó los 314 mil millones de dólares en 2024 y se estima que superará los 614 mil millones en 2029. Sin embargo, la eficacia de esta modalidad educativa no depende únicamente del formato digital, sino de múltiples factores pedagógicos, tecnológicos y personales que inciden directamente en la experiencia del estudiante. Si bien los estudios previos han contribuido a identificar factores relevantes mediante encuestas estructuradas y modelos teóricos consolidados, estos enfoques pueden verse limitados por sesgos de deseabilidad social o por restricciones en la formulación de respuestas.

En este contexto, las comunidades digitales emergen como espacios privilegiados para capturar percepciones espontáneas y no solicitadas sobre fenómenos educativos. Reddit, con más de 91 millones de usuarios activos diarios y una estructura organizada en comunidades temáticas, constituye una fuente de especial relevancia para el análisis de valoraciones, preocupaciones y expectativas expresadas en lenguaje natural. El presente Trabajo de Fin de Grado propone un análisis de las percepciones de los usuarios de Reddit en torno al aprendizaje *online*, aplicando técnicas de procesamiento de lenguaje natural sobre un corpus de publicaciones extraídas mediante la API oficial de la plataforma. La metodología adoptada combina el modelado de tópicos mediante *Latent Dirichlet Allocation* (LDA) para identificar los ejes temáticos predominantes en la discusión, y el análisis de sentimientos a través de VADER para evaluar la polaridad afectiva de las publicaciones.

En el modelado de tópicos se identifican 3 tópicos diferentes dentro del corpus de publicaciones, cuyas temáticas son: “Educación *Online* y Desarrollo Profesional”, “Experiencias Personales y Bienestar Emocional” y “Exámenes y Evaluación *Online*”. Cada uno de estos tópicos aborda aspectos relevantes de la experiencia de aprendizaje virtual, desde el desarrollo de competencias técnicas y profesionales hasta el impacto emocional y las preocupaciones específicas asociadas a la evaluación remota. Además, el análisis de n-gramas destaca la salud mental como una de las preocupaciones centrales del discurso, con el bigrama *mental health* ocupando la primera posición en términos de relevancia. El análisis de sentimientos

realizado revela una orientación mayoritariamente positiva, aunque con diferencias notables entre tópicos: las discusiones sobre educación formal y desarrollo profesional presentan un tono marcadamente positivo, mientras que las experiencias personales y de bienestar emocional exhiben mayor variabilidad y una puntuación media inferior, reflejando la diversidad de vivencias compartidas por los usuarios.

Palabras Clave: Aprendizaje *Online*, *Reddit*, Percepción Estudiantil, Análisis de Sentimientos, Modelado de Tópicos.

Abstract

Online learning has become one of the major transformations of the global education system, especially following the COVID-19 pandemic, which forced more than 1.6 billion students across 190 countries to migrate to virtual learning environments. This phenomenon has been accompanied by sustained growth in the e-learning sector, whose market value reached 314 billion dollars in 2024 and is estimated to exceed 614 billion by 2029. However, the effectiveness of this educational modality does not depend solely on the digital format, but on multiple pedagogical, technological, and personal factors that directly affect the student experience. While previous studies have contributed to identifying relevant factors through structured surveys and established theoretical models, these approaches may be limited by social desirability biases or constraints in response formulation.

In this context, digital communities emerge as privileged spaces for capturing spontaneous and unsolicited perceptions about educational phenomena. Reddit, with more than 91 million daily active users and a structure organized into thematic communities, constitutes a particularly relevant source for the analysis of opinions, concerns, and expectations expressed in natural language. This Bachelor's Thesis proposes an analysis of Reddit users' perceptions regarding online learning, applying natural language processing techniques to a corpus of posts extracted through the platform's official API. The methodology combines topic modeling through *Latent Dirichlet Allocation* (LDA) to identify the predominant thematic axes in the discussion, and sentiment analysis using VADER to evaluate the affective polarity of the posts.

The topic modeling identifies 3 different topics within the corpus of posts, whose themes are: "Online Education and Professional Development", "Personal Experiences and Emotional Well-being", and "Online Exams and Assessment". Each of these topics addresses relevant aspects of the virtual learning experience, from the development of technical and professional competencies to the emotional impact and specific concerns associated with remote assessment. Furthermore, the n-gram analysis highlights mental health as one of the central concerns in the discourse, with the bigram mental health ranking first in terms of relevance. The sentiment analysis reveals a predominantly positive orientation, although with notable differences between topics: discussions about formal education and professional development present a markedly positive tone, while personal experiences and emotional well-being

exhibit greater variability and a lower average score, reflecting the diversity of experiences shared by users.

Keywords: Online Learning, Reddit, Student Perception, Sentiment Analysis, Topic Modeling.

Índice general

1. Introducción	1
1.1. Motivación	1
1.2. Objetivos	5
1.2.1. Objetivo General	5
1.2.2. Objetivos Específicos	5
1.3. Estructura del documento	6
2. Percepción del Aprendizaje Online a través de Redes Sociales: Revisión de la Literatura	7
3. Metodología de Análisis de Datos	13
3.1. Adquisición y Preparación de los Datos	14
3.2. Pre-procesamiento de los datos	15
3.3. Exploración de N-Gramas	16
3.4. Modelado de Tópicos	18
3.5. Análisis de Sentimientos	19
4. Resultados	20
4.1. Adquisición y Preparación de los Datos	20
4.2. Pre-procesamiento de los datos	24
4.3. Análisis de N-Gramas	25
4.4. Modelado de Tópicos	28
4.5. Análisis de Sentimientos	33
5. Conclusiones	39
Bibliografía	45

Índice de figuras

1.1. Crecimiento proyectado del mercado global de <i>e-learning</i> (2024–2029). Fuente: (The Business Research Company, 2024). Elaboración propia.	2
1.2. Formatos de aprendizaje preferidos por los ciudadanos de Reino Unido. Fuente: (CIPHR, 2022). Elaboración propia.	2
1.3. Crecimiento global de usuarios activos en Reddit (2018-2026). Fuente: (Business of Apps, 2024). Elaboración propia.	5
3.1. Fases de la Metodología. Elaboración propia.	14
3.2. Representación del modelo LDA. Fuente: Journal of Machine Learning Research. (Blei, Ng, y Jordan, 2003). Elaboración propia.	18
4.1. Número de publicaciones de Reddit por año. Elaboración propia.	21
4.2. Número de publicaciones por idioma. Elaboración propia.	21
4.3. Top subreddits. Elaboración propia.	22
4.4. Top hashtags. Elaboración propia.	23
4.5. Top emojis. Elaboración propia.	23
4.6. Nube de palabras del corpus pre-procesado. Elaboración propia.	25
4.7. Unigramas más relevantes. Elaboración propia.	26
4.8. Bigramas más relevantes. Elaboración propia.	27
4.9. Trigramas más relevantes. Elaboración propia.	28
4.10. Índice de coherencia por número de tópicos. Elaboración propia.	29
4.11. Distancia intertópica $K=3$. Elaboración propia.	30
4.12. Distribución de publicaciones por tópico. Elaboración propia.	32
4.13. Distribución de las puntuaciones de sentimiento en el corpus. Elaboración propia.	33
4.14. Distribución de la puntuación compuesta. Elaboración propia.	34
4.15. Diagrama de caja de las puntuaciones de sentimiento del corpus. Elaboración propia.	36
4.16. Evolución temporal del sentimiento por tópico (media móvil 7 días). Elaboración propia.	37

Índice de tablas

2.1. Resumen de estudios previos relacionados.	12
4.1. Principales palabras clave, bigramas y trigramas por tópico.	32
4.2. Estadísticos descriptivos de las puntuaciones de sentimiento por tópico. Elaboración propia.	35

Acrónimos

<i>API</i>	Application Programming Interface
<i>Bi-LSTM</i>	Bidirectional Long Short-Term Memory
<i>BORU</i>	Best of Redditor Updates
<i>BoW</i>	Bag of Words
<i>BTM</i>	Biterm Topic Model
<i>CAGR</i>	Compound Annual Growth Rate
<i>CNN</i>	Convolutional Neural Network
<i>COVID-19</i>	Coronavirus Disease 2019
<i>DLDA</i>	Dynamic Latent Dirichlet Allocation
<i>DT</i>	Decision Tree
<i>IDF</i>	Inverse Document Frequency
<i>LDA</i>	Latent Dirichlet Allocation
<i>LSA</i>	Latent Semantic Analysis
<i>LSTM</i>	Long Short-Term Memory
<i>ML</i>	Machine Learning
<i>NB</i>	Naïve Bayes
<i>NLP</i>	Natural Language Processing
<i>PRAW</i>	Python Reddit API Wrapper
<i>RF</i>	Random Forest
<i>SMOTE</i>	Synthetic Minority Over-sampling Technique
<i>STM</i>	Structural Topic Model
<i>SVM</i>	Support Vector Machine
<i>TF</i>	Term Frequency
<i>TF-IDF</i>	Term Frequency–Inverse Document Frequency
<i>TFG</i>	Trabajo de Fin de Grado
<i>URL</i>	Uniform Resource Locator
<i>VADER</i>	Valence Aware Dictionary and sEntiment Reasoner

Capítulo 1

Introducción

1.1. Motivación

El aprendizaje *online* ha adquirido una relevancia creciente en los últimos años, consolidándose como una de las principales transformaciones del sistema educativo global. Este fenómeno se intensificó de forma abrupta a raíz de la pandemia de *COVID-19* (*Coronavirus Disease 2019*), cuando más de 1.600 millones de estudiantes en 190 países se vieron afectados por el cierre de instituciones educativas, lo que forzó una migración inmediata hacia entornos virtuales de enseñanza y aprendizaje (UNESCO, 2020). Esta digitalización repentina no solo permitió garantizar la continuidad educativa, sino que también mostró el potencial del aprendizaje *online* para fomentar modelos pedagógicos más flexibles, inclusivos e innovadores, capaces de adaptarse a las nuevas necesidades sociales y tecnológicas.

Desde una perspectiva económica, este proceso de transformación ha venido acompañado por un notable crecimiento del sector del *e-learning*, que se ha posicionado como uno de los mercados más dinámicos a nivel global. Tal como se muestra en la Figura 1.1, el valor del mercado mundial de aprendizaje *online* alcanzó los 314 mil millones de dólares en 2024 y se estima que superará los 614 mil millones en 2029, con una tasa de crecimiento anual compuesta (CAGR) del 14,9%. Esta evolución proyectada, que prácticamente duplicará el tamaño del mercado en apenas cinco años, refleja la consolidación del aprendizaje digital como un componente estructural y sostenible del sistema educativo contemporáneo. El incremento sostenido responde, entre otros factores, a la creciente demanda de formación continua, la expansión de modelos híbridos de enseñanza y la necesidad de desarrollar competencias digitales en un entorno laboral cada vez más automatizado y globalizado.

En línea con esta expansión global del mercado, las preferencias formativas también reflejan una creciente inclinación hacia el aprendizaje *online* en distintos contextos nacionales. Por ejemplo, en el Reino Unido, una encuesta realizada por *Censuswide* para CIPHR (empresa británica especializada en soluciones de gestión de recursos humanos) en enero de 2022 a una muestra de 2.000 adultos examinó los métodos de aprendizaje preferidos en ámbitos

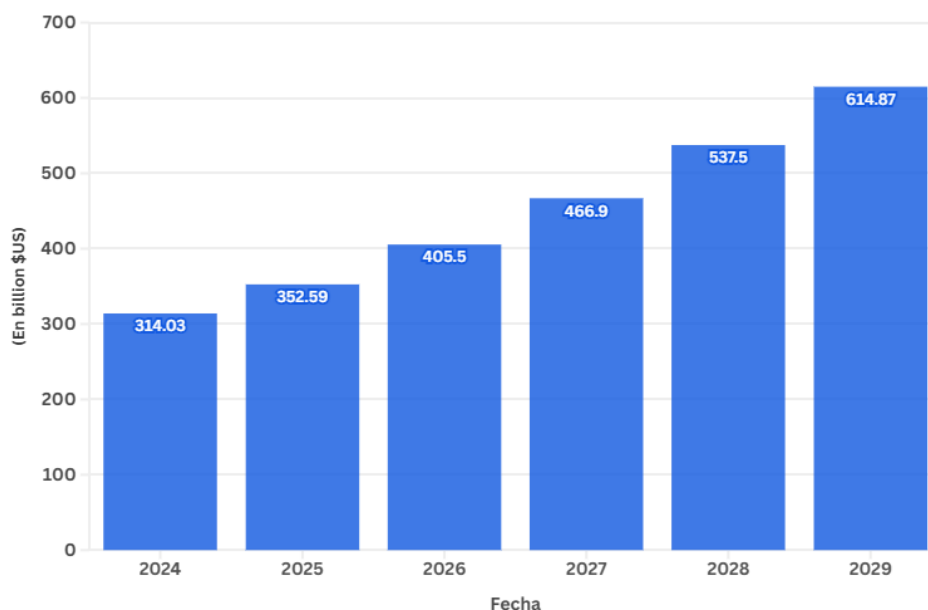


Figura 1.1: Crecimiento proyectado del mercado global de *e-learning* (2024–2029). Fuente: (The Business Research Company, 2024). Elaboración propia.

laborales, educativos y de ocio. Como se observa en la Figura 1.2, los vídeos educativos *online* y los cursos en línea fueron las modalidades más valoradas, superando incluso a las clases presenciales y a las tutorías individuales. Esta distribución muestra un claro desplazamiento hacia formatos digitales de autoaprendizaje, en los que la flexibilidad, la autonomía y el acceso asincrónico al contenido son percibidos como ventajas relevantes. Además, la popularidad de los vídeos y materiales audiovisuales apunta a una preferencia creciente por recursos que combinan inmediatez, claridad visual y facilidad de consumo, consolidando así el aprendizaje digital como una práctica integrada en múltiples esferas de la vida cotidiana.

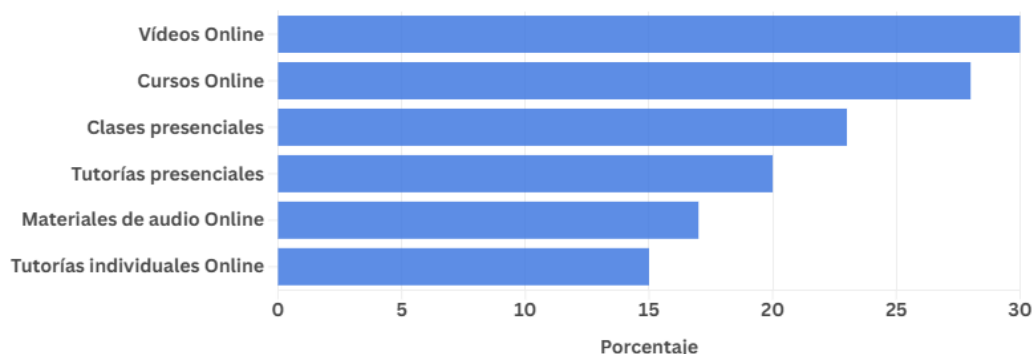


Figura 1.2: Formatos de aprendizaje preferidos por los ciudadanos de Reino Unido. Fuente: (CIPHR, 2022). Elaboración propia.

No obstante, a pesar de su crecimiento sostenido y de su creciente aceptación social, el aprendizaje *online* continúa generando debate en torno a su efectividad y a su impacto so-

bre la experiencia del estudiante. Una revisión rigurosa de la literatura previa sugiere que su eficacia no depende únicamente del formato digital, sino de múltiples factores pedagógicos, tecnológicos y personales. El meta-análisis realizado por (Means, Toyama, Murphy, Bakia, y Jones, 2009) para el Departamento de Educación de Estados Unidos recopiló más de mil estudios empíricos publicados entre 1996 y 2008, comparando el rendimiento académico en entornos presenciales, híbridos y completamente *online*. Sus resultados mostraron que los entornos virtuales pueden generar un rendimiento ligeramente superior al de la enseñanza tradicional, siempre que integren metodologías activas y materiales interactivos. Sin embargo, esta ventaja no es universal. Tal como señala (Dhawan, 2020) en su análisis del contexto universitario indio durante la pandemia, la eficacia del aprendizaje *online* depende en gran medida de variables humanas y contextuales como la autonomía del estudiante, la estabilidad de las plataformas tecnológicas y la formación del profesorado. Estos factores pueden actuar como potenciadores o limitantes de la calidad pedagógica, afectando directamente a la motivación y al compromiso del alumnado.

En consecuencia, las percepciones estudiantiles sobre la educación *online* no son homogéneas, sino que varían según el contexto y las condiciones específicas de implementación. En un estudio comparativo llevado a cabo en Sudáfrica, Hungría y el Reino Unido, (Cranfield, Tick, Venter, Blignaut, y Renaud, 2021) analizaron las respuestas de más de 900 estudiantes universitarios para identificar los factores que inciden en su nivel de satisfacción con la enseñanza digital. Los resultados evidenciaron que la accesibilidad tecnológica, la autonomía en el aprendizaje y la claridad comunicativa del profesorado fueron determinantes para una valoración positiva, mientras que la falta de interacción social y las distracciones del entorno doméstico emergieron como obstáculos frecuentes. En una línea complementaria, ElSayad (2024) adaptó y amplió el modelo teórico *Community of Inquiry*, incorporando el concepto de *learning presence*, entendido como la combinación de autorregulación e implicación emocional del estudiante. Aplicado a una muestra de 198 universitarios egipcios, este enfoque demostró que el aprendizaje *online* significativo exige no solo presencia docente y social, sino también un compromiso activo por parte del estudiante en la construcción de su propio proceso formativo.

Considerando que la eficacia del aprendizaje *online* no puede entenderse al margen de las percepciones, motivaciones y condiciones del propio estudiantado, resulta importante acceder a fuentes que recojan dichas experiencias con riqueza contextual y espontaneidad expresiva. Si bien los estudios previos han contribuido a identificar factores relevantes mediante encuestas estructuradas y modelos teóricos consolidados, estos enfoques pueden verse limitados por sesgos de deseabilidad social o por restricciones en la formulación de respuestas. En este contexto, las comunidades digitales emergen como espacios privilegiados para la exploración de percepciones reales y no solicitadas. En particular, las redes sociales se han consolidado como una de las principales fuentes para analizar percepciones y actitudes colectivas. Estas plataformas permiten acceder a opiniones expresadas de manera espontánea, sin

las limitaciones impuestas por instrumentos tradicionales como encuestas o entrevistas estructuradas. Su naturaleza abierta, dinámica y global las convierte en un escenario privilegiado para observar cómo las personas discuten, evalúan y reaccionan ante distintos fenómenos sociales y educativos en tiempo real.

Entre estas plataformas, Reddit destaca como un entorno especialmente relevante para el estudio de percepciones vinculadas al ámbito educativo. Organizada en comunidades temáticas denominadas *subreddits*, esta red agrupa a millones de usuarios que comparten de forma anónima sus experiencias, inquietudes y puntos de vista sobre temas específicos, generando un flujo constante de información espontánea y no condicionada. El estudio de (Li, Xie, Chiu, y Ho, 2023) analizó publicaciones relacionadas con la educación *online* en Reddit y evidenció diferencias significativas entre estudiantes y docentes: mientras los primeros expresaban preocupación por la soledad y la sobrecarga de trabajo, los segundos enfatizaban la falta de retroalimentación y los desafíos en la gestión del tiempo. De manera complementaria, (Rosamma y Rosamma Jr, 2024) identificaron en esta misma plataforma patrones de estrés y ansiedad asociados a experiencias académicas durante la enseñanza remota, destacando el valor de Reddit como observatorio digital para el análisis del bienestar estudiantil en contextos de aprendizaje virtual.

Actualmente, Reddit cuenta con más de 91 millones de usuarios activos diarios y más de 500 millones de usuarios mensuales, posicionándose como una de las plataformas sociales más visitadas a nivel mundial. Tal como se muestra en la Figura 1.3, su número de usuarios ha crecido de manera sostenida, pasando de aproximadamente 330 millones en 2018 a unos 1.500 millones proyectados para 2026, con un incremento anual cercano al 16 %. Este crecimiento no solo refleja la expansión global de la plataforma, sino también su papel central como espacio de conversación y análisis colectivo, donde millones de usuarios generan cada mes decenas de millones de publicaciones y comentarios que capturan percepciones auténticas, emociones y valoraciones sobre una amplia variedad de temas, entre ellos el aprendizaje *online*.

En conjunto, el crecimiento sostenido del aprendizaje *online*, las diferencias en su percepción según el contexto y los múltiples factores que inciden en su eficacia pedagógica evidencian la necesidad de seguir profundizando en el estudio de esta modalidad educativa. Las plataformas digitales como Reddit ofrecen una oportunidad única para explorar estas percepciones desde una perspectiva inductiva, abierta y basada en lenguaje natural, permitiendo acceder a discursos auténticos y no mediados sobre la experiencia educativa en entornos virtuales. En esta línea de investigación se enmarca el presente Trabajo de Fin de Grado, cuyo objetivo es analizar las percepciones, valoraciones y emociones expresadas por los usuarios de Reddit en relación con el aprendizaje *online*, mediante la aplicación de técnicas de NLP (*Natural Language Processing*) que permitan identificar los principales temas de discusión y el sentimiento asociado a ellos. Con ello, se busca contribuir a una comprensión más profunda y matizada de la experiencia estudiantil en entornos digitales, ofreciendo evidencia útil para

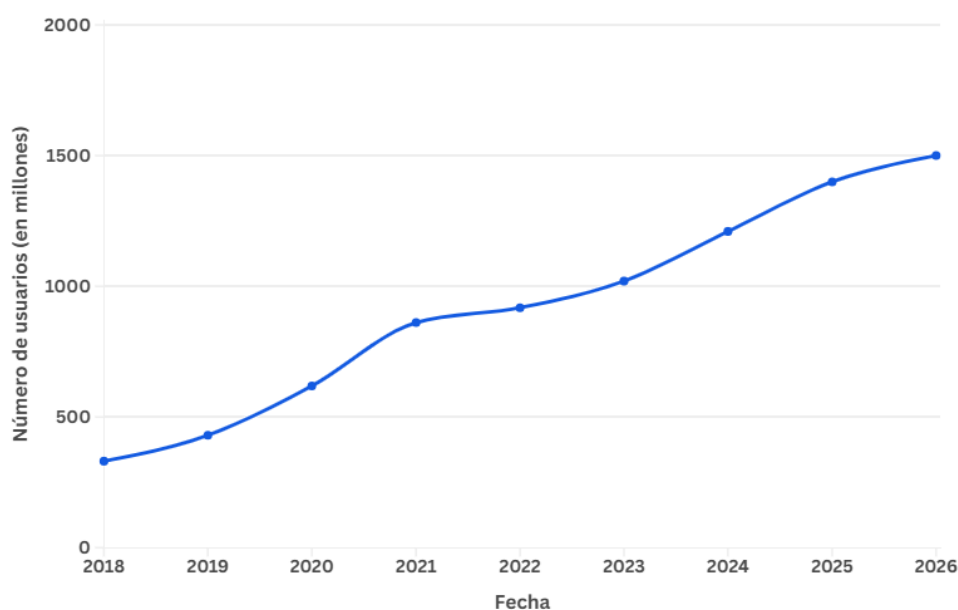


Figura 1.3: Crecimiento global de usuarios activos en Reddit (2018-2026). Fuente: (Business of Apps, 2024). Elaboración propia.

el diseño de estrategias pedagógicas más adaptadas a las necesidades reales del alumnado.

1.2. Objetivos

1.2.1. Objetivo General

El presente Trabajo de Fin de Grado en *Business Analytics* tiene como objetivo principal analizar las percepciones, valoraciones y experiencias expresadas por los usuarios de Reddit en torno al aprendizaje *online*, con el fin de comprender cómo los estudiantes interpretan, evalúan y se relacionan con esta modalidad educativa en entornos digitales. Este estudio busca ofrecer una visión integral de las actitudes y preocupaciones del alumnado, contribuyendo al diseño de estrategias pedagógicas y tecnológicas que promuevan un aprendizaje más inclusivo, participativo y eficaz.

1.2.2. Objetivos Específicos

- Contextualizar la importancia de conocer las percepciones estudiantiles sobre el aprendizaje *online* a partir de comunidades digitales, destacando el papel de plataformas como Reddit en la construcción de narrativas colectivas.
- Identificar las técnicas más utilizadas en la literatura académica para el análisis automático de textos, mediante una revisión sistemática que permita fundamentar metodológicamente el enfoque del estudio.

- Detectar las temáticas predominantes y los sentimientos asociados a la experiencia de aprendizaje *online* en Reddit, aplicando técnicas de modelado de tópicos y análisis de sentimientos a publicaciones y comentarios relevantes.

1.3. Estructura del documento

El presente documento se organiza en cinco capítulos que permiten desarrollar de manera estructurada los objetivos planteados. El Capítulo 2 ofrece una revisión del estado del arte en técnicas de minería de texto y analítica de datos aplicadas al estudio de percepciones públicas, con especial atención a investigaciones que abordan la experiencia estudiantil en entornos digitales y redes sociales. Esta revisión proporciona el marco conceptual y metodológico sobre el cual se apoya el análisis. El Capítulo 3 describe en detalle la metodología empleada, desde la selección y pre-procesamiento del conjunto de datos hasta la aplicación de técnicas de modelado de tópicos y análisis de sentimientos, ofreciendo una visión completa del enfoque adoptado. El Capítulo 4 presenta los resultados obtenidos en cada fase del análisis, destacando los análisis más relevantes en relación con las percepciones estudiantiles sobre el aprendizaje *online*. Finalmente, el Capítulo 5 describe las principales conclusiones del estudio, discute sus implicaciones teóricas y prácticas, y propone posibles líneas de investigación futuras que permitan ampliar y profundizar en la comprensión de la experiencia educativa en entornos virtuales.

Capítulo 2

Percepción del Aprendizaje Online a través de Redes Sociales: Revisión de la Literatura

En un escenario en el que el aprendizaje *online* se ha consolidado como una modalidad estructural en la educación superior y la formación continua, comprender las percepciones del estudiantado resulta relevante para evaluar su calidad y orientar decisiones pedagógicas e institucionales. Estas percepciones aportan información sobre dimensiones que no siempre quedan reflejadas en indicadores de rendimiento, como la satisfacción, la motivación, la sensación de acompañamiento, la carga percibida o las barreras tecnológicas y emocionales asociadas al proceso de aprendizaje. Para analizar estas experiencias pueden emplearse diversas fuentes, entre ellas encuestas, entrevistas, grupos focales, evaluaciones docentes y registros de interacción en plataformas educativas. Sin embargo, muchas de ellas recogen opiniones bajo marcos predefinidos y en contextos de respuesta condicionados. En contraste, las redes sociales y comunidades digitales constituyen una fuente complementaria de creciente relevancia, al ofrecer un acceso masivo a discursos espontáneos y no solicitados, donde los usuarios expresan de forma abierta sus valoraciones, preocupaciones y expectativas en torno al aprendizaje *online*, permitiendo observar tendencias, narrativas y cambios en el tiempo con un nivel de detalle difícil de alcanzar mediante instrumentos tradicionales.

Dado el potencial analítico de las comunidades digitales para capturar percepciones y prácticas educativas, diversos autores han propuesto metodologías orientadas a caracterizar, de manera sistemática, el papel de las redes sociales en los procesos de enseñanza y aprendizaje. En esta línea, se ha señalado que las herramientas de *social media* pueden actuar como catalizadores de cambio en la educación superior al favorecer dinámicas de apertura, interactividad y sociabilidad. Sin embargo, la evidencia empírica basada en muestras amplias a escala nacional continúa siendo limitada. Un ejemplo representativo es el estudio realizado por (Manca y Ranieri, 2016) con profesorado universitario en Italia, que adopta un enfoque

cuantitativo basado en encuesta y se apoya en una muestra amplia de 6.139 respuestas, con un instrumento que recoge frecuencia de uso, motivaciones, prácticas docentes y barreras percibidas para distintos tipos de plataformas (redes generalistas, servicios de *networking* académico-profesional, herramientas colaborativas y repositorios audiovisuales). Metodológicamente, el trabajo combina análisis descriptivos con pruebas inferenciales para examinar, por un lado, las variables sociodemográficas que se asocian con una mayor o menor frecuencia de uso y, por otro, la forma en la que se relacionan motivaciones, formas de utilización y obstáculos con la disciplina científica. Los resultados muestran una adopción todavía limitada de estas herramientas en la docencia universitaria, condicionada por resistencias culturales, dificultades pedagógicas y restricciones institucionales, junto con diferencias significativas entre áreas de conocimiento, lo que apunta a actitudes heterogéneas respecto a los beneficios y desafíos de las redes sociales en educación superior.

En una línea metodológica complementaria, otros trabajos han explotado el contenido generado en redes sociales como fuente primaria para evaluar percepciones públicas sobre el aprendizaje *online*, especialmente durante la pandemia. En este marco se sitúa el estudio de (Mujahid et al., 2021), que analiza la efectividad del *e-learning* a partir de opiniones expresadas en X mediante un corpus de 17.155 publicaciones relacionadas con la educación en línea. Desde un enfoque escalable, los autores combinan técnicas léxicas de análisis de sentimientos, *TextBlob*, *VADER* (*Valence Aware Dictionary and sEntiment Reasoner*) y *SentiWordNet*, para estimar la polaridad y la subjetividad del texto, con modelos supervisados de clasificación contruidos a partir de representaciones de texto basadas en *BoW* (*Bag of Words*) y *TF-IDF* (*Term Frequency–Inverse Document Frequency*). La evaluación comparativa de los modelos se realiza mediante métricas estándar de rendimiento identificando un desempeño particularmente elevado en clasificadores como *RF* (*Random Forest*) y *SVM* (*Support Vector Machine*), que alcanzan una exactitud de 0,95 cuando se emplea *BoW*. Adicionalmente, el estudio incorpora arquitecturas de *deep learning* como *CNN* (*Convolutional Neural Network*), *LSTM* (*Long Short-Term Memory*), *CNN-LSTM* y *Bi-LSTM* (*Bidirectional LSTM*), con el fin de contrastar su desempeño frente a los enfoques tradicionales. Como complemento interpretativo, los autores aplican modelado de tópicos para identificar problemáticas recurrentes asociadas al *e-learning*, destacando la incertidumbre respecto a la reapertura de campus, las dificultades de adaptación de ciertos estudiantes y las limitaciones de conectividad e infraestructura digital.

De forma similar, otros trabajos han extendido este tipo de enfoques hacia contextos lingüísticos y culturales menos representados en la literatura, mostrando que el análisis automatizado de sentimientos en redes sociales puede aportar evidencia relevante para la gestión de crisis y la toma de decisiones públicas. En este contexto, (Alharbi y Alkhateeb, 2021) examinan las orientaciones emocionales de la población árabe durante la pandemia a partir de 30.000 tuits extraídos entre noviembre de 2020 y enero de 2021, incluyendo publicaciones procedentes de distintas ciudades del mundo árabe. Metodológicamente, el estudio combina

técnicas de NLP y ML (*Machine Learning*) para clasificar el sentimiento de las publicaciones en tres categorías de polaridad (positiva, negativa y neutral). Para ello, los autores realizan una fase de anotación manual y emplean representaciones del texto como vectores de características, incluyendo TF-IDF y *word embeddings*. A continuación, comparan el desempeño de dos clasificadores, Naïve Bayes y LSTM, reportando un rendimiento superior del modelo LSTM, con una exactitud del 99 % en su configuración experimental.

Asimismo, algunos trabajos han profundizado en la dimensión temporal de estas percepciones, incorporando metodologías capaces de capturar no solo la polaridad del sentimiento, sino también su evolución y la dinámica de los temas discutidos a lo largo de periodos prolongados. En este campo, (Stracqualursi y Agati, 2022) analizan 25.100 publicaciones en X (anteriormente Twitter) en Italia entre marzo de 2020 y noviembre de 2021 para examinar, de manera conjunta, los sentimientos y los tópicos asociados al aprendizaje a distancia. Para ello, aplican un análisis de sentimientos en dos etapas mediante VADER y el paquete *syuzhet*, lo que permite capturar tanto la polaridad general como emociones específicas. Los resultados muestran una ligera predominancia de opiniones negativas, con variaciones temporales y una posterior reversión de la tendencia, destacando *confianza* como emoción positiva más frecuente y *miedo* y *tristeza* como principales emociones negativas. Además, emplean DLDA (*Dynamic Latent Dirichlet Allocation*) para identificar la evolución temática de la conversación, detectando tres ejes principales que incluyen demandas de medidas de apoyo, preocupaciones sobre la aplicación y el futuro del aprendizaje a distancia, y ansiedad vinculada a decretos y restricciones.

En una línea aplicada orientada a implementar de forma sistemática este tipo de análisis sobre grandes volúmenes de texto, otros autores proponen marcos metodológicos que integran análisis de sentimientos y modelado de tópicos para extraer información útil para los actores del sector educativo a partir de expresiones en redes sociales. En particular, el estudio desarrollado por (Devi, Dhaval, Moharkar, y Khanvilkar, 2022) recoge 6.000 datos de X mediante palabras clave relacionadas con educación *online* y complementa la recolección con respuestas obtenidas a través de un formulario dirigido a estudiantes universitarios, con el fin de disponer de una base empírica diversificada para el análisis. Para ello, se emplea un enfoque de análisis de sentimientos basado en aspectos (*aspect-based sentiment analysis*) para estimar la polaridad de las publicaciones, utilizando el paquete *TextBlob* en Python como herramienta de cálculo de sentimiento. De forma paralela, se aplican técnicas de identificación temática a partir de representaciones BoW y modelos como LDA (*Latent Dirichlet Allocation*) y LSA (*Latent Semantic Analysis*), incorporando medidas de coherencia temática para evaluar la consistencia semántica de los tópicos identificados. Finalmente, los autores reportan que la combinación de análisis de sentimientos y modelado de tópicos permite identificar de manera estructurada los principales focos de discusión y las preocupaciones recurrentes asociadas a la educación *online*, proporcionando una base informativa que puede apoyar a docentes, estudiantes y otros agentes educativos en la priorización de mejoras y en la respuesta

a necesidades detectadas en la conversación digital.

En coherencia con estas aproximaciones basadas en la combinación de análisis de sentimientos y modelado de tópicos, (Li et al., 2023) desarrollaron un estudio centrado en Reddit en el que analizaron más de 19.000 publicaciones con el objetivo de examinar diferencias en las percepciones de estudiantes y educadores respecto a la enseñanza *online*. Para ello, recopilaron mensajes en diversos *subreddits* vinculados al ámbito educativo y aplicaron VADER para estimar la distribución general de sentimientos, junto con LDA para identificar temáticas latentes en el corpus. Mediante esta estrategia de análisis, los autores detectaron tópicos recurrentes relacionados con la efectividad del aprendizaje, la interacción social, la fatiga digital y la motivación, entre otros. En términos de resultados, el estudio muestra diferencias sistemáticas entre ambos colectivos. Mientras el profesorado tendía a enfatizar aspectos asociados a la gestión del proceso de enseñanza y al uso de herramientas tecnológicas, los estudiantes manifestaban con mayor frecuencia preocupaciones vinculadas al nivel de compromiso, la carga de trabajo y la experiencia cotidiana del aprendizaje en entornos virtuales.

La tabla 2.1 resume los principales trabajos revisados que emplean datos de redes sociales para analizar percepciones asociadas al aprendizaje *online*. En concreto, la tabla recoge, para cada referencia, el tamaño y tipo de *dataset*, el método de adquisición de datos, el objetivo del estudio, los algoritmos utilizados y los resultados más relevantes. A partir de la literatura presentada, se observan tres resultados principales. Primero, en términos de adquisición y volumen, predomina la extracción automatizada desde plataformas sociales mediante APIs (*Application Programming Interface*) o herramientas de *scraping*/servicios externos, con tamaños muestrales que oscilan desde varios miles de publicaciones (~6.000 tuits) hasta decenas de miles (17.155, 25.100 o 30.000 tuits), y con trabajos específicos sobre Reddit que superan las 19.000 publicaciones; en menor medida aparecen enfoques basados en recopilación manual o diseños híbridos que combinan redes sociales con encuestas. Segundo, en cuanto a algoritmos, se aprecia una pauta consistente de combinar técnicas de análisis de sentimientos con estrategias de descubrimiento temático. Junto a métodos léxicos (*TextBlob*, *SentiWordNet*), varios estudios incorporan modelos supervisados (SVM, *random forest*) y redes neuronales (LSTM, Bi-LSTM, CNN), y complementan la interpretación con modelado de tópicos (LDA, LSA) o variantes dinámicas (DLDA). Tercero, destaca el uso recurrente de VADER en análisis de sentimientos aplicados a redes sociales, debido a que integra un léxico amplio con reglas capaces de manejar rasgos del lenguaje informal, como emoticonos, mayúsculas o signos de exclamación repetidos, permitiendo analizar grandes cantidades de texto sin entrenamiento previo y resultando especialmente adecuado para mensajes cortos y emocionalmente cargados, como los de Reddit (Youvan, 2024). De manera análoga, LDA aparece como el modelo de tópicos más extendido en este tipo de estudios. De hecho, una revisión sistemática reporta que LDA se empleó en 154 trabajos (79,79 %), muy por encima de alternativas como STM (*Structural Topic Model*) o BTM (*Biterm Topic Model*) (Laureate, Buntine, y Linger, 2023). Esta prevalencia se asocia a su capacidad de identificar temas de

forma no supervisada y a la interpretabilidad de su estructura probabilística, lo que facilita el análisis comparativo de grandes volúmenes de texto incluso en contextos con alta diversidad temática.

En este contexto, y tomando como referencia las tendencias metodológicas identificadas en la literatura, el presente Trabajo de Fin de Grado se plantea como una aplicación integrada de análisis de sentimientos y modelado de tópicos para profundizar en la comprensión de las percepciones sobre el aprendizaje *online* en comunidades digitales. En particular, se propone utilizar Reddit como fuente principal de datos, dada su estructura conversacional y el carácter espontáneo de las publicaciones, con el objetivo de capturar valoraciones y experiencias expresadas de manera no solicitada por los usuarios. Siguiendo la evidencia previa, el estudio adopta VADER como técnica base para estimar la polaridad del sentimiento en mensajes breves e informales y emplea LDA como herramienta para identificar los ejes temáticos predominantes en la discusión. De este modo, el trabajo busca no solo caracterizar el tono general asociado al aprendizaje *online*, sino también vincular dicho tono con las temáticas emergentes, proporcionando una visión más completa de las preocupaciones, expectativas y actitudes del estudiantado en entornos digitales.

Tabla 2.1: Resumen de estudios previos relacionados.

Referencia	Tamaño del Dataset	Método de Adquisición	Objetivo	Algoritmo	Resultados
(Manca y Rani, 2016)	6.139 respuestas	Recopilación manual a través de encuestas al profesorado.	Explorar percepciones sobre el aprendizaje colaborativo digital y la interacción estudiantil.	Frecuencia de palabras; Codificación cualitativa manual.	El estudio concluye que el uso de redes sociales en la educación superior es aún marginal. La adopción de estas herramientas se ve limitada por una combinación de factores que incluyen variables sociodemográficas, cuestiones institucionales, perspectivas pedagógicas, barreras prácticas y valores personales.
(Mujahid et al., 2021)	17.155 tuits	Extracción mediante API de X; filtrado por palabras clave relacionadas con educación <i>online</i> .	Evaluar la percepción pública sobre la enseñanza remota durante la pandemia.	Análisis de sentimientos: TextBlob, VADER, SentiWordNet; DT (<i>Decision Tree</i>), SVM, RF, LSTM, Bi-LSTM. Modelado de tópicos: LSA.	Los usuarios identificaron problemáticas centrales como la incertidumbre sobre el retorno a clases presenciales, la dificultad de adaptación del alumnado con necesidades especiales, y la falta de infraestructuras digitales adecuadas.
(Alharbi y Alkhateeb, 2021)	30.000 tuits aprox.	Extracción mediante API; filtrado por idioma árabe y palabras clave de COVID-19.	Analizar la percepción y emociones públicas sobre las medidas sanitarias durante la pandemia.	Análisis de sentimientos: Naïve Bayes, LSTM.	LSTM alcanzó un 98.9 % de precisión frente al 76 % de NB (<i>Naïve Bayes</i>). Aunque no centrado en educación, ilustra cómo el análisis de percepciones en redes sociales requiere modelos más complejos y sensibles al matiz emocional, metodología posteriormente aplicada al aprendizaje <i>online</i> .
(Stracqualursi y Agati, 2022)	25.100 tuits	Tuits recopilados mediante TrackMyHashtag, filtrando por hashtags relacionados con “aprendizaje a distancia”, idioma italiano y fechas entre marzo de 2020 y noviembre de 2021	Evaluar la evolución temporal del sentimiento hacia la educación <i>online</i> en Italia.	Análisis de sentimientos: VADER, NRC Emotion Lexicon. Modelado de tópicos: DLDA.	Predominio inicial de sentimiento negativo que mejora con la llegada de modelos híbridos. DLDA reveló tres áreas clave: soporte tecnológico, impacto académico-emocional y ansiedad por medidas gubernamentales.
(Devi et al., 2022)	6.000 tuits + encuestas estudiantiles	Tuits obtenidos con <i>snsrape</i> y recolección paralela de respuestas de estudiantes.	Estudiar actitudes hacia la educación <i>online</i> combinando redes sociales y opiniones directas de alumnos.	Análisis de sentimientos: TextBlob. Modelado de tópicos: LDA, LSA, Bag of Words.	Confirman que integrar análisis de sentimientos y modelado de tópicos permite detectar tensiones sobre accesibilidad, motivación y eficacia del aprendizaje remoto, consolidando la tendencia metodológica observada en la literatura.
(Li et al., 2023)	19.000 publicaciones en Reddit	Datos recopilados de Reddit mediante Pushshift.io y la Reddit API oficial, combinando archivos históricos completos con solicitudes API en tiempo real.	Comparar percepciones de estudiantes y docentes sobre la enseñanza <i>online</i> .	Análisis de sentimientos: VADER. Modelado de tópicos: LDA.	Estudiantes preocupados por carga y motivación, docentes centrados en herramientas y gestión.

Capítulo 3

Metodología de Análisis de Datos

Este capítulo describe en detalle la metodología empleada en este Trabajo de Fin de Grado, centrado en el análisis de las percepciones de los usuarios de *Reddit* sobre el aprendizaje *online*. La metodología se estructura en varias etapas, cada una compuesta por actividades específicas que organizan el proceso de investigación. Estas etapas se presentan en el diagrama de flujo mostrado en la Figura 3.1. En resumen, el procedimiento incluye la adquisición y preparación de datos, el preprocesamiento textual, la exploración descriptiva mediante n-gramas, el modelado de tópicos con LDA y el análisis de sentimientos con VADER.

En la primera etapa, la recolección y preparación de los datos, se seleccionan una serie de palabras clave relacionadas con el aprendizaje *online*, para garantizar que la muestra sea representativa y se extraen publicaciones y comentarios utilizando la API de *Reddit*. Esta fase se enfoca en garantizar que los datos sean representativos y adecuados para el análisis. Una vez obtenidos, se procede con el pre-procesamiento de los datos, que incluye la limpieza, tokenización y normalización de los textos para asegurar que el análisis posterior sea coherente. El análisis descriptivo de n-gramas se lleva a cabo a continuación, con el fin de identificar las combinaciones de palabras más frecuentes en el corpus de texto. Los n-gramas permiten captar las relaciones contextuales entre términos y facilitarán la aplicación de técnicas de aprendizaje no supervisado. En este caso, el análisis de n-gramas se realiza mediante el cálculo de la métrica *Term Frequency-Inverse Document Frequency* (TF-IDF), que pondera la importancia de las palabras dentro de un conjunto de documentos. Esta métrica ayuda a identificar términos que son más relevantes para los temas tratados en las publicaciones, descartando las palabras comunes y poco informativas.

En la etapa de modelado de tópicos, se utiliza *Latent Dirichlet Allocation* (LDA), un enfoque probabilístico que permite identificar los temas recurrentes en las publicaciones sobre aprendizaje *online*, clasificando los comentarios en diferentes grupos temáticos. Finalmente, para el análisis de sentimientos, se emplea VADER, una herramienta que clasifica las emociones en los comentarios como positivas, negativas o neutrales, y asigna una puntuación de polaridad a cada publicación.

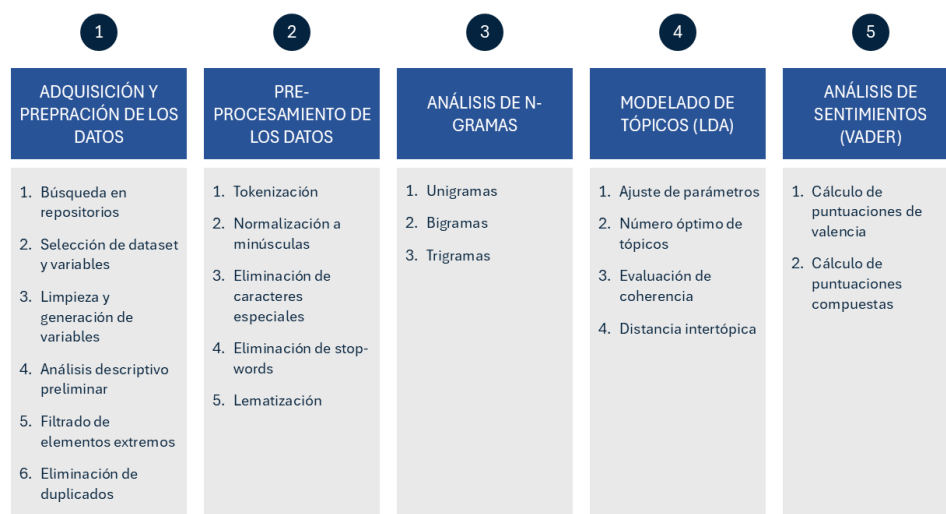


Figura 3.1: Fases de la Metodología. Elaboración propia.

3.1. Adquisición y Preparación de los Datos

Antes de llevar a cabo el modelado de tópicos y el análisis de sentimientos, resulta imprescindible disponer de un conjunto de datos de alta calidad que sirva como base para el estudio. En este trabajo, la recopilación de datos se ha centrado en publicaciones extraídas directamente de la plataforma Reddit, específicamente relacionadas con el aprendizaje *online*. Para la obtención de los datos, se ha utilizado la API oficial de *Reddit*, accediendo a ella a través de la librería *PRAW (Python Reddit API Wrapper)*, una interfaz de *Python* ampliamente empleada en el ámbito académico para la extracción automatizada de publicaciones de esta plataforma. La extracción de publicaciones se ha llevado a cabo mediante una estrategia de búsqueda sistemática basada en palabras clave, diseñada para capturar la mayor diversidad posible de experiencias y percepciones sobre el aprendizaje *online*. Este enfoque permitió recuperar publicaciones y comentarios relevantes de distintas comunidades de *Reddit*, asegurando que el conjunto de datos reflejara conversaciones directamente vinculadas al objeto de estudio.

Para maximizar la representatividad del corpus, cada palabra clave fue utilizada como consulta de búsqueda en el conjunto de todos los subreddits de *Reddit*, aplicando tres criterios de ordenación diferentes. El primero de ellos, *new*, ordena los resultados por fecha de publicación, priorizando las publicaciones más recientes y garantizando así la captación de las percepciones más actuales sobre el aprendizaje *online*. El segundo criterio, *relevance*, clasifica los resultados según el grado de coincidencia entre la consulta y el contenido de la publicación, lo que permite recuperar aquellas publicaciones cuyo texto se encuentra más estrechamente vinculado con las palabras clave definidas. Por último, el criterio *top* ordena los resultados en función de la puntuación neta recibida por cada publicación, calculada como la

diferencia entre los votos positivos y negativos otorgados por la comunidad, lo que permite captar las publicaciones que han generado un mayor impacto o resonancia entre los usuarios. La combinación de estos tres criterios reduce el sesgo que podría derivarse de un único método de ordenación, asegurando que el corpus incluya tanto publicaciones recientes como aquellas más relevantes temáticamente y las que han generado un mayor nivel de interacción en la plataforma. Adicionalmente, se aplicó un filtro temporal para conservar exclusivamente las publicaciones posteriores al 1 de enero de 2020, dado que la pandemia COVID-19 supuso un punto de inflexión en la adopción y percepción del aprendizaje *online* a nivel global, tal como se ha documentado en el Capítulo 1.

De cada publicación se han extraído las siguientes variables: el título, la fecha de creación, el texto o cuerpo del mensaje, el subreddit de origen, el autor y la palabra clave que originó su recuperación. Dado que una misma publicación puede ser recuperada por múltiples palabras clave, se ha llevado a cabo un proceso de eliminación de duplicados basado en el título, conservando únicamente la primera aparición de cada registro. Asimismo, se han descartado las publicaciones con campos de texto vacíos o marcados como eliminados. En total, se extrajeron 15.579 publicaciones.

3.2. Pre-procesamiento de los datos

En esta segunda etapa se realiza el pre-procesamiento del conjunto de datos con el objetivo de garantizar su calidad y adecuación para los análisis posteriores. En primer lugar, se lleva a cabo una limpieza inicial del *dataset*, eliminando observaciones vacías y depurando el contenido textual para reducir el ruido propio de las conversaciones en plataformas digitales. En esta fase se eliminan elementos que no aportan valor analítico, como las URL (*Uniform Resource Locator*), referencias a usuarios y comunidades (por ejemplo, patrones del tipo *u/usuario* o *r/subreddit*), caracteres especiales y signos de puntuación que pueden introducir inconsistencias en el modelado temático. Adicionalmente, se normaliza el texto para homogeneizarlo, convirtiendo todo el contenido a minúsculas y aplicando normalización de caracteres para evitar que variaciones ortográficas o símbolos afecten al análisis. Dado que el corpus analizado se encuentra mayoritariamente en inglés, se decide limitar el análisis textual a este idioma. Esta decisión permite trabajar con un conjunto lingüísticamente homogéneo y aprovechar herramientas de procesamiento de lenguaje natural ampliamente validadas y optimizadas para inglés, como los modelos de lematización de *spaCy* y los diccionarios empleados en *sentiment analysis*. De esta manera, se mejora la comparabilidad de los resultados a lo largo de todo el estudio.

En segundo lugar, se procede a la eliminación de *stopwords* (palabras vacías) en inglés, es decir, términos muy frecuentes que no aportan nada al contenido (por ejemplo, “*a*”, “*of*”, “*and*” o “*you*”). Dado el carácter conversacional de Reddit, esta lista se complementa con

stopwords adicionales de uso recurrente en lenguaje informal (por ejemplo, “like”, “really”, “people”) y términos meta propios de la plataforma (por ejemplo, “post”, “comments”, “thread”), con el objetivo de evitar que dominen los resultados y reduzcan la interpretabilidad de los tópicos.

Posteriormente, se aplica la tokenización, dividiendo cada texto en unidades mínimas (*tokens*) sobre las que se ejecutan transformaciones posteriores. A continuación, se realiza la lematización mediante *spaCy*, reduciendo cada palabra a su forma base o lema con el fin de unificar variantes morfológicas y mejorar la consistencia del corpus. Este paso resulta especialmente útil para técnicas posteriores como el análisis de *n-gramas* y el modelado de tópicos, ya que facilita la identificación de términos clave y categorías de discusión. Además, se incorporan normalizaciones específicas del dominio del aprendizaje *online* (por ejemplo, unificando expresiones como “*e-learning*”, “*online learning*” o “*learning management system*” bajo una nomenclatura común), con el objetivo de capturar con mayor precisión conceptos centrales del tema estudiado. Finalmente, tras la generación del texto lematizado, se revisa el conjunto resultante y se eliminan aquellas observaciones en las que el contenido queda vacío como consecuencia del filtrado. Con ello se obtiene un corpus limpio y consistente, preparado para las etapas siguientes de análisis descriptivo, modelado de tópicos y análisis de sentimientos.

Es importante señalar que parte de este pre-procesamiento se orienta específicamente al modelado temático y análisis de términos, donde la normalización y eliminación de ruido son relevantes. Sin embargo, para el análisis de sentimientos se conserva una versión del texto con menor grado de transformación, ya que elementos como la puntuación, intensificadores, expresiones coloquiales o ciertos signos pueden aportar matices relevantes para evaluar con precisión el tono emocional de las publicaciones.

3.3. Exploración de N-Gramas

Una vez depurado el corpus textual, el siguiente paso consiste en explorar las combinaciones de palabras que aparecen con mayor frecuencia en las publicaciones, con el fin de obtener una primera visión del conjunto de datos antes de aplicar técnicas de modelado más complejas. Para ello, se recurre al análisis de N-gramas, un concepto ampliamente utilizado en el procesamiento de lenguaje natural que hace referencia a secuencias de n palabras consecutivas dentro de un texto. Cuando $n = 1$, se habla de unigramas (palabras individuales), cuando $n = 2$, de bigramas (pares de palabras), y cuando $n = 3$, de trigramas (tríos de palabras). A modo ilustrativo, dada la frase “*I really enjoy studying from home*”, algunos de sus bigramas serían “*really enjoy*” y “*studying from*”, mientras que “*enjoy studying from*” constituiría un trigramas. El interés de analizar no solo palabras aisladas sino también sus combinaciones reside en que muchas expresiones relevantes para el estudio adquieren su significado completo

únicamente cuando se consideran en conjunto, como ocurre con “*mental health*” o “*time management*”.

No obstante, la simple frecuencia de aparición de un N-grama no resulta suficiente para determinar su relevancia dentro del corpus. Las palabras muy comunes pueden acumular frecuencias elevadas sin aportar información diferenciadora entre documentos. Para superar esta limitación, se utiliza la métrica TF-IDF (*Term Frequency-Inverse Document Frequency*), que pondera la importancia de cada término combinando dos factores complementarios.

El primero de ellos, denominado Frecuencia de Término, cuantifica el número de veces que un término aparece dentro de un documento individual. Sea t un término y d un documento, esta frecuencia se define como se muestra en la Ecuación (3.1):

$$TF(t, d) = f(t, d) \quad (3.1)$$

donde $f(t, d)$ denota el número de ocurrencias del término t en el documento d .

El segundo factor, la Frecuencia Inversa de Documento, penaliza aquellos términos que se distribuyen de manera uniforme a lo largo de todo el corpus, otorgando mayor peso a las palabras que se concentran en un subconjunto reducido de documentos. Dado un corpus D compuesto por N documentos, su cálculo se expresa según la Ecuación (3.2):

$$IDF(t) = \ln \left(\frac{N}{df(t)} \right) \quad (3.2)$$

donde $df(t)$ representa el número de documentos del corpus que contienen el término t . La puntuación final de cada término se obtiene como el producto de ambos componentes, tal como se recoge en la Ecuación (3.3):

$$TF-IDF(t, d) = TF(t, d) \times IDF(t) \quad (3.3)$$

De esta forma, un valor elevado de TF-IDF señala que el término en cuestión resulta frecuente en un documento específico pero infrecuente en el conjunto global, lo que lo convierte en un indicador distintivo de su contenido. Por el contrario, términos omnipresentes en el corpus reciben puntuaciones bajas, al carecer de capacidad para diferenciar unas publicaciones de otras. En el presente estudio, esta métrica se aplica de forma independiente a unigramas, bigramas y trigramas extraídos del corpus de publicaciones de *Reddit*. El análisis de unigramas permite identificar los conceptos individuales que dominan la conversación. Los bigramas facilitan la detección de asociaciones recurrentes entre dos conceptos, como “*remote work*” o “*self discipline*”. Por su parte, los trigramas aportan un grado adicional de especificidad, al capturar expresiones más elaboradas como “*lack social interaction*” o “*work life balance*”.

3.4. Modelado de Tópicos

Tras la exploración de N-gramas, se aplica *Latent Dirichlet Allocation* (LDA) para identificar de forma automática los temas de discusión latentes en las publicaciones de *Reddit*. LDA es un modelo probabilístico no supervisado que parte de la idea de que cada publicación del corpus aborda una combinación de temas, y que cada tema, a su vez, se compone de un conjunto de palabras con distintas probabilidades de aparición. De esta manera, el modelo es capaz de agrupar publicaciones que comparten patrones léxicos similares, revelando los ejes temáticos en torno a los cuales se articula la conversación sobre el aprendizaje *online*.

El proceso generativo de LDA se representa gráficamente en la Figura 3.2. En ella se observa cómo, a partir de un hiperparámetro α , se genera para cada documento una distribución de tópicos θ que refleja el peso relativo de cada tema en dicha publicación. Para cada posición de palabra dentro del documento, se selecciona un tópico z de acuerdo con θ , y a partir de ese tópico se genera la palabra observada w , cuya distribución viene controlada por el hiperparámetro β . Los recuadros del diagrama indican que este proceso se repite N veces por documento y M veces por corpus.

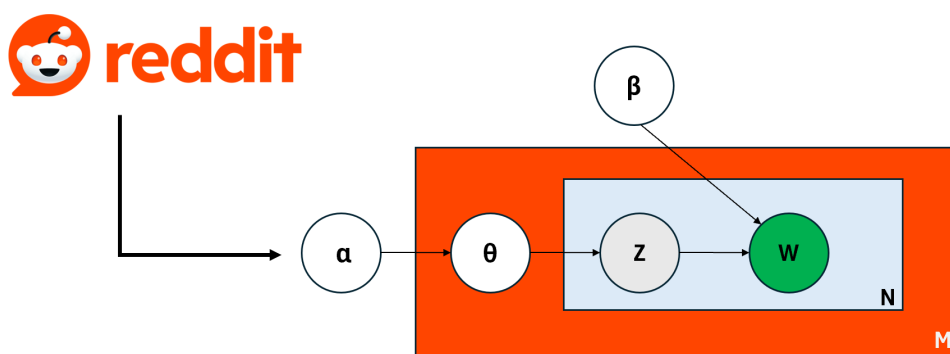


Figura 3.2: Representación del modelo LDA. Fuente: Journal of Machine Learning Research. (Blei et al., 2003). Elaboración propia.

LDA trabaja bajo el enfoque *Bag of Words*, lo que implica que no considera el orden de las palabras sino únicamente su frecuencia. Antes de ejecutar el modelo, se filtran del diccionario los términos que aparecen en menos de 2 documentos o en más del 95 % del corpus, ya que no aportan capacidad discriminativa entre tópicos. Para determinar el número adecuado de tópicos (k), se entrena el modelo con distintos valores de k y se evalúa la calidad de cada configuración mediante el índice de coherencia, que mide en qué grado las palabras principales de cada tópico están semánticamente relacionadas entre sí. El valor de k se selecciona aplicando el método del codo sobre la curva de coherencia, eligiendo el punto a partir del cual la mejora deja de ser significativa. Una vez entrenado el modelo definitivo, se valida la separación entre tópicos mediante *pyLDavis*, una herramienta que muestra cada tópico como un círculo en un plano bidimensional, permitiendo comprobar visualmente que no existen

solapamientos relevantes entre las categorías identificadas.

3.5. Análisis de Sentimientos

La última etapa de la metodología consiste en evaluar el tono emocional de las publicaciones mediante un análisis de sentimientos, con el fin de complementar la información temática obtenida del modelado de tópicos con una valoración de las actitudes y emociones que los usuarios de *Reddit* expresan en relación con el aprendizaje *online*. Para ello, se emplea VADER, un modelo basado en léxico específicamente diseñado para analizar el sentimiento en textos procedentes de redes sociales (Youvan, 2024). VADER resulta particularmente adecuado para este estudio por varias razones. A diferencia de otros modelos léxicos, VADER incorpora reglas que tienen en cuenta elementos propios del lenguaje informal en línea, como el uso de mayúsculas para expresar énfasis, los signos de exclamación repetidos, las negaciones, los emoticonos y las abreviaturas coloquiales. Estas características le permiten captar matices emocionales que se perderían con herramientas diseñadas para textos formales. Además, al tratarse de un modelo preentrenado, no requiere una fase de aprendizaje supervisado, lo que lo hace directamente aplicable al corpus sin necesidad de etiquetar datos manualmente.

Es importante destacar que, por las razones mencionadas, el texto que se suministra a VADER recibe un pre-procesamiento diferente al descrito en la sección 3.2. En este caso, se conservan las mayúsculas, los signos de puntuación, los emoticonos y las *stopwords*, ya que todos estos elementos aportan información relevante sobre la intensidad y la dirección del sentimiento. Únicamente se eliminan las URLs y las menciones a usuarios y subreddits, que no contribuyen al contenido emocional del mensaje. VADER asigna a cada publicación una puntuación compuesta (*compound score*) que sintetiza la polaridad global del texto en un valor normalizado entre -1 (sentimiento más negativo) y $+1$ (sentimiento más positivo). A partir de esta puntuación, cada publicación se clasifica en una de tres categorías: positiva si la puntuación es igual o superior a $0,05$, negativa si es igual o inferior a $-0,05$, y neutra si se sitúa entre ambos umbrales.

Una vez obtenidas las puntuaciones de sentimiento, se realiza un análisis cruzado con los tópicos identificados mediante LDA en la etapa anterior. Este cruce permite evaluar no solo la distribución general del sentimiento en el corpus, sino también el tono emocional asociado a cada categoría temática. De este modo, se obtiene una visión integrada que combina la dimensión temática con la dimensión afectiva, permitiendo identificar, por ejemplo, si los tópicos relacionados con la interacción social tienden a asociarse con emociones negativas o si las publicaciones sobre plataformas educativas presentan un tono más positivo. Este enfoque consistente con las prácticas metodológicas identificadas en la revisión de la literatura (Devi et al., 2022; Li et al., 2023; Stracqualursi y Agati, 2022), proporciona evidencia útil para comprender la experiencia del estudiantado en entornos de aprendizaje virtual.

Capítulo 4

Resultados

Este capítulo presenta los resultados obtenidos tras la implementación de la metodología descrita en el Capítulo 3. En primer lugar, se realiza una caracterización descriptiva del corpus (evolución temporal, distribución lingüística y principales comunidades de publicación), junto con el análisis exploratorio de elementos expresivos como tokens tipo “*hashtag*” y *emojis*. Este enfoque preliminar permite comprender la composición del conjunto de datos y establecer criterios de depuración y segmentación antes de aplicar técnicas de *NLP* de mayor profundidad.

4.1. Adquisición y Preparación de los Datos

En una primera fase de recopilación se obtuvieron 26.000 entradas textuales de Reddit. Posteriormente, se llevó a cabo un proceso de depuración orientado a garantizar la consistencia del corpus. En esta etapa, se eliminaron los registros duplicados, conservando únicamente la primera aparición de cada entrada, y se descartaron aquellas observaciones con campos de texto vacíos o marcados como eliminados por los usuarios. Como resultado de esta limpieza inicial, el conjunto de datos se redujo a 15.579 entradas únicas, que constituyen la base sobre la cual se desarrollaron las etapas posteriores del análisis. Antes de aplicar las técnicas de *NLP*, se realizaron una serie de análisis descriptivos preliminares con el fin de caracterizar el corpus y comprender mejor su estructura. En concreto, estos análisis incluyeron: (i) la evolución temporal del volumen de entradas, (ii) la distribución lingüística del contenido, (iii) la identificación de las comunidades (*subreddits*) con mayor actividad y (iv) una exploración de elementos expresivos, como *hashtags* y *emojis*.

La Figura 4.1 muestra el número de entradas agregadas por año. Se observa una presencia moderada en los primeros años del intervalo (2020-2022), coincidiendo con el período más intenso de la pandemia, seguida de un incremento notable en los años más recientes (2023-2025). Este patrón sugiere una intensificación sostenida de la conversación en torno al aprendizaje *online*, posiblemente vinculada a la consolidación de esta modalidad educativa

más allá del contexto de emergencia sanitaria.

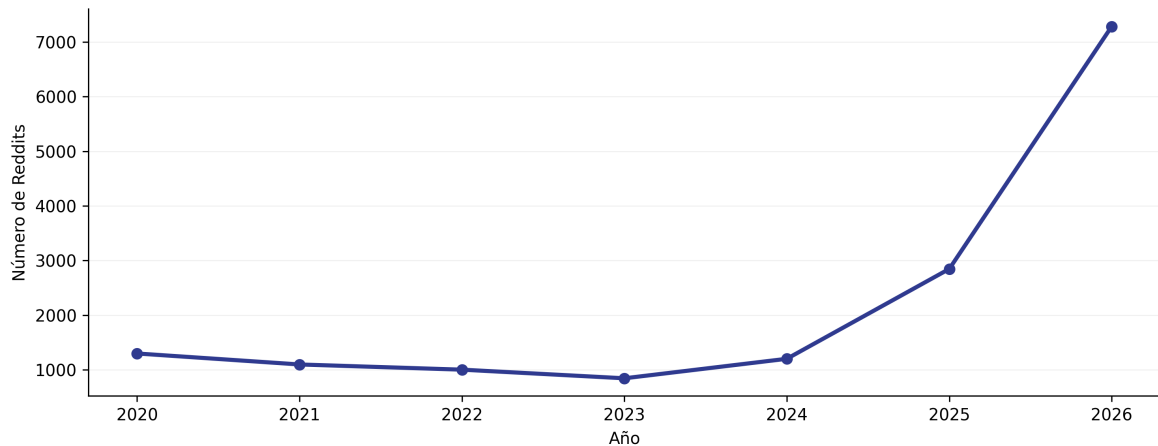


Figura 4.1: Número de publicaciones de Reddit por año. Elaboración propia.

Con el fin de evaluar la diversidad lingüística del corpus, se aplicó un procedimiento automático de detección de idioma mediante la librería *langdetect* de *Python*. La Figura 4.2 muestra la distribución de publicaciones por idioma, donde el inglés concentra el 98 % de las observaciones, mientras que el resto de lenguas (español, tagalo, portugués y francés, entre otras) representan conjuntamente el 2 % restante. Este patrón es consistente con la composición demográfica de *Reddit*, donde el inglés actúa como lengua predominante. Dado este resultado, se decidió limitar el análisis textual a las publicaciones en inglés, lo que permite trabajar con un conjunto lingüísticamente homogéneo.

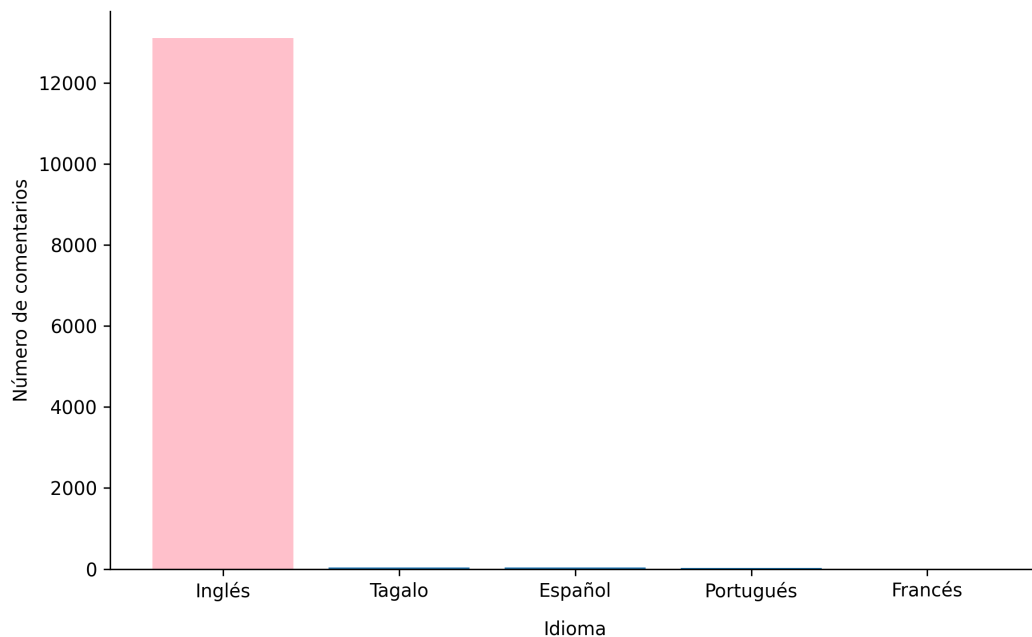


Figura 4.2: Número de publicaciones por idioma. Elaboración propia.

La Figura 4.3 presenta los *subreddits* con mayor volumen de publicaciones dentro del conjunto de datos. Los resultados muestran una concentración en un número reducido de comunidades, seguida de una distribución en “cola larga” en la que el resto de *subreddits* aportan un volumen comparativamente menor. Este patrón sugiere que la conversación relacionada con el aprendizaje *online* no se produce únicamente en comunidades estrictamente educativas, sino que también aparece en espacios generalistas o de carácter narrativo.

En este sentido, destaca la presencia de *BestofRedditorUpdates* entre los *subreddits* con mayor actividad. Este *subreddit* (conocido también como *BORU*) funciona principalmente como un agregador de contenido: recopila y republica historias publicadas originalmente en otros *subreddits* junto con sus actualizaciones posteriores, de modo que el lector pueda seguir la evolución completa de un caso en un único hilo. Por tanto, su aparición en el *top* no implica necesariamente que el aprendizaje *online* sea el tema central del *subreddit*, sino que refleja que dicho concepto aparece con frecuencia dentro de relatos personales (por ejemplo, experiencias académicas o laborales) que posteriormente son compilados en esta comunidad.

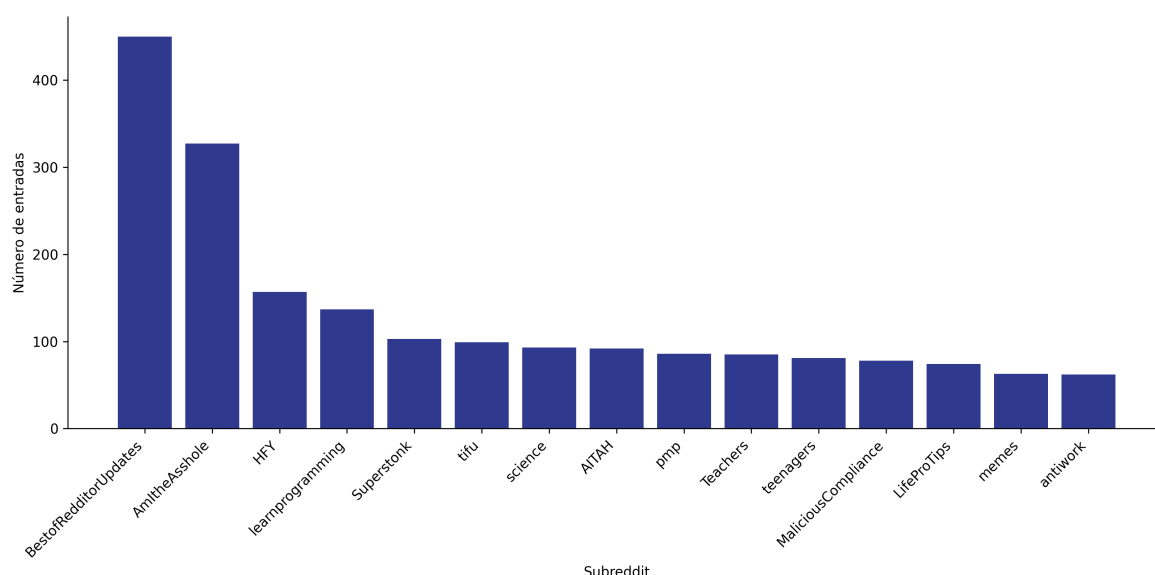


Figura 4.3: Top subreddits. Elaboración propia.

El análisis de los *hashtags* más frecuentes (Figura 4.4) muestra una concentración notable en etiquetas que, más que funcionar como marcadores temáticos propios de plataformas como X, parecen responder a patrones de formato y estructura dentro de Reddit. En particular, la presencia destacada de *#cast*, *#crew*, *#synopsis*, *#plot*, *#status* y *#links* sugiere que una parte relevante del corpus proviene de publicaciones altamente estructuradas, en las que el símbolo # se utiliza para organizar secciones o apartados del contenido. Este comportamiento es consistente con entradas de tipo recopilatorio o narrativas que siguen plantillas recurrentes y, por tanto, indica que la frecuencia de estas etiquetas puede estar capturando en mayor medida la *forma* del mensaje que su temática específica. En este sentido, estos resultados

deben interpretarse como un indicador exploratorio de estilos de publicación y de regularidades textuales, más que como una clasificación temática directa del aprendizaje *online*. De forma complementaria, el análisis de los *emojis* más frecuentes (Figura 4.5) aporta información sobre el componente expresivo del corpus. Los resultados muestran que los usuarios recurren a un conjunto reducido de símbolos para reforzar evaluaciones o enfatizar mensajes de manera rápida. La predominancia de ciertos *emojis* asociados a aprobación, intensidad o énfasis sugiere que parte de la conversación incorpora juicios marcados (positivos o de alta activación), lo cual resulta útil como contexto interpretativo para las fases posteriores de análisis de sentimiento y polaridad.

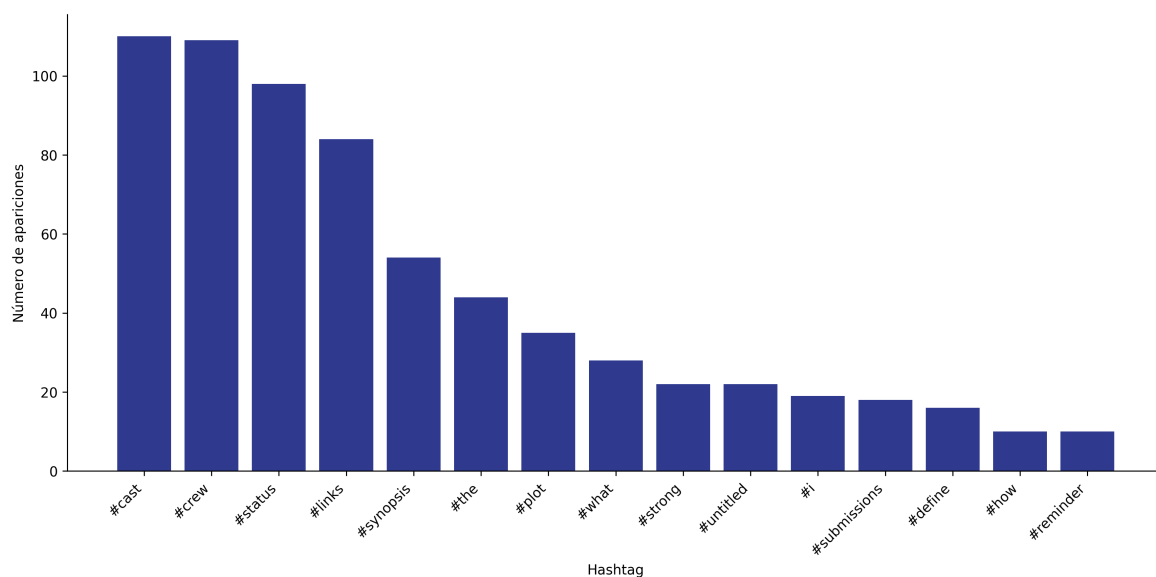


Figura 4.4: Top hashtags. Elaboración propia.

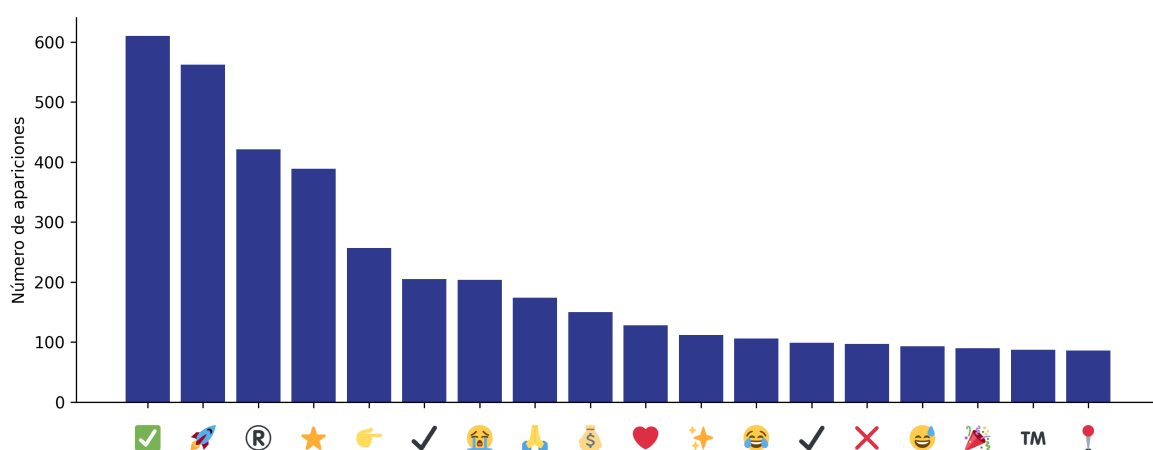


Figura 4.5: Top emojis. Elaboración propia.

4.2. Pre-procesamiento de los datos

Durante la fase de pre-procesamiento de datos, se lleva a cabo la limpieza necesaria antes de iniciar las fases de modelado de tópicos y de análisis de sentimientos. Para ello, el *dataset* principal se ha trasladado a *Python*, donde se realizarán las tareas de limpieza y los análisis posteriores. En primer lugar, se excluyen las publicaciones que no están escritas en inglés, lo que facilita el uso de diccionarios y librerías específicas de este idioma, haciendo el análisis más sencillo y eficiente. Luego, se realiza la eliminación de URLs, menciones a usuarios “*u/usuario*” y referencias a comunidades “*r/subreddit*”, así como caracteres especiales y números. A continuación, se utiliza la librería *spaCy* para eliminar las *stopwords* más comunes, las cuales fueron ampliadas con términos propios del lenguaje informal y expresiones específicas de *Reddit* como “*post*”, “*comment*”, “*thread*”, “*edit*” y “*update*”. Posteriormente, se aplica el proceso de lematización mediante el modelo *en_core_web_sm* para reducir las palabras a su forma base. Finalmente, se eliminan las publicaciones duplicadas y aquellas que quedaron vacías tras el filtrado. Durante el estudio, se encontraron algunas excepciones en cuanto a *stopwords* y lematización, las cuales se añadieron al proceso para abarcar casos especiales que no estaban contemplados en las librerías utilizadas. En este punto, es importante tener en cuenta que el pre-procesamiento de datos difiere en la etapa de modelado de tópicos y en la de análisis de sentimientos, de acuerdo a las explicaciones incluidas en el Capítulo 3. Por lo tanto, los pasos específicos para cada uno de ellos serán abordados y explicados en sus respectivos sub-apartados. El proceso de limpieza se realizó en tres fases. En primer lugar, la eliminación de URLs, menciones, caracteres especiales y palabras de una o dos letras redujo el corpus de 15.579 a 13.114 publicaciones, al descartar aquellas cuyo contenido textual quedó vacío tras las transformaciones. En segundo lugar, se aplicó el filtro de idioma inglés, eliminando 218 registros en otros idiomas y obteniendo 12.896 publicaciones. Finalmente, tras el proceso de lematización, el corpus final quedó constituido por 12.893 publicaciones aptas para el análisis.

Con el fin de obtener una visualización inicial y descriptiva del corpus resultante, se ha elaborado un *wordcloud* (Figura 4.6) que representa las 100 palabras más frecuentes del conjunto de textos. En esta visualización, el tamaño de cada término refleja su frecuencia relativa dentro del corpus. A primera vista, se observa una clara predominancia de palabras estrechamente vinculadas al ámbito educativo y al proceso de aprendizaje, como *exam*, *course*, *test*, *learn*, *class*, *school*, *student*, *homework*, *college* y *university*, lo que confirma la coherencia temática del corpus con el objeto de estudio. Junto a estos términos, también destacan palabras relacionadas con la dimensión académica práctica y evaluativa, como *work*, *write*, *study*, *research*, *project*, *practice* y *skill*, lo que sugiere que las conversaciones no solo giran en torno al aprendizaje en sentido general, sino también a sus aplicaciones concretas, tareas y procesos de evaluación. Asimismo, aparecen términos como *online*, *program*, *system*, *tool* y *support*, que refuerzan la orientación del corpus hacia entornos digitales de aprendizaje y

cir, palabras individuales que aparecen en el *dataset*. Al analizar estos unigramas, se destaca la presencia de términos directamente asociados al ámbito educativo y a la experiencia del aprendizaje, como *learn*, *course*, *school*, *class*, *student* y *study*. También destacan palabras vinculadas al contexto laboral (*work*, *job*, *experience*), lo que sugiere una conexión frecuente entre la formación *online* y el desarrollo profesional. Resulta especialmente relevante la presencia de términos asociados al ámbito académico formal como *exam*, *test*, *university*, *college* y *degree*, lo que indica que una parte significativa de las discusiones se centra en la evaluación y en las credenciales educativas. Asimismo, la aparición de términos como *build*, *project* y *practice* apunta a un interés por el desarrollo de habilidades prácticas y la aplicación del conocimiento adquirido. La presencia del término *hard* sugiere que los usuarios también discuten sobre las dificultades asociadas al aprendizaje *online*. Por último, la aparición del término *online* en una posición destacada confirma la coherencia temática del corpus con los objetivos del estudio.

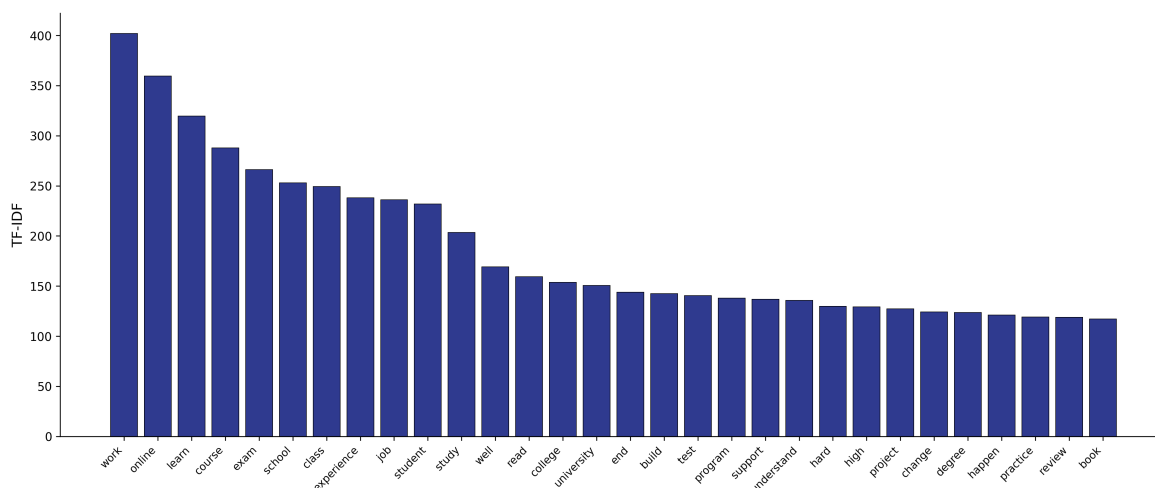


Figura 4.7: Unigramas más relevantes. Elaboración propia.

Posteriormente, al enfocarse en el análisis de los bigramas, se pueden observar los resultados en la Figura 4.8. Al examinar detalladamente dicha figura, se aprecia que los resultados son coherentes con el análisis de unigramas previo. Por un lado, se encuentran bigramas directamente asociados al contexto educativo, como *online course*, *high school*, *online school* y *class online*. También aparecen referencias a plataformas de aprendizaje reconocidas como *khan academy* y *udemy course*, lo que refleja la relevancia de estos recursos en las discusiones sobre formación *online*. Por otro lado, destaca la presencia del bigrama *mental health*, que ocupa la primera posición en términos de relevancia, lo que evidencia una preocupación significativa de los usuarios por los efectos del aprendizaje *online* sobre el bienestar psicológico. Este hallazgo se refuerza con la presencia de *health issue* entre los bigramas más frecuentes. También se observan bigramas vinculados al ámbito académico y profesional como *computer science*, *machine learning*, *community college* y *work experience*, lo que sugiere que los

usuarios discuten tanto sobre áreas de estudio específicas como sobre la aplicabilidad laboral de su formación. La presencia de *social medium* indica que los usuarios frecuentemente relacionan el aprendizaje *online* con el uso de redes sociales. Asimismo, bigramas como *practice test*, *practice exam* y *lesson learn* reflejan un interés por los métodos de estudio y evaluación. Finalmente, la aparición de *work hard*, *hard work* y *problem solve* sugiere que los usuarios valoran el esfuerzo y las habilidades de resolución de problemas como elementos clave en su experiencia de aprendizaje.

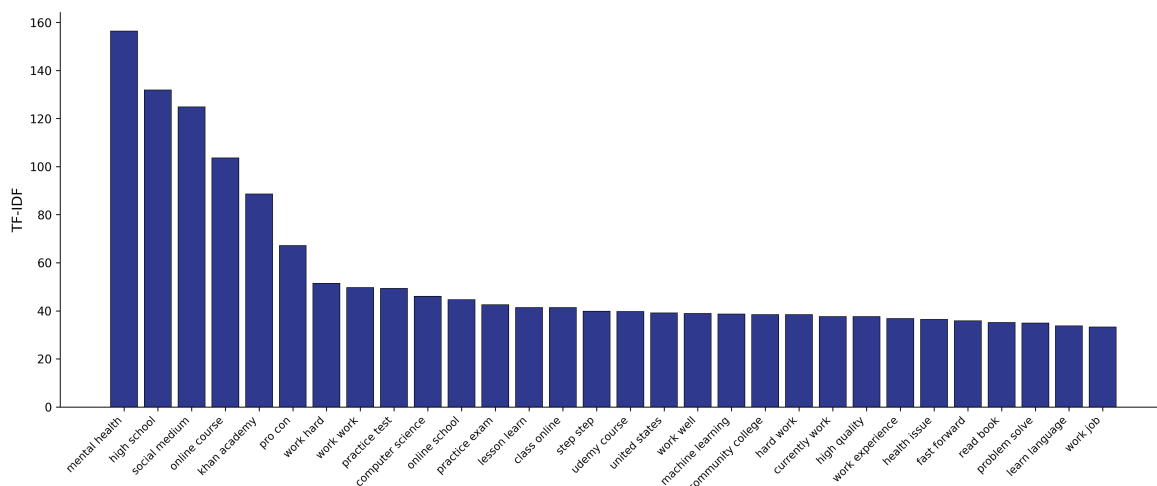


Figura 4.8: Bigramas más relevantes. Elaboración propia.

Finalmente, al llevar a cabo el análisis de trigramas y examinar los resultados presentados en la Figura 4.9, se observan resultados consistentes con los análisis anteriores. Por un lado, se identifican múltiples trigramas relacionados con la salud mental, como *mental health issue*, que ocupa la primera posición, junto con *mental health support*, *mental health professional*, *mental health problem* y *struggle mental health*, lo que evidencia que los usuarios discuten tanto sobre los problemas de bienestar como sobre los recursos de apoyo disponibles. También aparecen numerosos trigramas vinculados al contexto educativo formal, como *graduate high school*, *high school student*, *high school college* y *local community college*, reflejando discusiones sobre las distintas etapas del proceso formativo. Asimismo, la presencia de *online exam class*, *class assignment homework* y *struggle online exam* indica que la evaluación y las tareas académicas constituyen un foco importante de discusión. Por otro lado, se identifican trigramas relacionados con el ámbito tecnológico como *open source project*, *large language model* y *machine learning model*, confirmando el interés por la adquisición de competencias técnicas. La presencia de *native english speaker* sugiere discusiones sobre el aprendizaje de idiomas o sobre barreras lingüísticas en entornos educativos *online*.

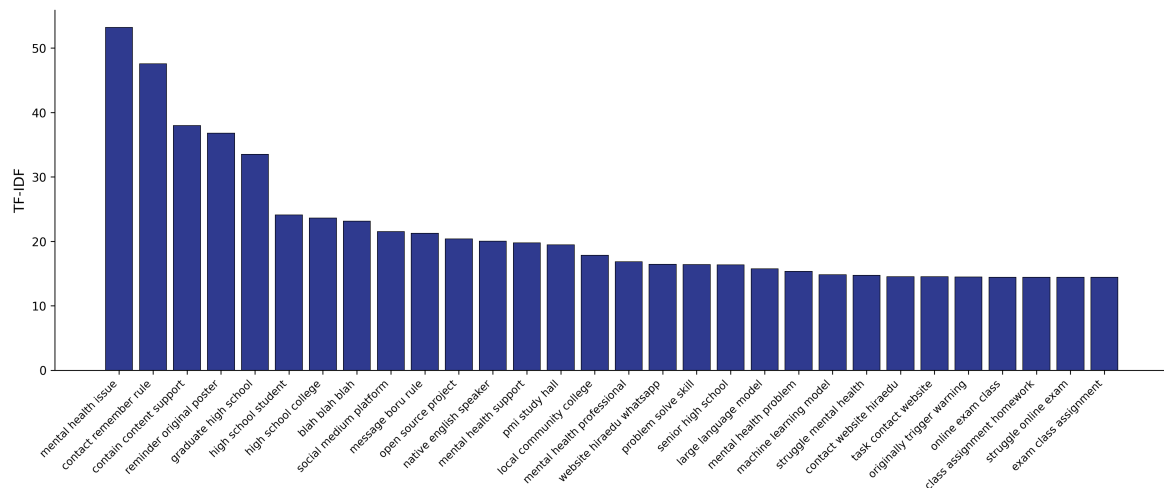


Figura 4.9: Trigramas más relevantes. Elaboración propia.

4.4. Modelado de Tópicos

El modelado de tópicos es una técnica utilizada para identificar las principales categorías de discusión presentes en un corpus de texto. En este caso, se aplicó la técnica de LDA para llevar a cabo dicho análisis en las diversas publicaciones que componen el *dataset* de Reddit. Antes de adentrarnos en el modelado propiamente dicho, como se ha comentado anteriormente, se ha llevado a cabo una fase específica de pre-procesamiento de datos destinada a preparar el corpus. Durante esta etapa, es importante recordar que se han aplicado diversas transformaciones, como la normalización de las palabras a minúsculas, la eliminación de caracteres especiales, números, menciones, *hashtags*, *stopwords* expandidas y palabras poco frecuentes o demasiado frecuentes, además de realizar la lematización. Tras este proceso de filtrado, el corpus quedó constituido por 12.893 publicaciones aptas para el análisis.

Una vez completados los pasos anteriores, se procede a determinar el número óptimo de tópicos, denotado como k , presentes en el *dataset*. Para lograr esto, se ha realizado una comparación basada en el valor de coherencia de varios modelos utilizando diferentes valores de k . En el presente estudio, se evaluaron modelos con valores de k comprendidos entre 2 y 12 tópicos, tal como se muestra en la Figura 4.10. Al analizar los resultados, se observa que los valores de coherencia oscilan entre 0,47 y 0,53, con variaciones notables según el número de tópicos seleccionado.

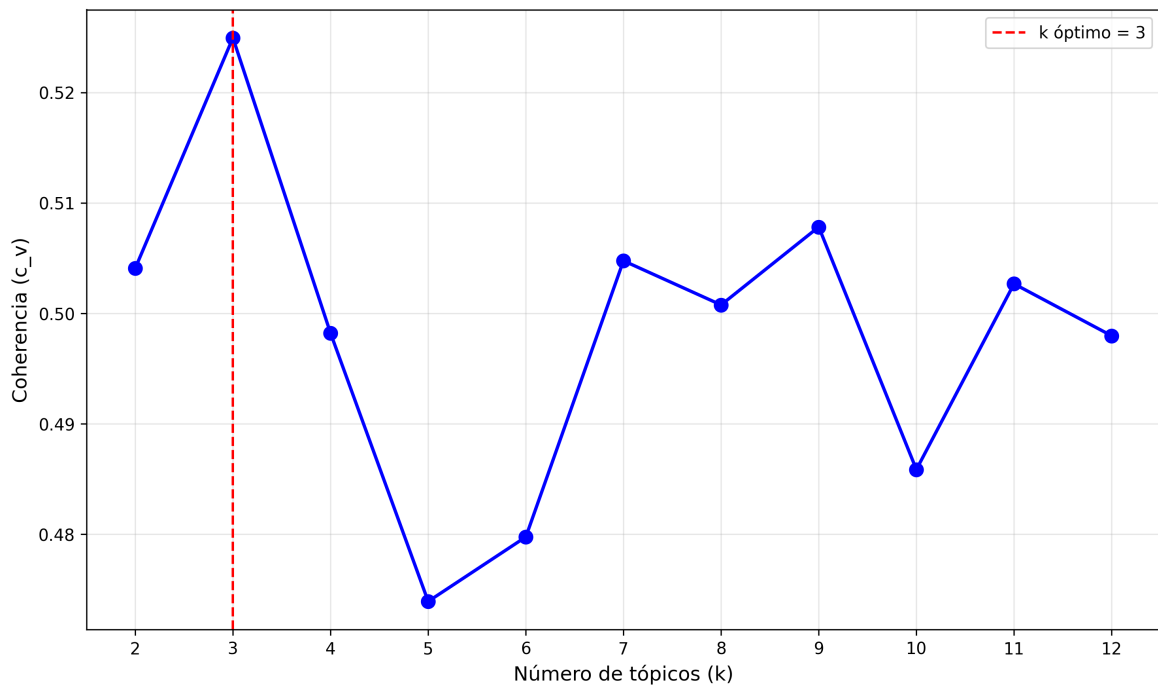


Figura 4.10: Índice de coherencia por número de tópicos. Elaboración propia.

El valor máximo de coherencia se alcanza con $k=3$ tópicos (0,5250), por lo que se seleccionó este número como el óptimo para el modelo final. Esta elección se fundamenta tanto en el criterio cuantitativo de máxima coherencia como en la interpretabilidad de los tópicos resultantes, que presentan una separación temática clara y coherente con los objetivos del estudio. Cabe destacar que valores superiores de k no solo presentaban coherencias inferiores, sino que también generaban categorías redundantes o difíciles de distinguir entre sí, lo cual dificultaría el análisis posterior. En el contexto de este estudio, se realizó una evaluación adicional mediante la visualización de la distancia intertópica, representada en la Figura 4.11. En esta representación, cada tópico aparece como un círculo cuyo tamaño es proporcional a su prevalencia estimada en el corpus, mientras que la distancia entre los círculos refleja su grado de similitud semántica. Así, los tópicos más próximos comparten una mayor proporción de vocabulario, mientras que los más alejados representan temáticas más diferenciadas.

Al examinar la figura, se observa que los tres tópicos se distribuyen de forma claramente separada en el plano bidimensional, lo que sugiere una adecuada diferenciación temática y respalda la elección de $k = 3$ como una solución interpretable y coherente. Asimismo, se puede apreciar que los Tópicos 1 y 2 presentan círculos de mayor tamaño, lo que indica una mayor prevalencia estimada dentro del corpus, mientras que el Tópico 3 aparece representado por un círculo notablemente más pequeño, reflejando un peso relativo mucho menor en el conjunto de documentos analizados.

En particular, el Tópico 1 (*Educación Online y Desarrollo Profesional*) y el Tópico 2 (*Experiencias Personales y Bienestar Emocional*) presentan tamaños similares y constitu-

yen conjuntamente más del 96 % del corpus, lo que refleja que estas dos dimensiones son las predominantes en las discusiones sobre aprendizaje *online* en Reddit. Por su parte, el Tópico 3 (*Exámenes y Evaluación Online*) muestra la menor prevalencia relativa, lo que indica su carácter más específico y focalizado dentro del *dataset*. Por su parte, la separación observada entre los círculos sugiere que el modelo ha logrado identificar categorías temáticas razonablemente diferenciadas, lo que aporta solidez a la interpretación de los tópicos obtenidos.

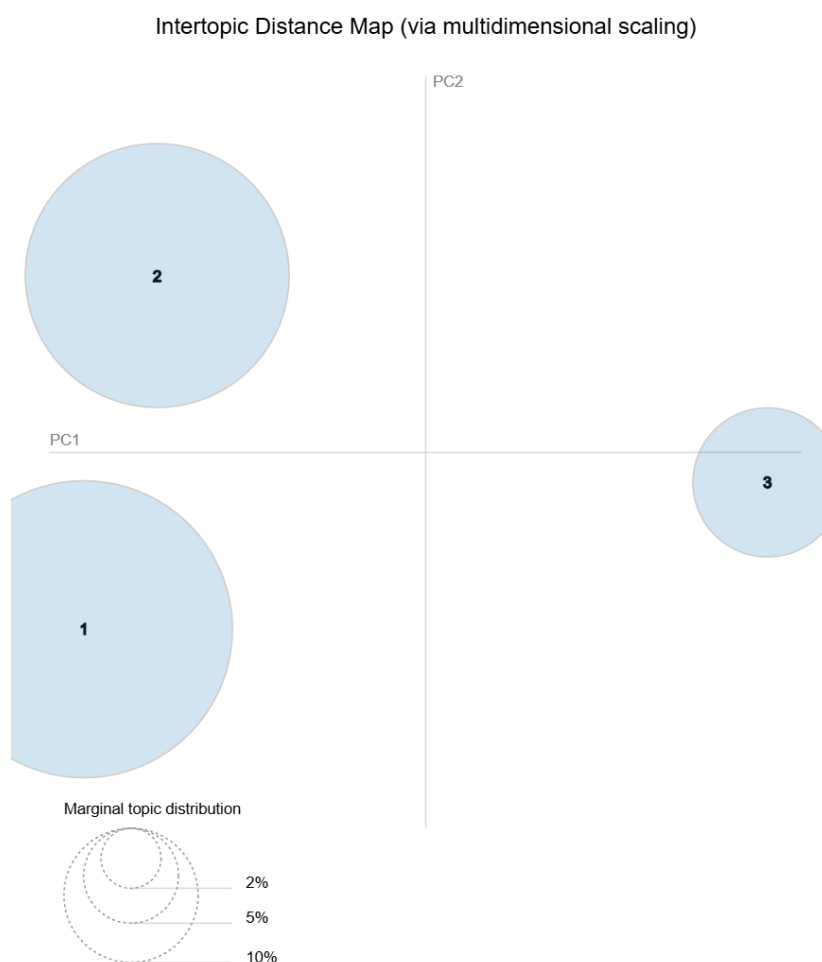


Figura 4.11: Distancia intertópica $K=3$. Elaboración propia.

Una vez completado el modelado de los 3 tópicos, se llevó a cabo un proceso de etiquetado para cada uno de ellos, con el objetivo de identificar su categoría principal. Para ello, se evaluaron las palabras más frecuentes, así como los principales bigramas y trigramas asociados a cada tópico. Estos elementos clave permitieron determinar la idea principal que representa cada tópico. En la Tabla 4.1 se presenta un resumen que muestra las principales categorías de cada tópico, así como las palabras clave, bigramas y trigramas más relevantes. Tras su análisis, se identificaron los siguientes tópicos principales en el *dataset* de este Trabajo de Fin de Grado:

- **Tópico 1:** Educación *Online* y Desarrollo Profesional
- **Tópico 2:** Experiencias Personales y Bienestar Emocional
- **Tópico 3:** Exámenes y Evaluación *Online*

En primer lugar, el Tópico 1 se enfoca en las publicaciones relacionadas con la educación *online* y el desarrollo profesional, constituyendo una de las categorías principales del corpus con 6.124 publicaciones (47,5 %). Como se puede apreciar en las palabras clave identificadas, este tópico incluye términos directamente asociados al aprendizaje en línea como “*work*”, “*learn*”, “*experience*”, “*system*”, “*build*”, “*design*”, “*course*”, “*high*”, “*base*” y “*model*”. Los bigramas más relevantes refuerzan esta interpretación, destacando expresiones como “*online course*”, “*khan academy*”, “*udemy course*”, “*machine learning*” y “*practice exam*”. Los trigramas confirman la orientación hacia el aprendizaje técnico y profesional, con expresiones como “*open source project*”, “*problem solve skill*”, “*machine learning model*”, “*large language model*” y “*native english speaker*”. Este tópico captura las discusiones sobre cursos en línea, plataformas educativas, desarrollo de habilidades técnicas y aprendizaje de tecnologías emergentes.

El Tópico 2, con 6.276 publicaciones (48,7 %), agrupa las discusiones relacionadas con experiencias personales y bienestar emocional, constituyendo la categoría más amplia del corpus. Las palabras clave más frecuentes incluyen “*work*”, “*end*”, “*happen*”, “*learn*”, “*mean*”, “*place*”, “*well*”, “*away*”, “*human*” y “*school*”, términos que sugieren narrativas personales y reflexiones sobre situaciones vividas. Los bigramas más relevantes evidencian una fuerte presencia de contenido relacionado con la salud mental y las experiencias escolares, incluyendo “*mental health*”, “*high school*”, “*health issue*” y “*online school*”. Los trigramas confirman esta orientación, destacando expresiones como “*mental health issue*”, “*graduate high school*”, “*struggle mental health*” y “*high school student*”. Este tópico refleja cómo los usuarios de Reddit comparten sus experiencias emocionales asociadas al aprendizaje *online*.

Finalmente, el Tópico 3 se centra en exámenes y evaluación *online*, con 493 publicaciones (3,8 %). Aunque representa la categoría más pequeña del corpus, su especificidad temática es notable. Las palabras clave más representativas incluyen “*exam*”, “*test*”, “*online*”, “*homework*”, “*student*”, “*class*”, “*ati*”, “*math*”, “*university*” y “*cheat*”, términos que reflejan las preocupaciones específicas relacionadas con la evaluación en entornos virtuales. Los bigramas confirman esta orientación, destacando expresiones como “*online exam*”, “*khan academy*”, “*insurance exam*”, “*ati exam*”, “*lockdown browser*” y “*practice test*”. Los trigramas más relevantes incluyen “*exam class assignment*”, “*struggle online exam*” y “*targeted medical surgical*”. Este tópico captura las discusiones sobre la experiencia de realizar exámenes en línea, las herramientas de supervisión remota como *Lockdown Browser*, la preparación para pruebas estandarizadas y las preocupaciones éticas relacionadas con la integridad académica en entornos digitales.

Tabla 4.1: Principales palabras clave, bigramas y trigramas por tópico.

Tópico	Categoría	Palabras Clave	Bigramas	Trigramas
1	Educación <i>Online</i> y Desarrollo Profesional	“work”, “learn”, “experience”, “system”, “build”, “design”, “course”, “high”, “base” y “model”	“online course”, “khan academy”, “udemy course”, “machine learning”, “practice exam” y “mental health”	“open source project”, “problem solve skill”, “machine learning model”, “large language model” y “native english speaker”
2	Experiencias Personales y Bienestar Emocional	“work”, “end”, “happen”, “learn”, “mean”, “well”, “place”, “away”, “human” y “school”	“mental health”, “high school”, “social medium”, “health issue”, “online school” y “learn lesson”	“mental health issue”, “graduate high school”, “struggle mental health”, “high school college” y “high school student”
3	Exámenes y Evaluación <i>Online</i>	“exam”, “test”, “online”, “homework”, “student”, “class”, “ati”, “math”, “university” y “cheat”	“online exam”, “khan academy”, “insurance exam”, “ati exam”, “lockdown browser” y “practice test”	“exam class assignment”, “struggle online exam”, “targeted medical surgical”, “online exam class” y “class assignment homework”

En la Figura 4.12 se presenta la frecuencia absoluta de cada tópico dentro del *dataset*. Se puede observar que las publicaciones se distribuyen de manera relativamente equilibrada entre los dos primeros tópicos, con el Tópico 2 (Experiencias Personales y Bienestar Emocional) presentando una ligera predominancia sobre el Tópico 1 (Educación *Online* y Desarrollo Profesional). El Tópico 3 (Exámenes y Evaluación *Online*), aunque minoritario, representa una categoría temática específica y relevante para comprender las preocupaciones de los estudiantes respecto a la evaluación en entornos virtuales.

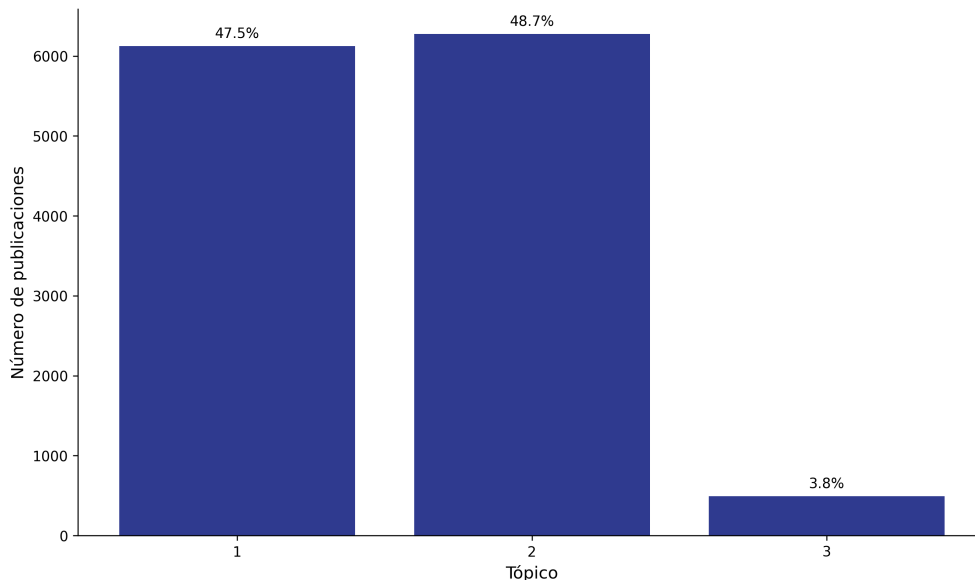


Figura 4.12: Distribución de publicaciones por tópico. Elaboración propia.

4.5. Análisis de Sentimientos

Una vez identificados los tópicos principales del corpus, se realizó un análisis de sentimientos con el objetivo de estimar la polaridad emocional de las publicaciones asociadas a cada categoría temática. Para ello, se empleó VADER, cuya fundamentación y adecuación para el análisis de texto informal en redes sociales se describen en la Sección 3.5. El análisis de sentimientos se aplicó sobre las publicaciones que contaban tanto con el texto original como con la asignación de tópico derivada del modelado LDA. Para cada publicación, se calculó la puntuación compuesta de sentimiento. Adicionalmente, se convirtieron los emoticonos presentes en el texto a su representación textual, lo cual permite que VADER capture la carga emocional de estos elementos visuales frecuentemente utilizados en plataformas como Reddit.

La Figura 4.13 muestra la distribución de las puntuaciones compuestas de sentimiento para el conjunto total de publicaciones mediante un histograma acompañado de su correspondiente curva de densidad. En términos generales, la distribución presenta una forma claramente bimodal, con dos concentraciones principales situadas en los extremos del rango de puntuación. El pico más elevado se localiza en el extremo positivo, en valores próximos a +1, lo que indica una elevada concentración de publicaciones con un sentimiento marcadamente positivo. Asimismo, se observa un segundo pico en el extremo negativo, cercano a -1, aunque de menor altura, lo que evidencia también la existencia de un grupo relevante de publicaciones con una carga emocional fuertemente negativa.

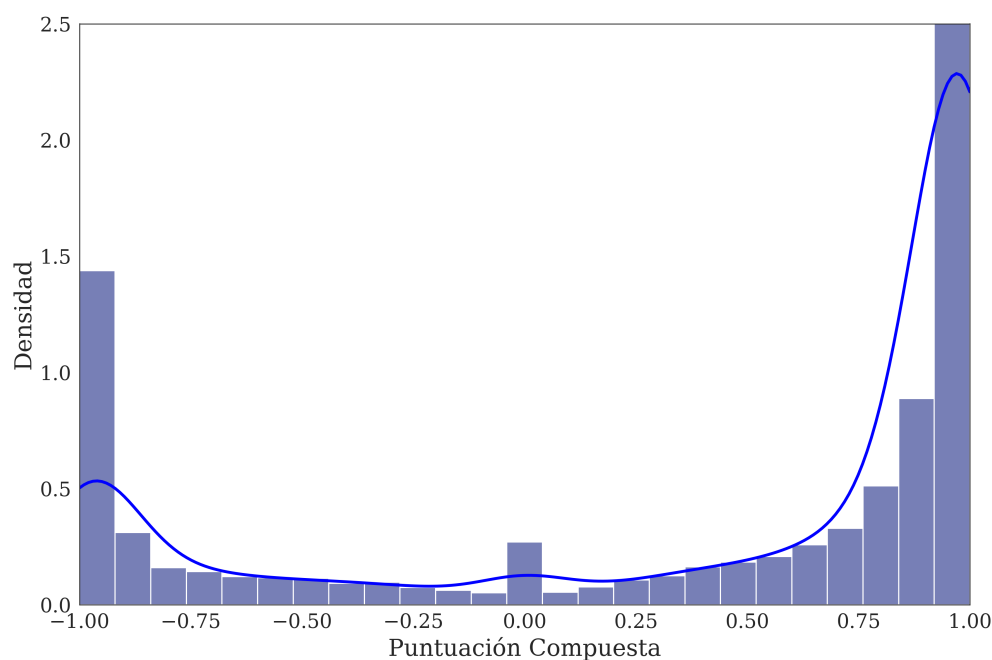


Figura 4.13: Distribución de las puntuaciones de sentimiento en el corpus. Elaboración propia.

La comparación entre ambos extremos permite apreciar que la concentración de observaciones es mayor en la zona positiva que en la negativa. Esto sugiere que, aunque en el corpus están presentes tanto valoraciones muy favorables como muy desfavorables, predominan las publicaciones con una orientación positiva. Por otro lado, la parte central de la distribución, correspondiente a valores próximos a la neutralidad, presenta una densidad considerablemente menor, lo que indica que las publicaciones neutrales son menos frecuentes que aquellas con una polaridad emocional más definida. Este patrón refleja una notable polarización emocional en el discurso analizado. Los usuarios tienden a expresar opiniones intensas sobre el aprendizaje *online*, ya sea en sentido positivo o negativo, en lugar de mantener posturas intermedias. Esta tendencia resulta coherente con la dinámica habitual de las comunidades digitales y las redes sociales, donde las experiencias más satisfactorias o más frustrantes son precisamente las que generan una mayor probabilidad de ser compartidas.

La Figura 4.14 complementa esta visualización mediante un diagrama de caja que resume los estadísticos principales de la distribución de puntuaciones. Se observa que la mediana (línea central de la caja) se sitúa en torno a 0,95, confirmando la tendencia positiva predominante. La media, representada por el rombo, se ubica aproximadamente en 0,50, un valor inferior a la mediana, lo cual refleja el efecto de las puntuaciones negativas extremas sobre el promedio. El rango intercuartílico (altura de la caja) abarca desde aproximadamente 0,15 hasta 1,0, indicando que el 50 % central de las publicaciones presenta puntuaciones positivas. Los bigotes se extienden hasta los valores extremos de -1 y +1, confirmando la presencia de publicaciones con sentimientos fuertemente polarizados en ambas direcciones.

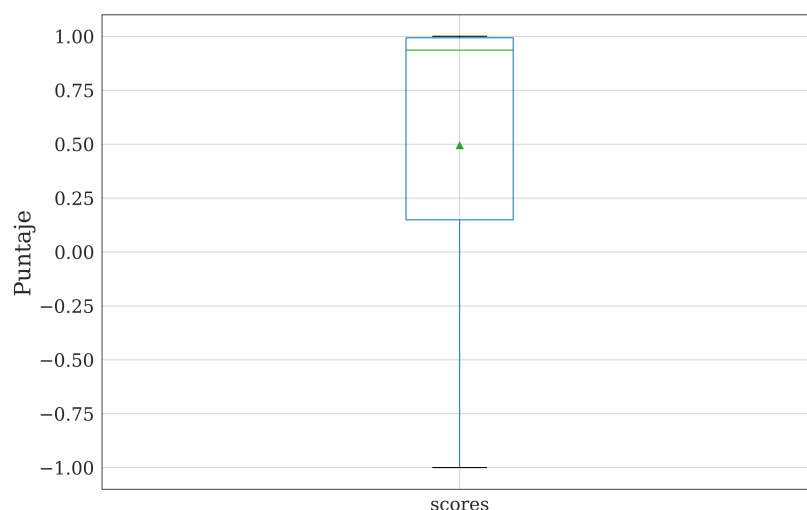


Figura 4.14: Distribución de la puntuación compuesta. Elaboración propia.

La predominancia del pico positivo sobre el negativo indica que, en términos generales, las percepciones sobre el aprendizaje *online* en las comunidades de Reddit analizadas son mayoritariamente favorables, aunque existe una proporción relevante de experiencias negativas que no debe ser ignorada en el diseño de estrategias pedagógicas.

Con el objetivo de profundizar en el análisis de las percepciones de los usuarios sobre el aprendizaje *online*, se compararon las puntuaciones de sentimiento según el tópico asignado a cada publicación. La Tabla 4.2 recoge los estadísticos descriptivos de las puntuaciones compuestas para cada uno de los tres tópicos identificados en la fase previa de modelado.

Tópico	Media	Desv. Típica	Mín.	Máx.	Coef. Asimetría
1	0,692	0,538	-0,9999	1,0	-1,985
2	0,310	0,842	-1,0000	1,0	-0,670
3	0,427	0,613	-0,9973	1,0	-1,011

Tabla 4.2: Estadísticos descriptivos de las puntuaciones de sentimiento por tópico.
Elaboración propia.

Los resultados revelan diferencias notables en la polaridad emocional entre los distintos tópicos. El **Tópico 1** (Educación *Online* y Desarrollo Profesional) presenta la puntuación media más elevada (0,692), lo cual sugiere que las discusiones sobre cursos en línea, plataformas educativas y desarrollo de habilidades técnicas tienden a expresarse en términos particularmente positivos. Este resultado puede explicarse por el entusiasmo característico de los usuarios que comparten logros académicos, recomendaciones de recursos formativos y experiencias satisfactorias con plataformas como Khan Academy o Udemy. Además, este tópico presenta la desviación típica más baja (0,538), lo que indica una mayor homogeneidad en las opiniones expresadas.

El **Tópico 3** (Exámenes y Evaluación *Online*) exhibe una puntuación media positiva intermedia de 0,427, reflejando una percepción moderadamente favorable hacia las experiencias de evaluación en entornos virtuales. Sin embargo, la presencia de términos como “*cheat*” y “*lockdown browser*” en este tópico sugiere que también existen preocupaciones relacionadas con la integridad académica y las herramientas de supervisión remota, lo cual podría explicar la puntuación media inferior respecto al Tópico 1.

En contraste, el **Tópico 2** (Experiencias Personales y Bienestar Emocional) muestra la puntuación media más baja del corpus (0,310), aunque sigue siendo ligeramente positiva. Este resultado es coherente con la naturaleza del tópico, que agrupa discusiones sobre salud mental, autoestima, aislamiento social y dificultades emocionales asociadas al aprendizaje *online*. La elevada desviación típica (0,842) indica una alta variabilidad en las opiniones, lo cual refleja la diversidad de experiencias personales compartidas por los usuarios: mientras algunos expresan superación y resiliencia, otros manifiestan frustración o malestar emocional. El coeficiente de asimetría proporciona información adicional sobre la forma de la distribución de sentimientos en cada tópico. Todos los tópicos presentan valores negativos, lo cual indica una asimetría hacia la izquierda, es decir, una cola más larga en la zona de sentimientos negativos. El Tópico 1 muestra el coeficiente de asimetría más pronunciado (-1,985), lo que refleja una fuerte concentración de puntuaciones en valores positivos con casos aislados

de sentimientos muy negativos. En contraste, el Tópico 2 presenta el coeficiente de asimetría más cercano a cero (-0,670), indicando una distribución más simétrica y confirmando la mayor variabilidad emocional observada en las discusiones sobre bienestar personal y salud mental.

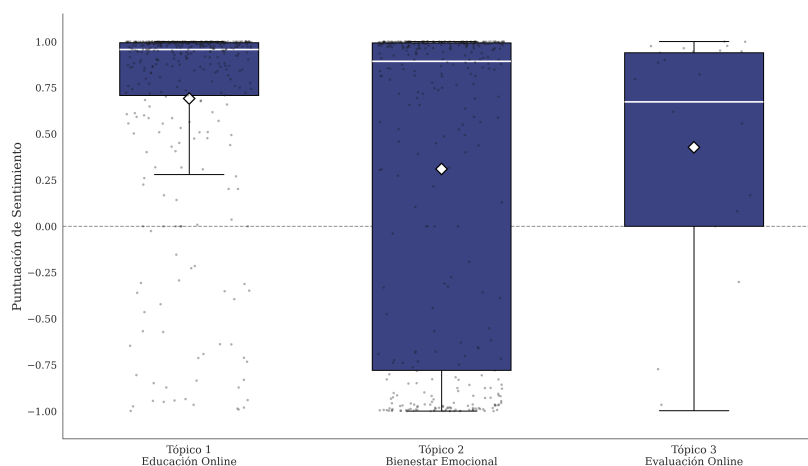


Figura 4.15: Diagrama de caja de las puntuaciones de sentimiento del corpus. Elaboración propia.

La Figura 4.15 presenta la distribución de las puntuaciones de sentimiento mediante diagramas de caja para cada tópico. Se puede observar que todos los tópicos presentan una mediana situada en la zona positiva del espectro, aunque con diferentes grados de dispersión. El Tópico 2 destaca por su mayor rango intercuartílico, lo cual confirma la heterogeneidad de las experiencias emocionales asociadas a este grupo temático. Asimismo, se aprecia que el Tópico 1 presenta una distribución más concentrada en valores positivos, mientras que el Tópico 2 muestra una distribución más simétrica que se extiende hacia valores negativos. El Tópico 3, por su parte, presenta una distribución intermedia con una mediana elevada pero con presencia de valores negativos.

Con el objetivo de examinar la evolución temporal del sentimiento expresado en Reddit en relación con el aprendizaje online, se ha calculado la puntuación media diaria de VADER para cada uno de los tres tópicos identificados. Como se muestra en la Figura 4.16, la línea fina recoge el sentimiento diario bruto, mientras que la línea gruesa representa la media móvil de 7 días, que suaviza el ruido puntual y permite observar la tendencia emocional subyacente. El área sombreada refleja la variabilidad diaria del sentimiento a lo largo del período analizado.

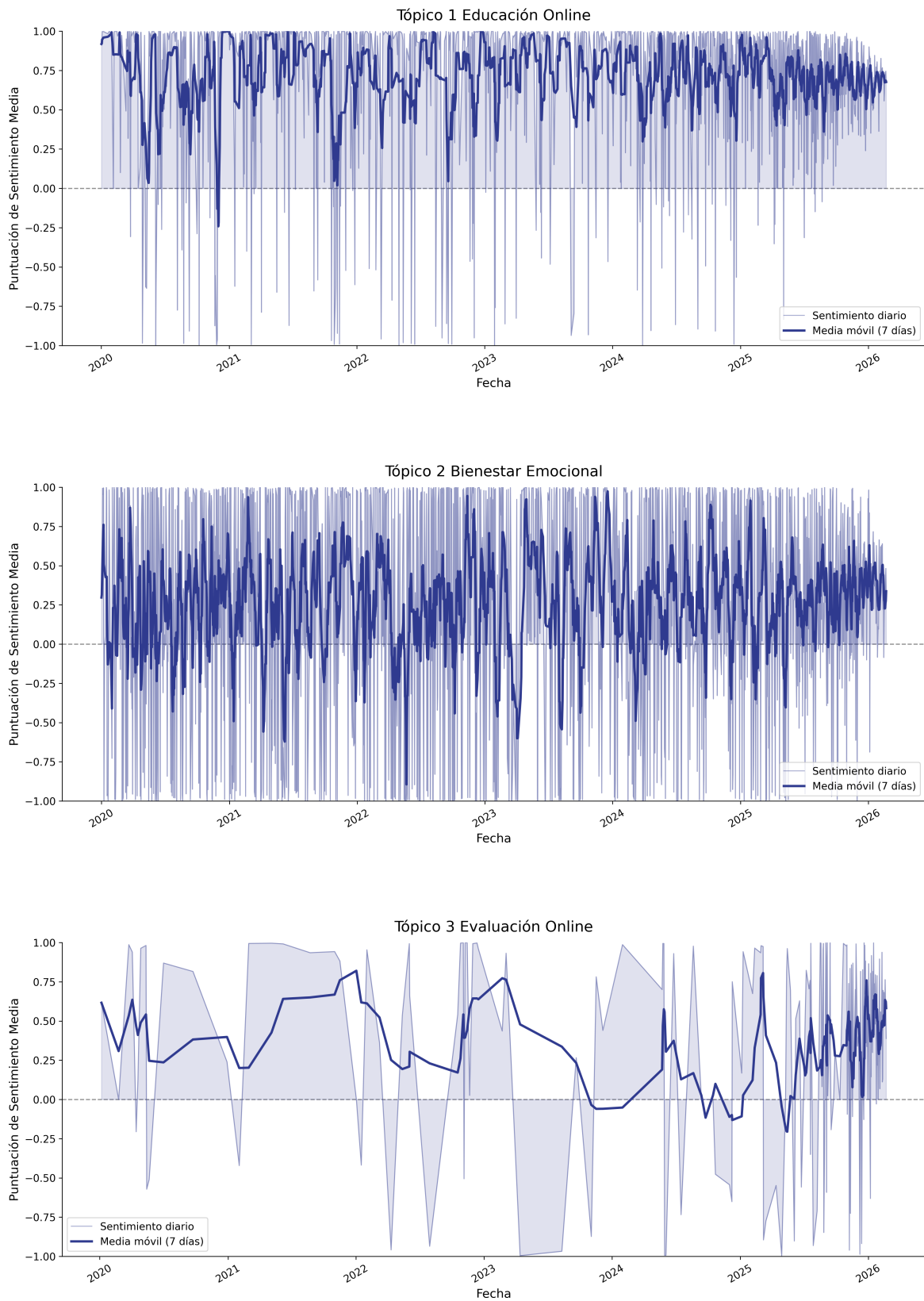


Figura 4.16: Evolución temporal del sentimiento por tópico (media móvil 7 días).
Elaboración propia.

En el caso del **Tópico 1 (Educación Online)**, la evolución temporal revela el patrón más estable y positivo de los tres tópicos. La media móvil se mantiene de forma prácticamente ininterrumpida por encima de 0,5 durante todo el período 2020-2026, sin registrar en ningún momento una tendencia sostenida hacia el terreno negativo. Si bien el sentimiento diario presenta picos descendentes puntuales que alcanzan valores negativos, estos son excepcionales y no alteran la tendencia general. Este comportamiento sugiere que las percepciones sobre aspectos generales de la educación *online*, son consistentemente favorables entre los usuarios de Reddit, con independencia del año considerado.

El **Tópico 2 (Bienestar Emocional)** presenta un patrón claramente más volátil. A diferencia del Tópico 1, la media móvil oscila con mayor amplitud a lo largo del tiempo, aproximándose con frecuencia al umbral de cero e incluso cruzándolo en determinados períodos, particularmente en torno a 2022 y 2023. El sentimiento diario bruto registra caídas negativas frecuentes y de mayor profundidad que en el tópico anterior, lo que refleja la naturaleza dual de las publicaciones sobre bienestar.

Por último, el **Tópico 3 (Evaluación Online)** es el que muestra una dinámica temporal más irregular y discontinua. La menor densidad de publicaciones asociadas a este tópico se traduce en amplias zonas del gráfico con escasa cobertura temporal, lo que hace que la media móvil presente saltos bruscos y períodos de alta inestabilidad. A partir de 2023 se observan episodios más frecuentes de sentimiento negativo, con la media móvil cruzando el cero en varias ocasiones y alcanzando valores negativos sostenidos en 2024 y 2025.

Capítulo 5

Conclusiones

El aprendizaje *online* ha adquirido una relevancia creciente en los últimos años, consolidándose como una de las principales transformaciones del sistema educativo global. Este proceso, acelerado por la pandemia de COVID-19, no solo ha permitido garantizar la continuidad educativa en contextos de crisis, sino que también ha impulsado modelos pedagógicos más flexibles e innovadores. En este escenario, comprender las percepciones del estudiantado sobre esta modalidad resulta fundamental, ya que se requiere conocer sus experiencias, valoraciones y preocupaciones para orientar la implementación de políticas educativas y estrategias pedagógicas más eficaces. Las redes sociales, en particular, se han convertido en una herramienta útil y efectiva para extraer esta información, al proporcionar una plataforma donde los usuarios pueden expresar opiniones y compartir experiencias de manera libre y espontánea. Al analizar las discusiones y sentimientos expresados en estas plataformas, es posible obtener información relevante sobre las actitudes del estudiantado hacia el aprendizaje *online*, lo que puede guiar la toma de decisiones y el diseño de experiencias educativas más alineadas con las necesidades reales del alumnado.

En este contexto, el presente Trabajo de Fin de Grado se ha centrado en brindar respuestas en este campo. Se ha logrado comprender la percepción del estudiantado acerca del aprendizaje *online* a través de la red social Reddit, mediante la realización de un modelado de tópicos y un análisis de sentimientos. Este análisis se basó en un conjunto de datos recopilados de diversas comunidades temáticas vinculadas al ámbito educativo, un espacio que resulta especialmente adecuado para capturar discursos auténticos y no mediados.

Con el fin de alcanzar este objetivo, se realizó una revisión de la literatura que abarca diversos estudios relacionados con el análisis de percepciones sobre el aprendizaje *online* a través de redes sociales. Gracias a este análisis, se concluyó que la técnica LDA ha sido ampliamente aplicada para el modelado de tópicos, debido a su capacidad para analizar grandes volúmenes de datos, su flexibilidad y los resultados altamente interpretables que ofrece. Además, se identificó que la técnica más adecuada para el análisis de sentimientos en este estudio es la herramienta VADER. Este modelo se destaca por su léxico detallado, que considera as-

pectos como la puntuación, las mayúsculas y los elementos propios del lenguaje informal, garantizando así un análisis fiable en mensajes breves y emocionalmente cargados como los de Reddit.

Posteriormente, se siguió un enfoque metodológico para identificar los temas de discusión y las percepciones y sentimientos de los usuarios en relación con el aprendizaje *online*. Para ello, se recopilaron 12.893 publicaciones en inglés de la red social Reddit mediante su API oficial. Seguidamente, se realizó el preprocesamiento de los datos para limpiar y garantizar la calidad del conjunto de datos. Tras esta fase de limpieza, se procedió al análisis de n-gramas para identificar los términos más relevantes del conjunto de datos, analizando unigramas, bigramas y trigramas relacionados con la temática del aprendizaje *online*. Finalmente, se llevó a cabo el modelado de tópicos utilizando la técnica LDA y se realizó un análisis de sentimientos empleando VADER, cuantificando las emociones asociadas a las publicaciones del conjunto de datos analizado.

En relación al modelado de tópicos, se identificaron tres tópicos principales. El primer tópico, denominado Educación *Online* y Desarrollo Profesional, se centra en las discusiones sobre cursos en línea, plataformas educativas y aprendizaje de tecnologías emergentes. Este tópico refleja el interés de los usuarios por la formación continua y el desarrollo de habilidades técnicas, haciendo referencia a plataformas reconocidas como Khan Academy o Udemy, así como a ámbitos profesionales como el *machine learning* o la programación. El segundo tópico, clasificado bajo la categoría de Experiencias Personales y Bienestar Emocional, aborda las narrativas compartidas por los usuarios sobre sus vivencias en entornos virtuales, incluyendo cuestiones relacionadas con la salud mental, las dificultades emocionales y los procesos de superación académica. Este tópico resalta la importancia de la dimensión afectiva del aprendizaje *online*, poniendo de manifiesto que la experiencia educativa digital trasciende los aspectos estrictamente académicos o tecnológicos. Finalmente, el tercer tópico, identificado como Exámenes y Evaluación *Online*, se refiere a las discusiones sobre la realización de pruebas en entornos virtuales, las herramientas de supervisión remota como *Lockdown Browser* y las preocupaciones relativas a la integridad académica. Este tópico destaca la relevancia de la evaluación como uno de los elementos más sensibles de la experiencia educativa en línea.

En cuanto al análisis de sentimientos, tras la evaluación de los diferentes tópicos y el tratamiento específico de los datos para su limpieza, se observó una notable predominancia de sentimientos positivos asociados a las publicaciones del *dataset*, aunque también se identificó una presencia destacable de sentimientos negativos, lo que refleja una marcada polarización emocional en el discurso analizado. Este resultado podría revelar un apoyo generalizado al aprendizaje *online* por parte de los usuarios de Reddit, destacando al mismo tiempo la existencia de experiencias críticas que conviven con valoraciones favorables. Asimismo, se realizó un análisis de la distribución del sentimiento en cada uno de los tres tópicos identificados. A pesar de la predominancia de sentimientos positivos en todos ellos, se encontró que el tópico

2, Experiencias Personales y Bienestar Emocional, presentaba una mayor variabilidad emocional, con una notable coexistencia de sentimientos positivos y negativos. Este tópico, que recoge las narrativas personales del estudiantado sobre sus vivencias en entornos virtuales, refleja la heterogeneidad de las experiencias emocionales, mientras algunos usuarios comparan procesos de superación, otros manifiestan frustración, ansiedad o malestar asociados a la modalidad *online*. Este sentimiento mixto sugiere que la experiencia educativa digital, aunque valorada positivamente en términos generales, lleva aparejadas tensiones emocionales significativas que merecen una atención específica.

Por su parte, el tópico 1, denominado Educación *Online* y Desarrollo Profesional, no solo concentra una parte importante de las publicaciones, sino también el mayor sentimiento positivo asociado, con una media superior respecto al resto. Esto refleja un fuerte apoyo a las iniciativas formativas digitales, a las plataformas de aprendizaje y al desarrollo de competencias profesionales a través de recursos en línea. Los usuarios parecen estar especialmente a favor de las oportunidades que ofrece el aprendizaje *online* para la adquisición de nuevas habilidades y la mejora profesional. El tópico 3, relativo a la evaluación en entornos virtuales, presentó por su parte un sentimiento intermedio, reflejando percepciones moderadamente favorables, aunque con la presencia de preocupaciones específicas vinculadas a las herramientas de supervisión y a la integridad académica.

Como línea de investigación futura, resultaría sumamente interesante ampliar el análisis a otras plataformas digitales y comunidades en línea para contrastar los patrones identificados en Reddit con los de otros entornos conversacionales, lo que permitiría obtener una mayor cantidad de datos e incrementar la comprensión acerca de cómo el estudiantado percibe el aprendizaje online en distintos contextos. Asimismo, la incorporación de técnicas de análisis de sentimientos más sofisticadas, como modelos basados en transformers o enfoques de detección de emociones específicas, podría aportar una granularidad mayor en la caracterización del componente afectivo del discurso. Estos análisis más detallados, basados en fuentes y técnicas más diversas, podrían proporcionar una imagen más completa y precisa de la percepción del estudiantado en relación con el aprendizaje online y sentarían las bases para desarrollar estrategias pedagógicas e institucionales más efectivas, que atiendan tanto las dimensiones académicas como las emocionales de la experiencia educativa digital.

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, Pablo Soriano Fernández, estudiante de E2-Analytics de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado “Análisis de Opiniones sobre el Aprendizaje Online en Comunidades Digitales”, 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. Brainstorming de ideas de investigación: Utilizado para idear y esbozar posibles áreas de investigación.
2. Crítico: Para encontrar contra-argumentos a una tesis específica que pretendo defender.
3. Referencias: Usado conjuntamente con otras herramientas, como Science, para identificar referencias preliminares que luego he contrastado y validado.
4. Metodólogo: Para descubrir métodos aplicables a problemas específicos de investigación.
5. Interpretador de código: Para realizar análisis de datos preliminares.
6. Estudios multidisciplinares: Para comprender perspectivas de otras comunidades sobre temas de naturaleza multidisciplinar.
7. Constructor de plantillas: Para diseñar formatos específicos para secciones del trabajo.
8. Corrector de estilo literario y de lenguaje: Para mejorar la calidad lingüística y estilística del texto.

9. Generador previo de diagramas de flujo y contenido: Para esbozar diagramas iniciales.
10. Sintetizador y divulgador de libros complicados: Para resumir y comprender literatura compleja.
11. Generador de datos sintéticos de prueba: Para la creación de conjuntos de datos ficticios.
12. Generador de problemas de ejemplo: Para ilustrar conceptos y técnicas.
13. Revisor: Para recibir sugerencias sobre cómo mejorar y perfeccionar el trabajo con diferentes niveles de exigencia.
14. Generador de encuestas: Para diseñar cuestionarios preliminares.
15. Traductor: Para traducir textos de un lenguaje a otro.

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 2026

Firma: Pablo Soriano Fernández

Referencias

- Alharbi, N. H., y Alkhateeb, J. H. (2021). Sentiment analysis of arabic tweets related to covid-19 using deep neural network. En *2021 international congress of advanced technology and engineering (icoten)* (pp. 1–11).
- Blei, D. M., Ng, A. Y., y Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993–1022.
- Business of Apps. (2024). *Reddit revenue and usage statistics (2024)*. Descargado de <https://www.businessofapps.com/data/reddit-statistics/> (Accessed: 2025-02-19)
- CIPHR. (2022). *Uk & global virtual learning statistics 2022*. Descargado de <https://www.ciphr.com/infographics/virtual-learning-statistics> (Survey data from CIPHR via Censuswide, January 2022)
- Cranfield, D. J., Tick, A., Venter, I. M., Blignaut, R. J., y Renaud, K. (2021). Higher education students' perceptions of online learning during covid-19—a comparative study. *Education Sciences*, 11(8), 403.
- Devi, S., Dhaval, C., Moharkar, L., y Khanvilkar, S. (2022). Impact of online education and sentiment analysis from twitter data using topic modeling algorithms. *International Journal of Applied Sciences and Smart Technologies*, 4(1), 21–34.
- Dhawan, S. (2020). Online learning: A panacea in the time of covid-19 crisis. *Journal of educational technology systems*, 49(1), 5–22.
- ElSayad, G. (2024). Higher education students' learning perception in the blended learning community of inquiry. *Journal of Computers in Education*, 11(4), 1061–1088.
- Laureate, C. D. P., Buntine, W., y Linger, H. (2023). A systematic review of the use of topic models for short text social media analysis. *Artificial Intelligence Review*, 56(12), 14223–14255.
- Li, S., Xie, Z., Chiu, D. K., y Ho, K. K. (2023). Sentiment analysis and topic modeling regarding online classes on the reddit platform: educators versus learners. *Applied Sciences*, 13(4), 2250.
- Manca, S., y Ranieri, M. (2016). Facebook and the others. potentials and obstacles of social media for teaching in higher education. *Computers & education*, 95, 216–230.
- Means, B., Toyama, Y., Murphy, R., Bakia, M., y Jones, K. (2009). Evaluation of evidence-

- based practices in online learning: A meta-analysis and review of online learning studies.
- Mujahid, M., Lee, E., Rustam, F., Washington, P. B., Ullah, S., Reshi, A. A., y Ashraf, I. (2021). Sentiment analysis and topic modeling on tweets about online education during covid-19. *Applied Sciences*, 11(18), 8438.
- Rosamma, K., y Rosamma Jr, K. (2024). Analyzing online conversations on reddit: A study of stress and anxiety through topic modeling and sentiment analysis. *Cureus*, 16(9).
- Soriano, P. (2026). *Pablo_Soriano-TFG*. Descargado de https://github.com/pablosoriano03-collab/Pablo_Soriano-TFG
- Stracqualursi, L., y Agati, P. (2022). Tweet topics and sentiments relating to distance learning among italian twitter users. *Scientific reports*, 12(1), 9163.
- The Business Research Company. (2024). *E-learning global market report 2025*. <https://www.thebusinessresearchcompany.com/report/e-learning-global-market-report>. (Accessed: 2025-02-17)
- UNESCO. (2020). *Policy brief: Education during covid-19 and beyond*. <https://unsdg.un.org/resources/policy-brief-education-during-covid-19-and-beyond>. (United Nations Sustainable Development Group)
- Youvan, D. C. (2024). Understanding sentiment analysis with vader: a comprehensive overview and application. *AI and Data Science Journal*.