



Facultad de Ingeniería

**MINERÍA DE DATOS PARA EL ANÁLISIS DE SENTIMIENTO EN REDES
SOCIALES: TECNICAS, EVALUACION PRACTICA Y BASE PARA FUTURAS
INVESTIGACIONES MULTIDISCIPLINARIAS.**

Proyecto de título para optar al grado de Ingeniero en informática con mención en
gestión de la información.

Sebastián Aníbal Zapata Fuenzalida

Profesora Guía: Delia Ibacache

San Felipe, Chile

2023

DEDICATORIA

Querida familia,

En este recorrido académico, ustedes han sido mi faro constante, iluminando el camino con su apoyo incondicional y los valores que nos unen: respeto, honestidad, humildad, perseverancia y responsabilidad. Cada paso que he dado en este viaje ha sido guiado por el legado de nuestra familia, una brújula que siempre apunta hacia el norte verdadero de nuestros principios.

A lo largo de esta travesía, la presencia divina ha sido palpable, como un faro en la oscuridad, recordándome que el aprendizaje es un regalo sagrado. En los momentos difíciles, su amor y mi fe en Dios han sido el sustento que ha fortalecido mi espíritu, recordándome que cada desafío es una oportunidad para crecer, independientemente del resultado, ya que en todo aprendizaje hay un tesoro.

Al contemplar el vasto universo de posibilidades que se despliega ante mí, no puedo dejar de reflexionar sobre el poder del conocimiento y la maravilla de la creación. Mi proyecto de título, como un destello de luz en la inmensidad, representa una modesta contribución a la riqueza del entendimiento humano.

Gracias a mi familia, mi refugio constante, a Dios, mi guía inquebrantable, y al universo, fuente de inspiración inagotable, dedico este proyecto de título. Que estas palabras sirvan como un recordatorio de que, al igual que las estrellas, cada uno de nosotros desempeña un papel único en el tejido cósmico. No dudes en buscar el tuyo.

Con gratitud eterna y amor infinito,

Sebastián Aníbal Zapata Fuenzalida.

Resumen

Este proyecto de título se concentra en el análisis del tono emocional presente en las publicaciones de redes sociales de la Universidad de Playa Ancha, con el propósito de comprender las dinámicas sentimentales dentro de la comunidad participante en estas plataformas. Dada la falta de estudios científicos en este ámbito, buscamos caracterizar estas experiencias para enriquecer la comunidad con iniciativas, programas y planes estratégicos, generando así interacciones significativas con el entorno territorial en el que se encuentra la universidad.

La metodología empleada para la minería de datos fue KDD (Knowledge Discovery in Databases), una herramienta capaz de transformar datos simples pero abundantes en información valiosa que facilita la toma de decisiones. Con la ayuda del análisis de sentimientos, que evalúa el tono emocional de los comentarios mediante un diccionario léxico-emocional, medimos en gran medida qué temas son importantes para la comunidad. La recopilación de datos se llevó a cabo mediante técnicas de web scraping y el acceso a APIs de Facebook, Instagram y YouTube, utilizando herramientas como pandas, selenium y facebook-scraper.

En cuanto a los resultados, identificamos cinco temas principales en los títulos y comentarios de las publicaciones, destacando áreas como Artes y Reconocimientos, Historia, Memoria, Logros y Reconocimientos en los comentarios. Un análisis detallado reveló una dinámica emocional activa en la comunidad que interactúa con la universidad a través de las redes sociales.

La discusión resalta la existencia de un entorno positivo y emocionalmente conectado, evidenciado por la celebración de logros, el reconocimiento de personas destacadas y expresiones de gratitud. Estos resultados no solo ofrecen una comprensión de la dinámica emocional actual, sino que también establecen una base sólida para futuras estrategias de investigación y planificación en la Universidad de Playa Ancha.

Abstract

This research focuses on analyzing the emotional tone present in the social media posts of the University of Playa Ancha, aiming to understand the emotional dynamics within the participating community on these platforms. Given the lack of scientific studies in this field, we aim to characterize these experiences to enrich the community with initiatives, programs, and strategic plans, fostering meaningful interactions with the territorial environment where the university is situated.

The methodology employed for data mining was KDD (Knowledge Discovery in Databases), a tool capable of transforming simple yet abundant data into valuable information that facilitates decision-making. With the assistance of sentiment analysis, which evaluates the emotional tone of comments through a lexical-emotional dictionary, we measure to a large extent which topics are important to the community. Data collection was carried out using web scraping techniques and accessing APIs of Facebook, Instagram, and YouTube, employing tools such as pandas, selenium, and facebook-scraper.

Regarding the results, we identified five main themes in the titles and comments of the posts, highlighting areas such as Arts and Recognitions, History, Memory, Achievements, and Recognitions in the comments. A detailed analysis revealed an active emotional dynamic in the community interacting with the university through social media.

The discussion emphasizes the existence of a positive and emotionally connected environment, evidenced by the celebration of achievements, recognition of distinguished individuals, and expressions of gratitude. These findings not only provide an understanding of the current emotional dynamics but also establish a solid foundation for future research and planning strategies at the University of Playa Ancha.

Tabla de contenido

DEDICATORIA.....	2
Resumen.....	3
Abstract.....	4
Introducción.....	1
Capítulo 1: Antecedentes del estudio.....	3
1.1 Planteamiento del problema.....	3
1.2 Situación Universidad de Playa Ancha.....	5
1.3 Objetivo.....	7
1.3.1 Objetivos generales.....	7
1.3.2 Objetivos Específicos.....	7
1.4 Propuesta de Solución.....	7
Capítulo 2 Marco Teórico.....	9
2.1 Big Data.....	9
2.1.1.- Qué es Big Data.....	9
2.1.2.- ¿Cuántos datos se generan todos los días?.....	12
2.1.3.- Cómo funciona big data.....	17
2.1.4.- Tecnologías de big data.....	18
2.1.5.- Casos de Uso del Big Data.....	21
2.1.6 Tipos de Datos en Proyectos Big Data.....	22
2.2.- Procesamiento del Lenguaje Natural.....	26
2.2.1 Definición de NLP.....	26
2.2.2 Definición de términos claves.....	27
2.2.2.1 Polaridad en el análisis de sentimiento.....	28
2.2.2.2 Emociones y su identificación en el texto.....	29
2.2.3 Niveles del PLN.....	31
2.2.4.- Aplicaciones del NLP.....	33
2.3 Análisis de Sentimientos.....	37
2.3.1.- Qué es el análisis de sentimiento y cómo funciona.....	37
2.3.1.1 Tipos de análisis de sentimientos.....	38
2.3.1.2.- Aplicaciones del análisis de sentimientos.....	41
2.3.1.3.- Herramientas para el análisis de sentimientos.....	42
2.4 Minería de datos.....	46
2.4.1.- ¿Por qué minar datos?.....	46
2.4.2.- ¿Qué tecnologías se utilizan?.....	48
2.4.3.- Metodologías para la minería de datos.....	50
2.4.3.1.- Introducción a las Metodologías.....	50
2.4.3.2.- Fases de la metodología KDD.....	51
2.4.3.3.- Fases de la metodología CRISP-DM.....	54
2.4.3.4.- Fases de la metodología SEMMA.....	56
2.4.3.5.- Ventajas y limitaciones.....	57
2.4.3.6.- Consideraciones sobre las metodologías de minería de datos.....	59
2.4.4.- Fases de la minería de datos.....	60

2.4.4.1.- Selección de Datos.....	60
2.4.4.1.1.- Sistemas de recolección de información de las redes sociales....	60
2.4.4.2.- Preprocesamiento de Datos.....	65
2.4.4.2.1.-Sistemas para la limpieza de datos.....	67
2.4.4.3.- Minería de Datos.....	73
2.4.4.4.- Presentación de Resultados.....	78
2.4.4.4.1.-Visualización de datos.....	79
2.4.4.4.2.- Análisis Explicativo: Comprendiendo la Audiencia, Objetivos y Métodos.....	80
2.4.4.4.3.- Elegir una Imagen Efectiva.....	81
2.4.4.5.- Toma de Decisiones.....	87
Capítulo 3 Enfoque metodológico.....	88
3.1. Descripción de la Metodología KDD para el Análisis de Sentimientos.....	88
3.2.- Comprensión del contexto.....	89
3.3.- Creación de un Conjunto de Datos Objetivo.....	90
3.4.- Limpieza y Preprocesamiento de Datos.....	91
3.5.- Coincidencia de Objetivos con Método de Extracción de Datos.....	92
3.6.- Minería de Datos.....	93
3.7.- Planificación.....	95
Capítulo 4 Presentación de resultados.....	102
4.1.- Implementación.....	102
4.1.1 - Proceso Detallado del Entorno de Desarrollo.....	102
4.1.2.- Selección de plataformas de redes sociales para el análisis de sentimientos....	104
4.1.3.- Recopilación, preparación y obtención de los datos de las redes sociales relevantes para el análisis.....	105
4.1.4.- Datos de interés de las redes sociales.....	105
4.1.5.- Almacenamiento, limpieza y preprocesamiento de datos para eliminar ruido y redundancia.....	108
4.1.6.- Técnicas de procesamiento de texto y análisis de sentimiento para formar un conjunto de datos más robusto.....	108
4.2.-Desarrollo.....	111
4.2.1.- Programación de extracción de los comentarios de las redes sociales.....	111
4.2.2.- Estandarización y Normalización del Conjunto de Datos: Un Paso Esencial para el Análisis.....	134
4.2.3.- Desarrollo del análisis.....	136
4.2.1.- Visualización de los resultados del análisis.....	147
Capítulo 5 Conclusiones.....	169
Referencias Bibliográficas.....	175

Índice de Tablas

Tabla 1: Volumen de Datos Generados Diariamente a Nivel Mundial en 2023 .

Tabla 2: Tráfico de datos por categoría.

Tabla 3: Categorías de Datos en el Big Data según IBM.

Tabla 4: Clasificación de las aplicaciones del procesamiento de lenguaje natural.

Tabla 5: Fases de cada metodología de minería de datos.

Tabla 6: Tareas y Modelos de entrenamiento usados por pysentimiento.

Índice de Figuras

- Figura 1:** Explorando las Dimensiones del Proceso Big Data.
- Figura 2:** Datos generados y proyecciones futuras.
- Figura 3:** Otros medios de minería de datos.
- Figura 4:** Modelos de Análisis de Sentimientos: Supervisados y no Supervisados.
- Figura 5:** Búsqueda de conocimiento en los datos.
- Figura 6:** Proceso de la metodología KDD.
- Figura 7:** Labores que llevan más tiempo en un análisis big data.
- Figura 8:** Descifrando Patrones: De lo Básico a lo Avanzado en Minería de Datos.
- Figura 9:** Gráfico de texto.
- Figura 10:** Tabla.
- Figura 11:** Mapa de calor.
- Figura 12:** Gráfico de dispersión.
- Figura 13:** Gráfico de líneas.
- Figura 14:** Gráfico de pendiente.
- Figura 15:** Gráfico de barras.
- Figura 16:** Carta Gantt: Planificación inicial.
- Figura 17:** Carta Gantt: Preparación.
- Figura 18:** Carta Gantt: Desarrollo de marco teórico general.
- Figura 19:** Carta Gantt: Identificación de las fuentes de datos.
- Figura 20:** Carta Gantt: Selección del conjunto de datos.
- Figura 21:** Carta Gantt: Procesamiento.
- Figura 22:** Carta Gantt: Transformación.
- Figura 23:** Carta Gantt: Minería de datos.
- Figura 24:** Carta Gantt: Interpretación y conocimiento.
- Figura 25:** Carta Gantt: Entrega final.
- Figura 26:** Librerías necesarias para extraer datos.
- Figura 27:** Código de extracción de datos en Facebook.
- Figura 28:** Código de extracción de datos en Facebook uno.
- Figura 29:** Código de extracción de datos en Facebook dos.
- Figura 30:** Código de extracción de datos en Facebook tres.
- Figura 31:** Código de extracción de datos en Facebook cinco.
- Figura 32:** Código de extracción de datos en Facebook seis.
- Figura 33:** Código de extracción de datos en Instagram uno.
- Figura 34:** Código de extracción de datos en Instagram dos.
- Figura 35:** Código de extracción de datos en Instagram tres.
- Figura 36:** Código de extracción de datos en YouTube uno.
- Figura 37:** Código de extracción de datos en YouTube dos.
- Figura 38:** Unificación de los datos.
- Figura 39:** Unificación de los datos.
- Figura 40:** Cantidad de comentarios y no comentarios por red social.
- Figura 41:** Cantidad de comentarios y no comentarios por red social.

- Figura 42:** Código del proceso de tokenización.
- Figura 43:** Tópicos encontrados en las publicaciones.
- Figura 44:** Tópicos encontrados en los comentarios.
- Figura 45:** Código que predice el sentimiento en los comentarios.
- Figura 46:** Cantidad de publicaciones por red social.
- Figura 47:** Participación activa vs pasiva en general.
- Figura 48:** Cantidad de publicaciones con uno o más comentarios por red social.
- Figura 49:** Cantidad total de comentarios por red social.
- Figura 50:** Frecuencia de comentarios por publicación y fuente.
- Figura 51:** Distribución de tono emocional en comentarios.
- Figura 52:** Distribución de tono emocional en comentarios.
- Figura 53:** Distribución de tópicos de los títulos y su tono emocional.
- Figura 54:** Distribución de comentarios por tópicos y sentimientos.
- Figura 55:** Distribución de participación por tópicos.
- Figura 56:** Evolución mensual de comentarios por sentimiento.
- Figura 57:** Cantidad de comentarios y su sentimiento en relación a los días de la semana.
- Figura 58:** Cantidad de comentarios por tópicos en diferentes meses.

Introducción

El presente proyecto de título se sumerge en una exploración profunda del matiz emocional presente en las publicaciones de las redes sociales de la Universidad de Playa Ancha (UPLA). El objetivo primordial es desentrañar las complejidades emocionales que influyen en la comunidad participante en estas plataformas, con la finalidad de caracterizar y comprender las experiencias emocionales que dan forma a las interacciones digitales dentro de la comunidad que se involucra en las redes sociales de la universidad.

La motivación que impulsa este estudio se encuentra en la carencia de investigaciones específicas que aborden la minería de datos y el análisis emocional en el contexto de las redes sociales universitarias. Nuestra misión es enriquecer la comprensión de las interacciones emocionales y catalizar iniciativas destinadas a fomentar conexiones más significativas entre la universidad y su entorno territorial.

En última instancia, lo que se quiere lograr con este proyecto es obtener información relevante que pueda ser aplicada de manera práctica. Por ejemplo, buscamos desarrollar métricas que permitan medir de manera efectiva la interacción en las publicaciones de las redes sociales. Estas métricas no sólo enriquecerán nuestra comprensión de las dinámicas emocionales, sino que también proporcionarán herramientas tangibles para evaluar y mejorar la efectividad de las interacciones en el entorno digital universitario.

A continuación una breve descripción de los capítulos que conforman este proyecto de título:

Capítulo 1: Antecedentes del estudio

Este capítulo establece el contexto general del análisis emocional en las redes sociales de la Universidad de Playa Ancha. Se aborda el planteamiento del problema, la situación específica en la universidad y se establecen los objetivos generales y específicos, así como la propuesta de solución.

Capítulo 2: Estado del arte

Aquí se explora el panorama actual relacionado con Big Data, centrándose en su definición, funcionamiento y tecnologías asociadas. Además, se profundiza en el Procesamiento del Lenguaje Natural (NLP), el análisis de sentimientos y la minería de datos, proporcionando información sobre metodologías, tecnologías y herramientas utilizadas en estos campos.

Capítulo 3: Metodología

Se describe la metodología utilizada para realizar el análisis emocional en las publicaciones de las redes sociales. Se destaca el uso de la minería de datos y la técnica KDD, así como la planificación detallada del proceso, desde la comprensión del contexto hasta la minería de datos propiamente dicha.

Capítulo 4: Presentación de resultados

En este capítulo, se presenta la implementación práctica de la metodología descrita anteriormente. Se detalla el proceso de desarrollo, la selección de plataformas de redes sociales, la recopilación y preparación de datos, así como la aplicación de técnicas de procesamiento de texto y la visualización de resultados. Se enfatiza la interpretación de los resultados obtenidos.

Capítulo 5: Conclusiones

Finalmente, se recopilan y resumen los hallazgos más relevantes del estudio. Se discute la existencia de un entorno emocionalmente positivo en la comunidad universitaria, destacando áreas específicas de interés. Estos resultados no solo proporcionan una comprensión profunda de la dinámica emocional actual en la Universidad de Playa Ancha, sino que también sientan las bases para futuras investigaciones y estrategias.

Capítulo 1: Antecedentes del estudio

1.1.- Planteamiento del problema

En el entorno universitario actual, la comunicación efectiva y la participación de todas aquellas personas que de algún modo interactúan en las redes sociales de la universidad son elementos fundamentales para el desarrollo y el éxito de una institución educativa. Sin embargo, muchas universidades enfrentan desafíos en la optimización de sus canales de comunicación y en la promoción de la participación activa en las redes sociales. Además, la imagen de una universidad desempeña un papel crucial tanto en su imagen pública como en la percepción general de su calidad académica.

Asimismo, la falta de una estrategia efectiva para el uso de estas plataformas digitales, puede limitar el alcance e influencia de la universidad, así como obstaculizar la generación de un diálogo constructivo y el intercambio de ideas entre la comunidad universitaria y la sociedad en general. Por lo tanto, es crucial implementar estrategias de comunicación que fomenten la participación activa, la interacción significativa y la difusión de información relevante en las redes sociales.

En este contexto, surge la necesidad de evaluar la efectividad de los canales de comunicación y participación en las redes sociales de la universidad, tomando en cuenta los principios y objetivos que guían la misión de las universidades estatales establecidos en la ley 21094 artículo 4 del ministerio de educación (Biblioteca del Congreso Nacional de Chile [BCN], 2018, p.2). Estos principios y objetivos guían a las universidades públicas a desempeñar un papel fundamental en la sociedad, y es esencial que mantengan una buena comunicación con su entorno territorial. El artículo 4 de la ley 21094 establece la misión de las universidades del estado([BCN],2018, p.2), que va más allá de la generación y transmisión del conocimiento. Estas instituciones tienen la responsabilidad de contribuir al desarrollo cultural, social, territorial, artístico, científico, económico y sustentable del país, tanto a nivel nacional como regional, con una perspectiva intercultural.

Una comunicación efectiva con el entorno territorial permite a la universidad identificar y satisfacer las necesidades e intereses generales de la sociedad a la que sirven. Al establecer una vinculación preferente y pertinente con la región en la que tiene su domicilio o desarrollan sus actividades, las universidades públicas pueden adaptar sus programas académicos, investigación e innovación a las demandas específicas de ese entorno. Esto implica una estrecha colaboración con actores locales, para generar un impacto significativo en el desarrollo regional.

Es importante destacar que esta comunicación efectiva no solo implica transmitir información, sino también escuchar activamente a la comunidad y responder a sus

necesidades. Las actividades deben ser receptivas a las demandas y desafíos presentes en su entorno territorial, utilizando esta retroalimentación para mejorar y fortalecer su labor académica y contribuir de manera más efectiva al desarrollo de la sociedad.

La colaboración entre la universidad y los actores locales no solo beneficia a la comunidad, sino que también enriquece la formación de los estudiantes. Al estar en contacto directo con la realidad del entorno territorial, los estudiantes tienen la oportunidad de desarrollar habilidades prácticas y adquirir conocimientos relevantes para su futura inserción laboral. Asimismo, esta conexión con el entorno les permite comprender y abordar los desafíos específicos de la región, promoviendo un sentido de pertenencia y compromiso con la comunidad.

Para alcanzar este objetivo, se hace necesario implementar estrategias basadas en dos enfoques esenciales: el análisis de big data y el análisis de sentimientos. Estos métodos no solo permiten recopilar información de las redes sociales de manera exhaustiva, sino que también posibilitan su procesamiento y análisis en profundidad.

En este contexto, se optó por abordar una cantidad específica de comentarios para cada red social, estableciendo un total de 500 comentarios por plataforma. De esta manera, se logra obtener una muestra representativa de los comentarios generados en las publicaciones durante el segundo semestre del 2023.

El análisis de big data proporciona una visión enriquecedora al evaluar la efectividad de las redes sociales de la universidad. Al recopilar y analizar los datos generados en estas plataformas, como publicaciones, comentarios e interacciones, se obtiene una comprensión más profunda de la actividad en las redes sociales. Esto incluye identificar las tendencias emergentes, los temas de interés y las conversaciones relevantes entre la comunidad universitaria y la sociedad en general.

Por otro lado, el análisis de sentimientos desempeña un papel crucial en la evaluación de la efectividad de las redes sociales. Al aplicar técnicas de procesamiento del lenguaje natural, se puede analizar el tono emocional y las opiniones expresadas en los comentarios y publicaciones de las redes sociales. Esto permite comprender la percepción de los usuarios sobre la universidad, identificar si la respuesta emocional es positiva, negativa o neutra, y evaluar cómo esto impacta la imagen pública de la institución.

La combinación de estos enfoques, el análisis de big data y el análisis de sentimientos, no sólo identificará áreas de mejora en la comunicación institucional, sino que también facilitará la toma de decisiones informadas para implementar estrategias de comunicación más efectivas. En última instancia, este enfoque integral promoverá una participación activa y significativa en las redes sociales,

fortaleciendo la conexión entre la universidad y quienes interactúan en estas plataformas

Es por ello que, en el presente estudio, se busca evaluar la efectividad de los canales de comunicación y participación en las redes sociales de la universidad, y poder detectar posibles mejoras en dichos aspectos mediante la aplicación de enfoques basados en el análisis de big data y el análisis de sentimientos. Estas metodologías permiten recopilar, procesar, y analizar grandes volúmenes de datos generados en redes sociales, con el objetivo de identificar áreas de mejoras en la comunicación entre la institución y quienes interactúan en redes sociales.

Además, resulta imperativo examinar detenidamente los comentarios y opiniones expresadas en las redes sociales, con el propósito de evaluar y suministrar a la universidad información relevante que le permita alinearse con los principios establecidos en la ley, fortaleciendo así su conexión con el entorno. En consecuencia, el objetivo primordial de esta investigación es emplear el análisis de sentimientos como una herramienta integral para evaluar la eficacia de los canales de comunicación y la participación en las redes sociales de la Universidad de Playa Ancha (UPLA, s.f).

1.2.- Situación Universidad de Playa Ancha

La Universidad de Playa Ancha es una institución de educación superior fundada en 1948, la UPLA se ha consolidado como una universidad reconocida por su compromiso con la excelencia académica y su contribución al desarrollo regional y nacional. La UPLA ofrece una amplia variedad de programas académicos en diversas áreas del conocimiento, incluyendo humanidades, ciencias sociales, ciencias naturales, ingeniería, educación, arte y salud.

La UPLA ha reconocido el poder de las redes sociales y los medios de comunicación para fortalecer la comunicación institucional, promover la participación de la comunidad universitaria y proyectar su imagen pública. La universidad ha establecido una presencia activa en las principales redes sociales (Upla Comunica, s.f.), como Facebook, Twitter, Instagram, YouTube y LinkedIn, utilizando estas plataformas para mantener un diálogo abierto y cercano con estudiantes, académicos, personal administrativo y la comunidad en general.

A través de las redes sociales, la UPLA comparte información relevante sobre sus programas académicos, eventos, investigaciones y logros, fomentando así la participación activa y la colaboración. Estas plataformas permiten a la universidad llegar a un público amplio y diverso, y brindan a la comunidad universitaria la oportunidad de interactuar, expresar sus opiniones y compartir experiencias.

Reconociendo la relevancia de la comunicación digital, la UPLA implementa planes anuales específicos para eventos clave, como el proceso de admisión, eventos y otras fechas específicas. Este enfoque estratégico se dirige principalmente a estudiantes, seguido por docentes y la comunidad UPLA en general. Además, se busca impactar positivamente a la comunidad externa, subrayando la voluntad de la UPLA de ser una institución comprometida con la sociedad.

Por otra parte, es importante resaltar la importancia de elementos específicos para un análisis integral de las interacciones en línea. Datos cuantitativos, métricas de rendimiento, comentarios y reacciones provenientes de las redes sociales son esenciales para optimizar las publicaciones y obtener una comprensión profunda de las interacciones en línea.

Sin embargo, la gestión efectiva de la reputación en línea presenta desafíos. La universidad debe abordar críticas y comentarios negativos de manera estratégica para proteger y fortalecer su imagen pública. La diversificación de contenido también es crucial, equilibrando la promoción de programas académicos con material atractivo y relevante para diversos segmentos de la comunidad.

El volumen masivo de datos de redes sociales es otro desafío, requiriendo herramientas y estrategias eficientes para filtrar y priorizar la información relevante. Además, la interpretación precisa del contexto, especialmente frente a expresiones culturales específicas o sarcasmo, es esencial para evitar malentendidos en el análisis de sentimientos.

La identificación de tendencias emergentes exige un análisis continuo y adaptativo para responder proactivamente a cambios significativos en la opinión pública. Manejar comentarios negativos, desarrollando protocolos para gestionar críticas constructivas y situaciones de crisis, es crítico para la protección de la reputación en línea.

Convertir los hallazgos del análisis de sentimientos en estrategias concretas también plantea desafíos. La UPLA debe desarrollar un puente efectivo entre los resultados del análisis y la implementación de cambios o mejoras. Este enfoque integral permitirá a la universidad no solo evaluar su efectividad en las redes sociales, sino también optimizar sus estrategias de comunicación y participación en línea.

1.3.- Objetivo

1.3.1.- Objetivos generales

Evaluar la efectividad de los canales de comunicación y participación en las redes sociales de la Universidad de Playa Ancha para determinar su impacto en la difusión de información, interacción con la comunidad universitaria y proyección de la imagen institucional.

1.3.2.- Objetivos Específicos

1.- Revisar y determinar las métricas que permiten medir la percepción de las comunicaciones en redes sociales.

2.- Analizar el alcance y la visibilidad de las publicaciones y contenidos compartidos por la UPLA en sus redes sociales, mediante el seguimiento de las métricas identificadas.

3.- Realizar un análisis de sentimientos en los comentarios y menciones relacionados con la UPLA en las redes sociales para medir la percepción general de la universidad. evaluando el tono emocional de los comentarios.

1.4.- Propuesta de Solución

La estrategia propuesta para optimizar los canales de comunicación y participación en las redes sociales de la Universidad de Playa Ancha se fundamenta en enfoques avanzados como el análisis de big data y el análisis de sentimientos. Este enfoque integral tiene como objetivo evaluar de manera exhaustiva la efectividad de las interacciones en línea, considerando diversos aspectos cruciales.

Se llevará a cabo un seguimiento meticuloso de métricas clave, incluyendo la cantidad de seguidores, interacciones como: comentarios, compartir, me gustas y la viralidad de las publicaciones. Este análisis proporcionará una evaluación detallada del nivel de participación y la difusión de la información compartida por la UPLA en sus redes sociales. Se identificará qué publicaciones generan mayor cantidad de comentarios, destacando aquellas que generan un mayor impacto en la comunidad. Este enfoque permitirá comprender qué contenidos son más atractivos y cómo pueden servir de modelo para futuras estrategias.

Mediante técnicas de procesamiento del lenguaje natural, se llevará a cabo un análisis de sentimientos en los comentarios y menciones relacionados con la UPLA en las redes sociales. Esta evaluación permitirá medir el tono emocional de las interacciones, identificando si existe una percepción pública positiva, negativa o neutral hacia la institución y sus actividades.

Se implementará un análisis detallado de las palabras más repetidas en los comentarios, identificando términos clave y temas recurrentes. Esta información proporcionará perspectivas valiosas para ajustar la estrategia de contenido y alinearse con los intereses de la comunidad. Se analizará el momento óptimo para publicar contenido, evaluando las horas en que la audiencia tiende a interactuar más en las redes sociales. Este enfoque temporal estratégico garantizará una mayor visibilidad y participación.

El análisis incluirá la gestión efectiva de comentarios negativos, estableciendo protocolos para abordar críticas constructivas y gestionar situaciones de crisis, fortaleciendo así la reputación en línea de la UPLA. Se presentarán visualizaciones gráficas para ilustrar de manera clara y concisa los datos clave. Estas representaciones visuales facilitarán la interpretación de los resultados y su integración en estrategias concretas. Se destaca que este análisis no es un evento aislado, sino un proceso continuo de mejora. La adaptación constante según las tendencias y las necesidades cambiantes de la comunidad garantizará la eficacia a largo plazo.

Con esta propuesta, se busca obtener una visión completa y detallada de la efectividad de los canales de comunicación, proporcionando a la UPLA herramientas precisas para mejorar la conexión con la comunidad universitaria y el entorno territorial.

Capítulo 2: Marco Teórico

2.1.- Big Data

En el contexto actual, el uso potente de tecnologías ha orquestado un crecimiento exponencial en la generación de datos. Estos datos, emergiendo a una velocidad vertiginosa y adoptando formatos diversos, plantean la urgente necesidad de extraer patrones, tendencias y conocimientos de manera eficiente. El propósito central radica en obtener información valiosa, esencial para la toma de decisiones en diferentes escenarios. En respuesta a esta urgencia, los métodos tradicionales han experimentado una evolución sustancial.

La sobrecarga informativa que caracteriza el entorno digital actual no solo se traduce en un aumento cuantitativo de datos, sino también en una velocidad de generación y diversificación de formatos sin precedentes. Este escenario impulsa la urgencia de adaptar enfoques y metodologías para aprovechar efectivamente el potencial inherente a estos datos.

Surge entonces el "Análisis Big Data" como respuesta a la complejidad y volumen de los datos contemporáneos. Este enfoque se centra en la extracción rápida de información valiosa, confiriendo a las organizaciones una capacidad mejorada para la toma de decisiones. Estas habilidades, a su vez, generan ventajas competitivas en una variedad de sectores.

Esta sección inicial busca proporcionar una visión general de las implicaciones de "Big Data", especialmente enfocada en el ámbito del análisis y en el beneficio que se puede obtener a través del uso de estas metodologías. Esta introducción establecerá los fundamentos para un análisis más detallado de oportunidades, retos y tendencias en las secciones subsiguientes.

A medida que nos sumergimos en el desarrollo de este tema, se explorará con mayor profundidad varios aspectos del análisis de grandes conjuntos de datos. Nos adentraremos en su historia, funcionamiento, tecnologías asociadas, casos de uso y tipos de datos relevantes en proyectos de "Big Data".

2.1.1.- Qué es Big Data

Bajo el artículo llamado "The Evolution of Big Data and the Future of the Data Platform"(Oracle, 2022), explica exactamente lo que es el Big data y que para ellos "El término "big data" abarca datos que contienen una mayor variedad y que se presentan en volúmenes crecientes y a una velocidad superior. Esto también se conoce como "las tres V"."

Dicho de otro modo, el big data está formado por conjuntos de datos de mayor tamaño y más complejos, especialmente procedentes de nuevas fuentes de datos. Estos conjuntos de datos son tan voluminosos que el software de procesamiento de datos convencional sencillamente no puede gestionarlos. Sin embargo, estos volúmenes masivos de datos pueden utilizarse para abordar diversos problemas que antes no hubiera sido posible solucionar.

Para comprender mejor la complejidad de big data, es esencial explorar los conceptos iniciales claves conocidos como "las tres V". Por su parte (Oracle, s.f.) define las "tres V" del big data como:

Volumen: La cantidad de datos importa. Con big data, tendrá que procesar grandes volúmenes de datos no estructurados de baja densidad. Puede tratarse de datos de valor desconocido, como feeds de datos de Twitter, flujos de clics de una página web o aplicación para móviles, o equipo con sensores.

Velocidad: La velocidad es el ritmo al que se reciben los datos y (posiblemente) al que se aplica alguna acción. La mayor velocidad de los datos normalmente se transmite directamente a la memoria, en vez de escribirse en un disco. Algunos productos inteligentes habilitados para Internet funcionan en tiempo real o prácticamente en tiempo real y requieren una evaluación y actuación en tiempo real.

Variedad: La variedad hace referencia a los diversos tipos de datos disponibles. Los datos convencionales eran estructurados y podían organizarse claramente en una base de datos relacional. Con el auge del big data, los datos se presentan en nuevos tipos de datos no estructurados. Los tipos de datos no estructurados y semiestructurados, como el texto, audio o video, requieren un preprocesamiento adicional para poder obtener significado y habilitar los metadatos.

Estos son fundamentales para comprender los desafíos y oportunidades asociados con el análisis de grandes volúmenes de información. Además, en los últimos años han surgido nuevas "V" que también son relevantes en el contexto del análisis de big data. A continuación, se presentan algunas de estas "V" adicionales.

Bajo el artículo "Las 10v's del big data"(datahack, 2020), define las siguientes "V" como:

Variabilidad: La variabilidad se refiere tanto a las inconsistencias en los datos como a las múltiples dimensiones de los mismos. En términos de inconsistencias, se busca identificar y abordar los valores atípicos o anómalos que puedan afectar los resultados del análisis. En cuanto a las dimensiones de los datos, es importante considerar que los datos provienen de diferentes fuentes y tipos, lo que puede

generar desafíos adicionales al integrar y analizar estos conjuntos de datos dispares.

Veracidad: La veracidad se relaciona con la procedencia y confiabilidad de los datos. Es esencial evaluar la calidad y la confiabilidad de la fuente de datos, así como su contexto, para garantizar que el análisis se base en información precisa y significativa. A medida que aumenta la complejidad y variedad de los datos, es importante tener en cuenta la veracidad de los mismos.

Validez: La validez se refiere a la calidad y precisión de los datos para su uso en el análisis. Es fundamental asegurarse de que los datos sean limpios, precisos y estén libres de errores o inconsistencias. Adoptar buenas prácticas de gobernanza de datos ayuda a garantizar una calidad de datos coherente, definiciones comunes y metadatos adecuados.

Vulnerabilidad: La vulnerabilidad se refiere a la seguridad de los datos. Con el aumento de la cantidad y la sensibilidad de los datos almacenados y procesados, es necesario tener en cuenta la seguridad y protección de la información para prevenir accesos no autorizados, hackers y actividades ilegales.

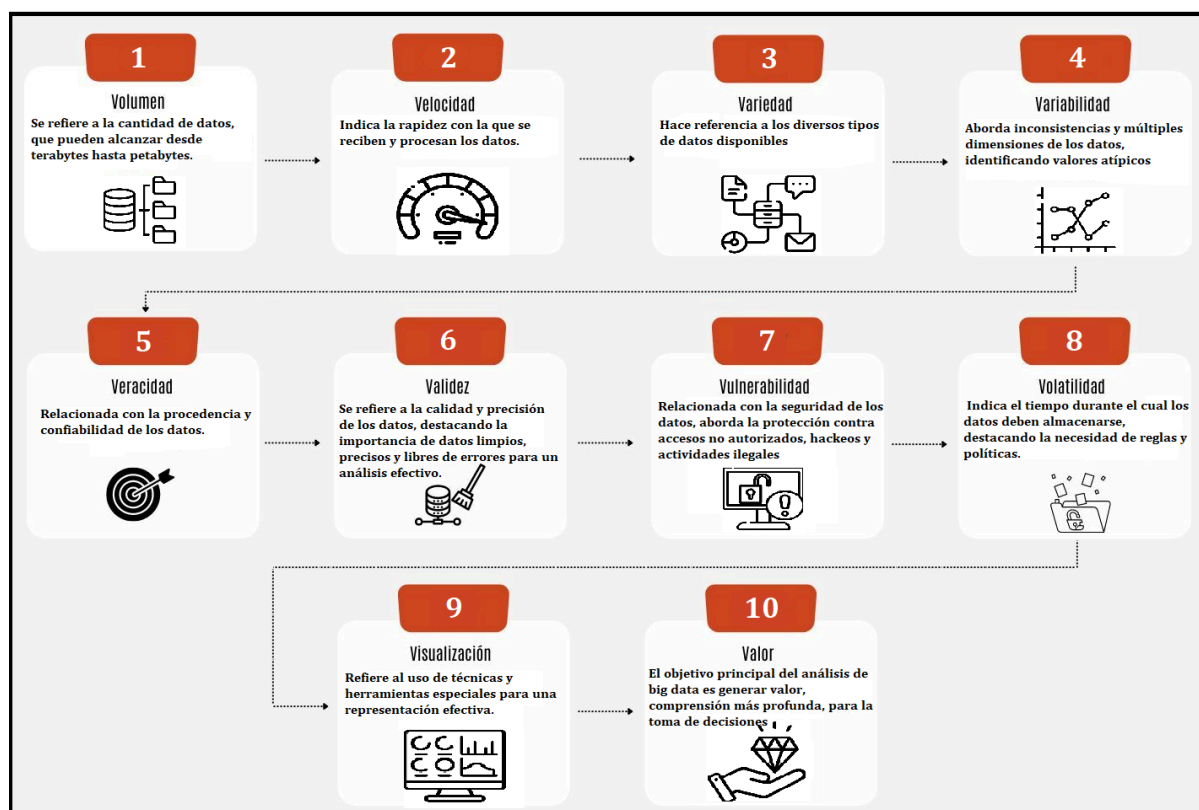
Volatilidad: La volatilidad se refiere al tiempo durante el cual los datos deben mantenerse o almacenarse. Antes del auge del big data, se solía guardar datos indefinidamente debido a los bajos costos de almacenamiento. Sin embargo, en la actualidad, es necesario establecer reglas y políticas para determinar la disponibilidad y la vigencia de los datos, así como garantizar una recuperación rápida de la información cuando sea necesario.

Visualización: La visualización se refiere a la complejidad de representar y visualizar grandes volúmenes de datos. Los métodos tradicionales de visualización pueden no ser adecuados para grandes conjuntos de datos, por lo que es necesario utilizar técnicas y herramientas especiales para representar de manera efectiva la información, como la agrupación, el uso de mapas, las coordenadas y los diagramas.

Valor: Finalmente, el valor es el resultado clave que se busca obtener del análisis de big data. Todas las características anteriores son importantes, pero en última instancia, el análisis de big data debe generar valor, como una comprensión más profunda de los clientes, la optimización de procesos, la mejora del rendimiento o la toma de decisiones más informadas. La representación gráfica ilustra el proceso mediante el cual el Big Data puede transformar grandes volúmenes de datos en información valiosa.

Figura 1.

Explorando las Dimensiones del Proceso Big Data



Nota. Elaboración propia.

Estas "V" adicionales amplían la comprensión de los desafíos y oportunidades asociados con el análisis de big data, y proporcionan un marco más completo para abordar la complejidad de los datos en la actualidad. Al considerar estas diversas dimensiones, las organizaciones pueden obtener una visión más clara y aprovechar al máximo el potencial de sus datos.

2.1.2.- ¿Cuántos datos se generan todos los días?

Según las últimas estimaciones realizadas en explodingtopics (Duarte, 2023), se generan asombrosos 328,77 millones de terabytes de datos cada día. Esta cifra abarca tanto datos recién generados como aquellos capturados, copiados o consumidos.

Para comprender mejor la magnitud de esta cantidad, veamos cómo se traduce en zettabytes: 120 zettabytes por año, 10 zettabytes por mes, 2,31 zettabytes por semana o 0,33 zettabytes por día. A continuación, se presenta una perspectiva de la cantidad de datos generados diariamente en diferentes unidades de medida Tabla 1.

Tabla 1*Datos Generados Diariamente a Nivel Mundial en 2023*

Unidad de medida	Datos Generados
Zettabytes	0.33
exabytes	328.77
petabytes	328.767,12
Terabytes	328,77 millones
Gigabytes	328.77 mil millones
Megabytes	328,77 billones
kilobytes	328,77 cuatrillones
bytes	328.77 quintillones

Nota: (Statista, 2023)

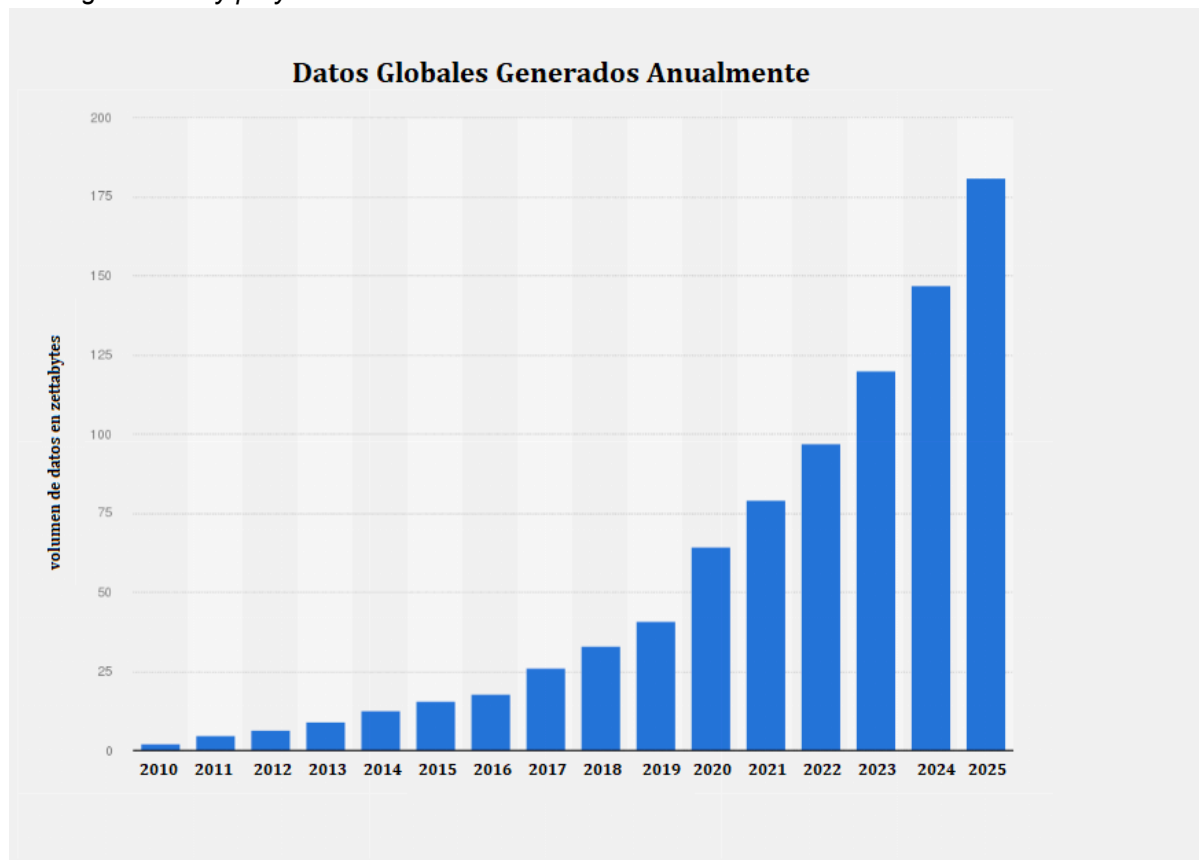
El volumen de datos generados anualmente ha experimentado un crecimiento exponencial desde 2010, marcando un hito significativo en el panorama de la información. De hecho, se estima que el 90% de los datos existentes en el mundo han sido generados en los últimos dos años, lo cual evidencia la aceleración de este fenómeno.

En tan solo 13 años, el volumen de datos ha experimentado un impresionante aumento, pasando de 2 zettabytes en 2010 a los actuales 120 zettabytes generados en 2023. Esto representa un incremento aproximado de 60 veces en este periodo. Las proyecciones futuras indican que este crecimiento no muestra signos de desaceleración. Se espera que el volumen de datos generados en 2025 alcance los 181 zettabytes, lo que representa un aumento de más del 150% en comparación con 2023.

Estas cifras que se pueden observar en la figura 2, destacan la magnitud del impacto que los datos tienen en nuestra sociedad y resaltan la importancia de la gestión y el análisis eficientes de esta vasta cantidad de información para impulsar la innovación y la toma de decisiones informadas en diversos sectores.

Figura 2

Datos generados y proyecciones futuras.



Nota. Creación De Datos Por Categoría.(Bernard Marr & Co, 2023).

De acuerdo con un informe de Sandvine (Lyn Cantor, 2022, p.13), el video es responsable de más de la mitad de todo el tráfico de datos a nivel mundial, representando un impresionante 53,72%. Es evidente que las redes sociales desempeñan un papel destacado en la difusión de contenido de vídeo. TikTok, una plataforma que se centra exclusivamente en videos, continúa expandiendo su base de usuarios de manera significativa año tras año. Además, Facebook ha experimentado una evolución notable, llegando al punto en que el 51% del contenido compartido en esta red social se basa en videos. Estos datos resaltan la creciente importancia y popularidad del formato de video en línea.

Al analizar el panorama global del tráfico de datos, podemos observar que, junto con las redes sociales (12,69%) y los juegos (9,86%), estas tres categorías en conjunto representan más de tres cuartas partes (76,27%) del total de tráfico de datos en Internet. Esto demuestra la influencia significativa que tienen el video, las redes sociales y los juegos en la forma en que consumimos y compartimos información en la era digital.

Tabla 2*Tráfico de datos por categoría*

Categoría	Proporción del tráfico de datos de internet
Video	53.72%
Redes sociales	12.69%
Web	9.86%
Juego de azar	5.67%
Mensajería	5.35%
Mercado	4.54%
Compartición de archivos	3.74%
Nube	2.73%
Red virtual privada(vpn)	1.39%
Audio	0.31%

Nota. (Lyn Cantor, 2022, p.13).

¿Cuántos datos se generan diariamente en las redes sociales? Las estadísticas de uso revelan que la cantidad es asombrosa. A continuación, se destacarán algunos datos relevantes sobre las plataformas de redes sociales.

El 51% del contenido de Facebook es vídeo. En el panorama actual, el contenido de video es el protagonista. No es de extrañar que más del 50% del contenido en Facebook sea video. Esta forma de contenido atrae la atención de los usuarios y les permite interactuar a través de comentarios, likes y comparticiones.

Más de 650 millones de tuits nuevos se publicaron todos los días en 2022. Es difícil calcular la cantidad exacta de tweets enviados diariamente, pero las estadísticas recientes de uso de las redes sociales indican que se publican más de 650 millones de tuits al día. Cada uno de estos tuits tiene el potencial de generar comentarios, recibir likes y volverse viral, lo que amplifica su alcance.

A partir del primer trimestre de 2022, el 79% de los usuarios de Twitter siguen marcas. Los usuarios de Twitter están interesados en mantenerse conectados con

las marcas y recibir las últimas actualizaciones. Muchas marcas crean contenido atractivo e informativo que los seguidores disfrutan ver y comentar. Los comentarios y likes en los tweets ayudan a medir la interacción de los usuarios con las marcas.

El 90% de los usuarios de Instagram siguieron una marca en 2022. Es interesante destacar que el 70% de los compradores revisan las marcas en Instagram antes de realizar una compra. Aquí, los comentarios y likes en las publicaciones son indicadores clave de la aceptación y el interés de los usuarios en el contenido de las marcas.

Estas estadísticas subrayan la importancia de aprovechar los comentarios, likes y la viralidad del contenido en el ámbito de las redes sociales. Estas interacciones no solo brindan a los organismos y a los usuarios una oportunidad de conectarse, expresar opiniones y compartir contenido relevante, sino que también pueden desempeñar un papel fundamental para que una organización pueda establecer una conexión significativa con su entorno y difundir información valiosa para la comunidad.

En un mundo cada vez más interconectado, las redes sociales se han convertido en un canal poderoso para establecer una comunicación bidireccional con el público objetivo. Los comentarios permiten a las organizaciones y a los usuarios entablar conversaciones en tiempo real, responder preguntas, agradecer el apoyo y recibir retroalimentación directa. Los likes, por su parte, indican la aceptación y el aprecio hacia el contenido compartido, generando un sentido de validación y reconocimiento.

La viralidad del contenido en las redes sociales puede tener un impacto significativo en la difusión de información valiosa para la comunidad. Cuando el contenido se comparte masivamente y se vuelve viral, alcanza a un público más amplio y puede generar un impacto duradero. Esto brinda a las organizaciones la oportunidad de compartir mensajes importantes, promover iniciativas sociales, crear conciencia sobre temas relevantes y fomentar la participación de la comunidad en causas significativas.

Es fundamental comprender que un correcto uso de las redes sociales no solo implica generar interacciones superficiales, sino también aprovechar su potencial para construir conexiones auténticas y significativas. Al hacerlo, las organizaciones pueden establecerse como fuentes confiables de información, promover el diálogo constructivo y brindar un valor agregado a la comunidad. En última instancia, el uso efectivo de las redes sociales puede fortalecer la relación entre una organización y su entorno, permitiendo una difusión más efectiva de información valiosa para el beneficio de todos.

2.1.3.- Cómo funciona big data

El análisis de big data es un proceso complejo que abarca varios pasos esenciales para transformar grandes volúmenes de datos en conocimientos prácticos y aprovechables. A continuación, se ofrece una breve introducción a cada uno de los procesos implicados (Tableau, sf):

Recopilación de datos: La recopilación de datos es el primer paso del análisis de big data. Las organizaciones pueden aprovechar la tecnología actual para recopilar datos de diversas fuentes, como almacenamiento en la nube, aplicaciones móviles, sensores IoT, redes sociales, entre otros. Estos datos pueden ser estructurados, por ejemplo, bases de datos, o no estructurados como texto, imágenes o videos. Algunos datos se almacenan para facilitar el acceso, mientras que otros datos sin procesar o complejos se pueden almacenar en un lago de datos asignándoles metadatos.

Procesamiento de datos: Una vez que los datos se recopilan, es necesario procesarlos para obtener resultados precisos en las consultas analíticas. Dado que la cantidad de datos disponibles está creciendo rápidamente, el procesamiento de datos se vuelve desafiante. Hay dos enfoques comunes para el procesamiento de datos: el procesamiento por lotes y el procesamiento en tiempo real (o de flujo). El procesamiento por lotes analiza grandes bloques de datos en un período de tiempo, mientras que el procesamiento en tiempo real analiza pequeños lotes de datos de manera continua y rápida. Estos enfoques permiten realizar un análisis más eficiente y tomar decisiones más rápidas.

Limpieza de datos: La limpieza de datos es un proceso esencial para garantizar la calidad de los datos utilizados en el análisis. Esto implica eliminar datos duplicados, corregir errores y garantizar que los datos estén en el formato correcto. La limpieza de datos ayuda a evitar errores y distorsiones en los resultados del análisis, asegurando que los conocimientos extraídos sean precisos y confiables.

Análisis de datos: Una vez que los datos están limpios y preparados, se pueden aplicar técnicas de análisis avanzadas para extraer conocimientos valiosos. Algunos métodos comunes de análisis de big data incluyen la minería de datos, el análisis predictivo y el aprendizaje profundo. La minería de datos busca patrones, relaciones y anomalías en grandes conjuntos de datos. El análisis predictivo utiliza datos históricos para realizar predicciones sobre eventos futuros. Por último, el aprendizaje profundo utiliza técnicas de inteligencia artificial y aprendizaje automático para identificar patrones complejos y abstractos en los datos.

2.1.4.- Tecnologías de big data

Tal como mencionamos anteriormente, el Big Data no consiste netamente en almacenar y almacenar grandes cantidades de datos, sino que se trata de procesarlos con el fin de generar información de utilidad. Es allí donde se encuentra la importancia de las herramientas de Big Data, las cuales más que ser importantes, son esenciales pues complementan el proceso.

De esta manera, las distintas herramientas son capaces de reducir notablemente el coste de almacenamiento, lo cual resulta ideal cuando hablamos de rendimiento e incremento de la eficiencia. En este sentido, gracias a ellas el análisis de los datos es aún más sencillo, por lo que también reduce la inversión de tiempo y esfuerzo, pues estaríamos realizando las mismas tareas, pero con una mayor rapidez.

A continuación, se detallarán algunas herramientas de Big Data más utilizadas y se especificará la importancia de cada una de ellas.

Herramientas de Procesamiento y Almacenamiento:

- **Apache Hadoop:** Hadoop es un framework que permite el procesamiento de un gran volumen de datos en lote, utilizando para ello modelos de programación simple.
- **Elasticsearch:** Elasticsearch permite el procesamiento de grandes cantidades de datos y ver la evolución de éstos en tiempo real.
- **Apache Storm:** Consiste en una herramienta Big Data con un sistema de computación distribuida en tiempo real que se encuentra orientado a procesar constantes flujos de datos.
- **MongoDB:** Es una base de datos NoSQL que fue optimizada con el objetivo de trabajar con grandes grupos de datos que tienden a variar con cierta frecuencia, o que sencillamente son semiestructurados o inestructurados.
- **Apache Spark:** Consiste en un motor de procesamiento de datos de código abierto, siendo la velocidad de esta herramienta su aspecto más atractivo; Spark cumple la función de analizar datos divididos en lotes y en tiempo real.

Lenguajes de Programación

- **Python:** Python siempre será una opción de lo más adecuada, esto se debe a que, además de ser un lenguaje en desarrollo constante, su código simple y

sus inmensas bibliotecas hacen que sea la opción preferida para la mayoría de los programadores.

- **Lenguaje R:** Es un retorno y lenguaje de programación que está enfocado principalmente en el análisis estadístico, puesto que resulta sumamente similar al lenguaje matemático; aunque también es utilizado con mucha frecuencia para el análisis de Big Data.

Bases de Datos y Almacenamiento Distribuido

- **Apache Cassandra:** Ofrece alta disponibilidad y escalabilidad sin necesidad de comprometer el rendimiento, se trata de una base de datos NoSQL.
- **Apache Drill:** Es un framework open-source que ha sido diseñado con el fin de alcanzar una escalabilidad alta en servidores, así como para poder procesar petabytes de datos y millones de registros en muy poco tiempo.
- **Apache Oozie:** Consiste en un sistema de flujo de trabajo que te permitirá definir un amplio rango de trabajos programados o escritos, y en distintos lenguajes.

Otras Herramientas Significativas:

- **MapReduce:** Esta herramienta Big Data está dividida en dos términos clave: “Mapa” y “Reducir”; el primero hace referencia a la conversión de un dataset en otro conjunto de datos(es *decir: clave/valor*). Por otro lado, el “reducir” se refiere a extraer los datos almacenados del “mapa” para combinar los diferentes pares clave/valor, convirtiéndolos en pares más pequeños.
- **Gridgain:** Esta es una herramienta importante que podrá proporcionarte métodos dinámicos de almacenamiento de datos y computación, como, por ejemplo, la agrupación dinámica, el balanceo de carga, la mensajería distribuida, el procesamiento MapReduce, la ejecución de cierre distribuido, la tolerancia a fallos e incluso escalabilidad lineal.
- **HPCC:** Según sus siglas: “*Clustering Computacional de Alto Rendimiento*”, ha ganado gran reconocimiento debido a que es capaz de ofrecer un rendimiento superior a Hadoop en entornos específicos.
- **Apache Hive:** Se trata de una infraestructura datawarehouse para Hadoop, que a su vez facilita la gestión y consulta de grandes volúmenes de datos que se encuentran en almacenamiento distribuido.

- **Apache Kafka:** Consiste en un sistema de publicación distribuida que fue especialmente diseñado con el fin de permitir que sólo un grupo de datos representa la columna vertebral en una gran organización. En este sentido, con un único broker de Kafka es posible controlar cientos de megabytes en escrituras y lecturas por cada segundo desde miles de clientes.
- **Apache Flume:** Este consiste en un servicio fiable y distribuido sumamente útil para agregar, recoger y trasladar grandes conjuntos de datos de la forma más eficiente posible.
- **Hcatalog:** Esta es una gran herramienta para análisis de Big Data, pues consiste en un servicio centralizado de administración y uso compartido de metadatos destinados a Apache Hadoop.
- **Impala Apache:** Es una herramienta Big Data que te proporcionará una plataforma unificada y familiar para realizar diversas consultas que se orienten a lotes o en tiempo real. De esta manera, te brindará la posibilidad de hacer consultas SQL interactivas y veloces de forma directa en los datos de Apache Hadoop.
- **JSON:** Según sus siglas en inglés: «JavaScript Object Notation». Se trata de una herramienta Big Data altamente popular en todas aquellas soluciones que manejan REST-API para administración mediante interfaz navegador. Es por ello que muchas de las bases de datos NoSQL en la actualidad prefieren almacenar datos en el formato JSON, ya que ha sido incluido entre los más fiables para los desarrolladores web.
- **Neo4j:** Consiste en una base de datos de gráficos, que a su vez cuenta con amplias mejoras de rendimiento de hasta 1000 veces o más en comparación con bases de datos relacionales.
- **Pig:** Este es un lenguaje que se basa en Hadoop y que fue desarrollado por Yahoo. Su mayor ventaja es lo sencillo que resulta aprender a manejarlo, siendo a su vez un lenguaje experto para los canales de datos demasiado extensos y profundos.
- **Tableau:** En este caso nos encontramos ante una herramienta de visualización de datos que se enfoca principalmente en la inteligencia empresarial, brindando así una gran ventaja a las diferentes compañías. En este sentido, con Tableau será posible crear gráficos de barras, diagramas de dispersión, mapas, etc; y lo más interesante es que para todo esto no tendrás necesidad de programar.

2.1.5.- Casos de Uso del Big Data

El procesamiento de Big Data tiene aplicaciones amplias y significativas en diversas áreas del conocimiento, abarcando disciplinas que van desde la literatura, la salud y la psicología hasta la biología, la matemática y la ingeniería. Para comprender mejor la utilidad y la diversidad de estas aplicaciones, se puede recurrir al artículo titulado "17 Use Cases for Graph Databases and Graph Analytics" (Oracle, 2021). Este recurso proporciona una perspectiva detallada que permitirá una definición más precisa y completa del impacto y las posibilidades que ofrece el Big Data en distintos campos.

Fabricación: La revolución digital ha transformado la industria manufacturera. Los fabricantes ahora están encontrando nuevas formas de aprovechar todos los datos que generan para mejorar la eficiencia operativa, optimizar los procesos comerciales y descubrir información valiosa que impulsará las ganancias y el crecimiento.

Retail: La competencia es feroz en el comercio minorista. Para mantenerse a la vanguardia, las empresas se esfuerzan por diferenciarse. Los grandes datos se utilizan en todas las etapas del proceso minorista, desde predicciones de productos hasta pronósticos de demanda y optimización en la tienda. Utilizando big data, los minoristas están encontrando nuevas formas de innovar.

Cuidado de la salud: Las organizaciones de atención médica están utilizando big data para todo, desde mejorar la rentabilidad hasta ayudar a salvar vidas. Las empresas de atención médica, los hospitales y los investigadores recopilan cantidades masivas de datos. Pero todos estos datos no son útiles de forma aislada. Se vuelve importante cuando se analizan los datos para resaltar tendencias y amenazas en patrones y crear modelos predictivos.

Telecomunicaciones: La popularidad de los teléfonos inteligentes y otros dispositivos móviles ha brindado a las empresas de telecomunicaciones enormes oportunidades de crecimiento. Pero también hay desafíos, ya que las organizaciones trabajan para seguir el ritmo de las demandas de los clientes de nuevos servicios digitales mientras administran un volumen de datos en constante expansión.

Servicios financieros: Los bancos y las empresas de servicios financieros con visión de futuro están capitalizando el big data. Desde la captura de nuevas oportunidades de mercado hasta la reducción del fraude, las organizaciones de servicios financieros han podido convertir el big data en una ventaja competitiva.

Redes sociales: Las plataformas de redes sociales generan grandes volúmenes de datos, como publicaciones, comentarios, likes y compartidos. El big data se utiliza para analizar estos datos y extraer información sobre patrones de comportamiento de los usuarios, preferencias, tendencias y opiniones, lo que permite a las empresas de redes sociales mejorar sus servicios, personalizar recomendaciones y segmentar audiencias para campañas publicitarias más efectivas.

Marketing digital: En el ámbito del marketing digital, el big data se utiliza para recopilar y analizar datos de comportamiento del usuario, preferencias de compra, interacciones en línea y datos demográficos. Esto permite a las empresas personalizar sus estrategias de marketing, ofrecer mensajes y ofertas dirigidas a segmentos específicos de audiencia y realizar campañas más efectivas.

Transporte y logística: Las empresas de transporte y logística utilizan big data para optimizar rutas de entrega, mejorar la eficiencia en el seguimiento de envíos, predecir la demanda y gestionar el inventario de manera más efectiva. Además, el análisis de datos en tiempo real puede ayudar a prevenir problemas y mejorar la seguridad en el transporte.

Energía: En el sector energético, el big data se utiliza para analizar datos de medidores inteligentes, sensores y dispositivos de generación de energía. Esto permite a las empresas identificar patrones de consumo, mejorar la eficiencia energética, predecir la demanda y optimizar la producción y distribución de energía.

2.1.6.- Tipos de Datos en Proyectos Big Data

En la era digital actual, como ya se mencionó anteriormente la cantidad de datos generados a diario es asombrosa. Desde las interacciones en redes sociales hasta los registros de transacciones comerciales, los datos fluyen constantemente y a un ritmo vertiginoso. Esta avalancha de información, conocida como "Big Data", ha transformado la forma en que las organizaciones abordan la toma de decisiones, la identificación de patrones y la obtención de conocimientos valiosos.

Sin embargo, no todos los datos son iguales. En el vasto mundo del Big Data, es fundamental comprender los diferentes tipos de datos con los que nos encontramos. Desde los datos estructurados, que siguen un formato predefinido, hasta los datos no estructurados, que son más desafiantes de analizar debido a su falta de estructura uniforme, cada tipo de dato presenta sus propias características y desafíos.

En esta sección, se profundizará los tipos de datos que se encuentran en proyectos de Big Data. Abordaremos no solo los datos estructurados y no estructurados, sino también una categoría intermedia conocida como datos semiestructurados. A través

de esta comprensión, estaremos preparados para abordar la diversidad de datos que enfrentaremos en la era del Big Data y aprovechar al máximo su potencial para la toma de decisiones, la innovación y la obtención de información crítica.

A continuación, se inspeccionará la clasificación y caracterización de los tipos de datos en proyectos de Big Data. En este análisis, descubriremos cómo cada uno de ellos desempeña un papel fundamental en el estudio de este emocionante campo, facilitando el análisis y la extracción de conocimiento.

Según (Castañeda et al ,sf) los datos estructurados son: “Datos que tienen bien definidos su longitud y su formato, como las fechas, los números o las cadenas de caracteres.” Por otro lado, (Joyanes, 2013), define a los datos estructurados como: “Datos con formato o esquema fijo que poseen campos fijos. En estas fuentes, los datos vienen en un formato bien definido que se especifica en detalle, y que conforma las bases de datos relacionales.”

Los datos estructurados son fácilmente organizados en tablas con campos bien definidos y formatos específicos. Algunos ejemplos claros de datos estructurados incluyen:

- Números de Teléfono: Los números de teléfono, como (123) 456-7890, tienen un formato fijo y una longitud específica.
- Fechas: Fechas como 01/01/2023 siguen un formato específico (día/mes/año) y tienen una estructura fija.
- Datos de Transacciones Financieras: Detalles sobre compras, transferencias y pagos en transacciones bancarias.
- Datos de GPS de Dispositivos Móviles: Datos de ubicación que pueden analizarse para entender patrones de movimiento de personas o vehículos en una ciudad.
- Bases de Datos Relacionales: Los datos almacenados en bases de datos SQL donde cada campo tiene un tipo de datos específico y un tamaño definido.

Para los datos no estructurados (Castañeda et al ,sf) los define como: “Datos en el formato tal y como fueron recolectados, sin un formato específico. No se pueden almacenar dentro de tablas ya que no se puede desgranar su información a tipos básicos de datos.” Por otro lado, (Joyanes, 2013), los define como: "Los datos no estructurados son datos sin tipos predefinidos. Se almacenan como "documentos" u "objetos" sin estructura uniforme, y se tiene poco o ningún control sobre ellos."

Los datos no estructurados carecen de un formato específico y no se pueden organizar fácilmente en tablas. Aquí tienes algunos ejemplos de datos no estructurados:

- Correos Electrónicos: El contenido de un correo electrónico puede variar significativamente y no sigue una estructura específica.
- Archivos de Audio y Video: Los archivos de audio y video contienen datos que no tienen una organización predefinida y no pueden ser desglosados en tipos de datos básicos.
- Redes Sociales: Los comentarios, fotos y videos en plataformas de redes sociales no siguen un formato estándar y son ejemplos de datos no estructurados.
- Datos de Imágenes y Videos: Imágenes médicas como resonancias magnéticas o radiografías.
- Datos de texto: Conjuntos grandes de datos de texto como libros, artículos de noticias o revisiones de productos.
- Datos de Clickstream: Datos que registran acciones de usuarios en un sitio web, como clics y tiempos de permanencia.
- Datos de Audio: Grabaciones de llamadas de centros de atención al cliente.

Mientras que los datos semiestructurados (Castañeda et al ,sf) los define como: “Datos que no se limitan a campos determinados, pero que contiene marcadores para separar los diferentes elementos. Es una información poco regular como para ser gestionada de una forma estándar. Estos poseen sus propios metadatos semiestructurados que describen los objetos y las relaciones entre ellos, y pueden acabar siendo aceptados por convención.” Por otro lado, (Joyanes, 2013), los define como: “Datos que no tienen formatos fijos, pero contienen etiquetas y otros marcadores que permiten separar los elementos datos. La lectura de datos semiestructurados requiere el uso de reglas complejas que determinan cómo proceder después de la lectura de cada pieza de información.”

Los datos semiestructurados tienen cierta estructura, pero no se ajustan completamente a un formato fijo. Algunos ejemplos de datos semiestructurados incluyen:

- Datos de Redes Sociales: Datos de Twitter que incluyen tweets, retweets, me gusta y comentarios.
- Datos de Sensores IoT (Internet de las cosas): Datos de sensores en máquinas industriales que monitorean temperatura, presión y otras condiciones.
- JSON (JavaScript Object Notation): Los datos en formato JSON tienen una estructura flexible con etiquetas que separan los diferentes elementos, lo que permite cierto grado de organización, pero no sigue una estructura tabular rígida.
- XML (eXtensible Markup Language): Los datos en formato XML contienen marcadores que definen la estructura del documento, lo que facilita la organización, pero no tiene una estructura estricta como las bases de datos relacionales.

- Archivos CSV con Encabezados Variables: Los archivos CSV pueden ser semiestructurados si tienen encabezados variables, lo que significa que diferentes filas pueden tener diferentes campos, lo que permite cierta flexibilidad en la estructura de los datos.

La diversidad de datos en proyectos Big Data se manifiesta a través de sus distintas estructuras: datos estructurados, no estructurados y semiestructurados. Estos actúan como el lienzo donde se despliegan las complejidades y matices de la información digital. Para comprender a fondo esta diversidad, exploramos la clasificación de (IBM, 2014, como se citó en Monleón-Getino, 2015), que nos presenta cinco tipos específicos de datos en el contexto del Big Data.

Tabla 3

Categorías de Datos en el Big Data según IBM

Datos	Descripción	Ejemplos
Web and Social Media	Contenido web e información que es obtenida de las redes sociales	www, Facebook, Twitter, LinkedIn, blogs
Machineto-Machine (M2M):	Tecnologías que permiten conectarse a otros dispositivos. M2M utiliza dispositivos como sensores o medidores que capturan algún acontecimiento en particular.	Velocidad, temperatura, presión, variables meteorológicas, variables químicas
Big Transaction Data	Incluye datos procedentes de transacciones masivas de los centros de atención telefónica, de banca, finanzas, atención a clientes, etc	Incluye registros de facturación, en telecomunicaciones los llamados registros detallados de las llamadas (Call Detail Record o CDR), etc.
Biometrics	Información biométrica. En el área de seguridad e inteligencia, los datos biométricos son sumamente importantes para los gobiernos, seguridad privada, etc	Huellas digitales, escaneo de la retina, reconocimiento facial, genética, etc
Human Generated	Datos digitales generados por las personas, en sentido genérico	Notas de voz, correos electrónicos, documentos electrónicos, resultados de estudios médicos, multas, etc

Nota. (IBM, 2014, como se citó en Monleón-Getino, 2015).

2.2.- Procesamiento del Lenguaje Natural

En la era digital actual, las redes sociales han dado lugar a un vasto océano de intercambio de ideas, opiniones y experiencias. La inmensa cantidad de comentarios generados en estas plataformas presenta un desafío intrigante: ¿cómo podemos comprender y clasificar la gama completa de emociones expresadas en este extenso flujo de texto digital? Es aquí donde entra en juego el fascinante campo del Procesamiento del Lenguaje Natural (PLN).

El PLN, una disciplina en la intersección de la inteligencia artificial y la lingüística computacional, se centra en dotar a las máquinas con la capacidad de entender y procesar el lenguaje humano. En este contexto, el análisis de sentimiento se destaca como una aplicación crucial del PLN, permitiendo la evaluación automatizada de las emociones transmitidas a través de los comentarios en redes sociales.

Este viaje de exploración nos sumergirá en los fundamentos del PLN y cómo se aplica específicamente al análisis de sentimiento en comentarios de redes sociales. Desde la identificación de opiniones positivas y negativas hasta la interpretación de matices emocionales en mensajes breves y directos, el desafío radica en enseñar a las máquinas a captar la complejidad inherente a la expresión humana en línea.

A lo largo de esta exploración, examinaremos cómo los avances en modelos de aprendizaje profundo han revolucionado la capacidad del PLN para discernir las sutilezas del lenguaje emocional en un entorno digital dinámico. Al comprender el estado actual de estas tecnologías, busquemos arrojar luz sobre cómo el PLN no solo transforma la forma en que interactuamos con la información, sino también cómo puede proporcionar percepciones valiosas sobre el pulso emocional de la sociedad digital.

2.2.1.- Definición de NLP

El procesamiento del lenguaje natural es una rama de la informática, la inteligencia artificial y la lingüística que se ocupa de las interacciones entre las computadoras y el lenguaje humano (natural) como el español, inglés, francés y alemán. Son sistemas de comunicación complejos y ricos que han evolucionado naturalmente entre los seres humanos. El PLN combina los campos de la informática, la inteligencia artificial y la lingüística para explorar las complejas interacciones entre las computadoras y el lenguaje humano. En esencia, se enfoca en cómo las máquinas pueden comprender, interpretar y generar el lenguaje humano de manera similar a cómo lo hacen los seres humanos.

Las lenguas naturales, por otro lado, son los idiomas que las personas utilizan en su comunicación cotidiana. Estos idiomas se adquieren y se desarrollan orgánicamente en las sociedades humanas, y se utilizan para expresar una amplia gama de conocimientos, emociones y respuestas en diversas situaciones. En cualquier forma de comunicación, los lenguajes naturales sirven como un medio vital para expresar nuestras ideas, emociones y pensamientos, permitiéndonos conectarnos con los demás y comprender nuestro entorno. El PLN busca replicar esta complejidad lingüística en el mundo digital, explorando cómo las computadoras pueden entender y utilizar el lenguaje humano de manera efectiva y significativa.

En este intrincado tejido lingüístico, el PLN despliega su magia. Según las definiciones de los expertos, como (Liddy , 2001) quien describe el PLN como: "El procesamiento del lenguaje natural es una gama de técnicas computacionales motivadas teóricamente para analizar y representar textos que ocurren naturalmente en uno o más niveles de análisis lingüístico con el fin de lograr un procesamiento del lenguaje similar al humano para una variedad de tareas o aplicaciones". Por otro lado, según (Vide, 2004, citado en Vilares, 2005) define al PLN como: "La rama de las ciencias computacionales encargada del diseño e implementación de los elementos software y hardware necesarios para el tratamiento computacional del lenguaje natural, entendiendo como tal todo lenguaje humano, en contraposición a los lenguajes formales"

Además, (Kumar, 2011) nos ofrece una visión clara: ""El procesamiento del lenguaje natural (PNL) es un campo de la informática y la lingüística que se ocupa de las interacciones entre las computadoras y los lenguajes humanos (naturales). Los sistemas de generación de lenguaje natural convierten la información de bases de datos informáticas en lenguaje humano legible. Los sistemas de comprensión del lenguaje natural convierten muestras del lenguaje humano en representaciones más formales, como árboles de análisis o lógica de primer orden, que son más fáciles de manipular para los programas de computadora".

En este viaje, nos sumergimos en la esencia misma de la comunicación humana, desglosando las palabras en su significado más íntimo y luego reconstruyéndose en un lenguaje digital que la

computadoras pueden entender. Este es el mundo fascinante y complejo del procesamiento del lenguaje natural, donde las fronteras entre la humanidad y la tecnología se desdibujan, permitiéndonos explorar y entender el lenguaje humano en su forma más pura y esencial.

2.2.2.- Definición de términos claves

En el ámbito del Procesamiento del Lenguaje Natural (PLN), es esencial comprender ciertos términos clave que forman la base para el análisis de

sentimientos. Estos términos no sólo proporcionan un marco conceptual, sino que también son fundamentales para abordar aspectos específicos de la tarea, como la polaridad en el análisis de sentimientos, la identificación de emociones en el texto y la aplicación de modelos de aprendizaje automático en NLP

Explorar el terreno del PLN nos lleva a comprender conceptos esenciales que son piedras angulares para el análisis de sentimientos y la comprensión profunda del texto en las redes sociales. En este contexto, abordaremos la importancia de la polaridad en el análisis de sentimientos, la identificación de emociones en el texto y la aplicación de modelos de aprendizaje automático en el PLN.

2.2.2.1.- Polaridad en el análisis de sentimiento

La importancia de "Lo que piensan los demás" siempre ha sido un elemento fundamental en nuestro proceso de toma de decisiones. Antes de la generalización del conocimiento de la World Wide Web, solicitábamos recomendaciones a amigos sobre mecánicos, discutíamos opciones de voto en elecciones locales. Sin embargo, con la llegada de Internet y la Web, se ha vuelto posible acceder a las opiniones y experiencias de un amplio grupo de personas más allá de nuestro círculo personal o críticos profesionales conocidos. Este cambio ha transformado significativamente la dinámica de cómo obtenemos información para nuestras decisiones.

En el ámbito específico del análisis de sentimientos, la polaridad se erige como una dimensión esencial para comprender las expresiones humanas en texto. Como señala (Pang y Lee, 2008, p.16) "Las etiquetas de polaridad son fundamentales para resumir el contenido del texto sobre un tema, ya sea positivo, negativo o neutral. La polaridad, en este contexto, se refiere a la dirección de la opinión expresada."

Por otro lado (Liu, 2012, p.17), en su obra "Sentiment Analysis and Opinion Mining", destaca la importancia de la polaridad al definirla como "Una opinión consta de dos componentes clave: un objetivo "g" y un sentimiento "s" sobre el objetivo, es decir, (g, s), donde g puede ser cualquier entidad o aspecto de la entidad sobre la cual se ha expresado una opinión. y s es un sentimiento positivo, negativo o neutral, o una puntuación de calificación numérica que expresa la fuerza/intensidad del sentimiento (por ejemplo, de 1 a 5 estrellas). Positivo, negativo y neutral se denominan orientaciones (o polaridades) de sentimiento (u opinión)".

2.2.2.2.- Emociones y su identificación en el texto

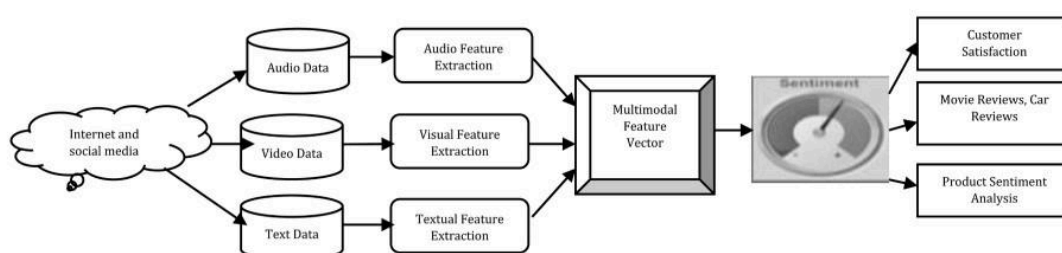
La identificación y comprensión de las emociones en el análisis de sentimientos constituyen un componente esencial para capturar la complejidad de las respuestas humanas en el vasto universo digital. Más allá de simplemente clasificar la polaridad de un texto como positiva, negativa o neutral, la capacidad de discernir emociones específicas agrega una capa de profundidad invaluable.

En palabras de (Cambria et al, 2012), en su obra "Sentic Computing: Techniques, Tools, and Applications", las emociones son el "pegamento" que une las palabras a las experiencias humanas. La identificación precisa de estas emociones en el texto permite no solo entender la orientación general del sentimiento, sino también desentrañar la riqueza de matices emocionales que caracterizan la comunicación humana. Es un proceso que va más allá de etiquetar un texto como feliz o triste; implica la captura de matices como la alegría, el disgusto, la sorpresa y el temor.

La importancia de identificar emociones en el texto radica en su capacidad para proporcionar percepciones más profundas sobre la respuesta del usuario. Como menciona (Gandhi et al., 2022, p. 424) en su artículo "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions", ve la importancia de las emociones: "Emociones y los sentimientos juegan un papel importante en nuestra vida diaria. Ellos ayudan en la toma de decisiones, aprendizaje, comunicación y conciencia de la situación en entornos centrados en el ser humano." Esto abarca mucho más allá de lo normalmente establecido en el campo de la minería de opiniones, ya que además, en su artículo, destaca la importancia que tienen otros medios de difusión de información en las redes sociales, como lo son los audios, videos y también el texto que se pueden ver en la figura 3.

Figura 3

Otros medios de minería de datos.



Nota. (Gandhi et al., 2022, p. 426).

Texto

En el ámbito del análisis de sentimiento multimodal, la modalidad de texto destaca como un componente esencial y dominante. Su relevancia radica en su capacidad para identificar los sentimientos subyacentes y desentrañar matices emocionales ocultos. Aunque el análisis de sentimientos textuales ha demostrado resultados prometedores, la prevalencia actual de datos compartidos en forma de videos plantea nuevos desafíos, ya que la interpretación basada solo en texto puede ser limitada en estos escenarios.

Visual

La modalidad visual desempeña un papel crucial al identificar sentimientos y opiniones subyacentes a través de características visuales. Por ejemplo, cuando se encuentra un texto como "este es un mouse bastante bueno", la interpretación sólo basada en datos textuales podría dificultar la distinción entre un ratón animal original y un ratón de ordenador. En este contexto, las características visuales son fundamentales, y la combinación de texto y elementos visuales crea un sistema bimodal, generando resultados más precisos en comparación con sistemas unimodales.

Audio

La modalidad de audio utiliza características acústicas para generar información textual y permite identificar datos de videos, así como el tono de los oradores. La combinación de las tres modalidades (texto, visual y audio) genera un modelo de análisis más robusto. En situaciones que involucran elementos como humor, sarcasmo o imágenes que desafían la detección de sentido común, la combinación de modalidades demuestra ser especialmente efectiva, permitiendo una identificación más precisa de los sentimientos del usuario.

En conclusión, la identificación de emociones en el texto no solo enriquece la comprensión del sentimiento humano en la era digital, sino que también potencia la capacidad de las máquinas para interactuar de manera más intuitiva y personalizada, especialmente en el contexto de comentarios en redes sociales. La habilidad de discernir las emociones expresadas en plataformas digitales permite una comprensión más profunda de las respuestas de los usuarios, contribuyendo a una interpretación más precisa y contextualizada de sus experiencias y opiniones.

Este enfoque no se limita solo al análisis de texto, sino que se extiende a nuevas modalidades como el análisis de emociones en videos y audios. La incorporación de estas modalidades multimodales ofrece una visión más completa de la expresión emocional, capturando matices que van más allá de las palabras escritas. Por ejemplo, la detección de emociones en videos puede revelar gestos faciales y

expresiones que complementan o incluso contradicen el contenido textual, proporcionando así una comprensión más holística.

En el dinámico mundo de las redes sociales, donde la interacción se diversifica a través de diferentes medios, la identificación de emociones en múltiples modalidades se vuelve esencial. Esta perspectiva ampliada no solo mejora la capacidad de las máquinas para adaptarse a la complejidad de las respuestas emocionales, sino que también abre nuevas posibilidades para la personalización de interacciones en línea y el desarrollo de sistemas más empáticos. Así, la identificación de emociones no solo enriquece la comprensión de textos, sino que también impulsa la evolución de tecnologías que se adaptan de manera más precisa a la expresión emocional en todas sus formas en el vasto paisaje digital.

2.2.3.- Niveles del PLN

Para cumplir su objetivo, un sistema de NLP necesitará hacer uso de una cantidad considerable de conocimiento acerca de la estructura del lenguaje. Este conocimiento se puede estructurar en niveles, como lo menciona (Liddy, 2001), quien afirma que: "El método más explicativo para presentar lo que realmente sucede dentro de un sistema de procesamiento del lenguaje natural es mediante el enfoque de los niveles del lenguaje. Esto también se conoce como modelo sincrónico del lenguaje y se distingue del modelo secuencial anterior, que plantea la hipótesis de que los niveles de procesamiento del lenguaje humano se suceden uno al otro de manera estrictamente secuencial.

La investigación psicolingüística sugiere que el procesamiento del lenguaje es mucho más dinámico, ya que los niveles pueden interactuar en una variedad de órdenes". Por lo que el análisis del lenguaje natural se divide en varios niveles amplios, como el análisis fonológico, morfológico, léxico, sintáctico, semántico, pragmático y del discurso. Estos se describirán bajo (Kumar, 2011) que los definió de la siguiente manera:

Análisis fonológico: La fonología es el análisis del lenguaje hablado. Por tanto, se trata del reconocimiento y generación de voz. La tarea principal del sistema de generación y reconocimiento de voz es tomar una forma de onda acústica como entrada y producir como salida una cadena de palabras. La fonología es una parte del análisis del lenguaje natural, que se ocupa de él. El área de la lingüística computacional que se ocupa del análisis del habla es la fonología computacional.

Análisis Morfológico: Es la fase más elemental de la PNL. Se trata de la formación de palabras. En esta fase, las palabras individuales se analizan según sus componentes llamados "morfemas". El morfema es el componente gramatical básico que forma las palabras. La tarea de dividir una palabra en sus morfemas se

llama análisis morfológico. Un morfema se define como una unidad mínima significativa en un idioma, que no puede dividirse en unidades más pequeñas.

Análisis léxico: En esta fase del análisis del lenguaje natural se comprueba la validez de las palabras según el léxico. Lexicon significa diccionario. Es una colección de todas las posibles palabras válidas del idioma junto con su significado. En PNL, la primera etapa del procesamiento del texto de entrada es escanear cada palabra de la oración y calcular (o buscar) toda la información lingüística relevante sobre esa palabra. El léxico proporciona las reglas y datos necesarios para llevar a cabo el análisis de la primera etapa. Se verifican los detalles de las palabras, como su tipo (sustantivo, verbo, adverbio, y otros detalles).

Análisis sintáctico: La sintaxis se refiere al estudio de las relaciones formales entre palabras de oraciones. En esta fase se comprueba la validez de una frase según las reglas gramaticales. Para realizar el análisis sintáctico se requieren conocimientos de gramática y técnica de análisis sintáctico. La gramática es una especificación formal de las reglas permitidas en el idioma y el análisis sintáctico es un método para analizar una oración para determinar su estructura de acuerdo con la gramática. La gramática más común utilizada para el análisis sintáctico de lenguajes naturales es la gramática libre de contexto (CFG), también llamada gramática de estructura de frase y gramática de cláusula definida. Estas gramáticas se describen en detalle en una sección separada.

Análisis Semántico: La semántica se ocupa del significado de las oraciones del lenguaje natural. En esta fase se comprende el significado de la frase. Si las computadoras desean comunicarse mediante lenguaje natural, se requiere una representación computacional de estas oraciones (llamada representación del conocimiento) para capturar el significado. La semántica del lenguaje natural consta de dos características principales:

1. Cómo representar los significados de una manera que permita todas las manipulaciones necesarias (particularmente la inferencia).
2. Cómo relacionar estas representaciones con esa parte del modelo lingüístico que se ocupa de la estructura (gramática o sintaxis).

En gran medida, el objetivo tanto de la semántica como de la representación del conocimiento es formalizar lo que la gente sabe sobre palabras y oraciones; y lo que saben sobre el significado de palabras, frases, etc. del lenguaje natural. Estos también están relacionados con el conocimiento sobre el mundo. Por ejemplo, para modelar una expresión en lenguaje natural, que analiza el tiempo y la duración, necesitamos una representación del tiempo que refleje la forma en que la gente habla sobre el tiempo y la forma en que piensa sobre el tiempo. Hay dos enfoques principales para encontrar una teoría del significado computacionalmente utilizable

para el lenguaje natural. Algunos de estos enfoques se analizarán más adelante en este capítulo.

Análisis pragmático y del discurso: La pragmática es el estudio de la relación entre el lenguaje y el contexto de uso. El contexto de uso incluye cosas tales como las identidades de las personas y los objetos. Por tanto, la pragmática incluye el análisis de cómo se utiliza el lenguaje para referirse a personas y cosas. El entorno contextual incluye el contexto del discurso. En esta fase de la PNL se comprende la intención principal del hablante detrás de un mensaje.

El discurso es una colección de oraciones, pero la colección arbitraria de múltiples oraciones no constituye discurso. En el discurso, las oraciones deben ser coherentes. La colección de oraciones bien formadas e interpretables individualmente a menudo forma el discurso coherente. Durante el análisis del discurso, se requiere la comprensión de los referentes. Los referentes se utilizan en la última parte del discurso, en lugar de sustantivos introducidos en la primera parte del discurso.

Los lenguajes naturales ofrecen muchas formas de referirse a entidades. Cada forma de oración ayuda al oyente a saber cómo debe procesarse con respecto al modelo del discurso y el conjunto de creencias sobre la palabra son palabras de reemplazo. La interpretación del discurso requiere que uno construya una representación evolutiva del estado del discurso llamada modelo de discurso que contenga la representación de entidades y las relaciones entre entidades.

2.2.4.- Aplicaciones del NLP

El Procesamiento del Lenguaje Natural (NLP) ha emergido como un campo fascinante que permite a las máquinas comprender, interpretar y generar lenguaje humano de manera inteligente. Su aplicación se extiende a lo largo y ancho, transformando diversas esferas de nuestras interacciones diarias. En esta sección, se verán las variadas aplicaciones del NLP que han redefinido la forma en que nos comunicamos y accedemos a la información en el vasto universo digital.

Desde la creación de asistentes virtuales que responden a nuestras preguntas en lenguaje natural hasta la capacidad de analizar grandes cantidades de texto en redes sociales para entender las opiniones públicas, el NLP ha demostrado ser una herramienta versátil y poderosa. Su influencia se extiende desde la traducción automática hasta la generación de contenido automatizado, abriendo nuevas posibilidades en otros sectores.

Esta sección profundizará en algunas de las aplicaciones más destacadas del NLP, explorando cómo esta tecnología ha allanado el camino para mejoras significativas

en la eficiencia, la accesibilidad a la información y la toma de decisiones en distintos campos. A través de ejemplos concretos y perspectivas respaldadas por investigaciones y expertos, descubriremos cómo el NLP continúa desempeñando un papel fundamental en la evolución de la interacción humano-máquina y en la manera en que conceptualizamos el procesamiento del lenguaje en la era digital.

Aunque pueda parecer algo redundante, según (Lane et al., 2018, p3), la definición del Procesamiento del Lenguaje Natural (NLP) se articula como "Un procesamiento generalmente implica traducir el lenguaje natural en datos (números) que una computadora puede usar para aprender sobre el mundo. Y esta comprensión del mundo a veces se utiliza para generar texto en lenguaje natural que refleja esa comprensión."

Ahora, después de establecer esta base, se detallará el análisis de las aplicaciones del NLP, explorando cómo esta tecnología ha allanado el camino para mejoras significativas en la eficiencia, la accesibilidad a la información y la toma de decisiones en distintos campos. A través de ejemplos concretos y perspectivas respaldadas por investigaciones y expertos, descubriremos cómo el NLP continúa desempeñando un papel fundamental en la evolución de la interacción humano-máquina y en la manera en que conceptualizamos el procesamiento del lenguaje en la era digital.

El Procesamiento del Lenguaje Natural (NLP) se adentra en el fascinante mundo de la comprensión y generación inteligente del lenguaje humano por parte de las máquinas. Su aplicación abarca un espectro amplio, transformando nuestra interacción con la información en el vasto universo digital. Habiendo esclarecido la esencia de NLP, ahora se verá cómo esta tecnología ha impregnado diversos aspectos de nuestra vida cotidiana.

Desde la creación de asistentes virtuales que responden a preguntas en lenguaje natural hasta la capacidad de analizar grandes volúmenes de texto en redes sociales para entender las opiniones públicas, el NLP ha demostrado ser una herramienta poderosa y versátil. Sus aplicaciones se extienden desde la traducción automática hasta la generación de contenido automatizado, abriendo nuevas perspectivas en una variedad de sectores.

En esta sección, se verán algunas de las aplicaciones más destacadas del NLP, examinando cómo esta tecnología ha catalizado mejoras sustanciales en la eficiencia, la accesibilidad a la información y la toma de decisiones en diversas disciplinas. A través de ejemplos concretos respaldados por investigaciones y expertos, descubriremos cómo el NLP continúa desempeñando un papel fundamental en la evolución de la interacción humano-máquina y en la conceptualización del procesamiento del lenguaje en la era digital.

Para dar inicio a este recorrido, consultamos la Tabla 4, que presenta una clasificación de diversas aplicaciones del NLP en distintos dominios.

Tabla 4

Clasificación de las aplicaciones del procesamiento de lenguaje natural.

Buscar	Web	Documentos	Autocompletar
Edición	Ortografía	Gramática	Estilo
Diálogo	Chatbot	Asistente	Planificación
Escribir	Índice	Concordancia	Tabla de contenido
Email	Filtro de spam	Clasificación	Priorización
Minería de textos	Resumen	Extracción de conocimiento	Diagnostico médico
Análisis de sentimientos	Monitoreo de la moral de la comunidad	Clasificación de reseñas de productos	Atencion al cliente
Predicción del comportamiento	Finanzas	Previsión electoral	Marketing
Escritura creativa	Poesía	letra de canciones	Guiones

Nota. (Lane et al., 2018, p8).

Buscar en la Web: Utilización de NLP para mejorar la precisión y relevancia de los resultados de búsqueda en la web.

Documentos: Análisis y procesamiento de documentos utilizando NLP para extraer información y comprender el contenido.

Autocompletar: Empleo de NLP para prever y sugerir palabras o frases mientras se escribe, como en motores de búsqueda o teclados de teléfonos móviles.

Edición: Uso de NLP para corregir errores de ortografía, gramática y estilo en documentos.

Diálogo: Implementación de chatbots o asistentes virtuales que utilizan NLP para entender y responder a las consultas de los usuarios.

Escribir: Generación de texto automáticamente mediante el uso de modelos de lenguaje entrenados por NLP.

Índice: Creación de índices para facilitar la búsqueda y organización de información en grandes conjuntos de datos.

Concordancia: Identificación y presentación de palabras o frases relacionadas con un término específico.

Tabla de Contenido: Creación automática de una estructura jerárquica para facilitar la navegación y comprensión de un documento.

Email: Aplicación de NLP en la gestión de correos electrónicos, incluyendo filtros de spam, clasificación y priorización.

Minería de Textos: Extracción de información valiosa y patrones a partir de grandes cantidades de texto.

Resumen: Generación automática de resúmenes concisos a partir de textos más extensos.

Extracción de Conocimiento: Utilización de NLP para identificar y extraer información relevante y estructurada.

Diagnóstico Médico: Aplicación de NLP en la interpretación de datos médicos para ayudar en el diagnóstico.

Análisis de Sentimientos: Evaluación y clasificación de la polaridad en textos para comprender las emociones y opiniones expresadas.

Monitoreo de la Moral de la Comunidad: Seguimiento y análisis de las actitudes y emociones de una comunidad en línea.

Clasificación de Reseñas de Productos: Utilización de NLP para categorizar y evaluar opiniones de productos.

Atención al Cliente: Implementación de chatbots o sistemas de respuesta automática para brindar asistencia al cliente.

Predicción del Comportamiento: Utilización de NLP en la anticipación de comportamientos futuros basados en patrones históricos.

Finanzas: Aplicación de NLP en el análisis de datos financieros, pronósticos y toma de decisiones.

Previsión Electoral: Utilización de NLP para analizar opiniones y prever resultados en elecciones.

Marketing: Aplicación de NLP en estrategias de marketing, como la personalización de mensajes y la comprensión del feedback del cliente.

Escritura Creativa: Generación de contenido creativo, como guiones, poemas y letras de canciones, mediante NLP.

2.3.- Análisis de Sentimientos

El análisis de sentimientos, también conocido como minería de opiniones, se ha convertido en una disciplina crucial en el ámbito del procesamiento de lenguaje natural y la inteligencia artificial. Este campo se enfoca en comprender y evaluar las emociones, actitudes y opiniones expresadas en el lenguaje humano, especialmente a través de plataformas en línea y redes sociales. El crecimiento exponencial de la información disponible en texto ha impulsado la necesidad de técnicas automatizadas que puedan discernir y clasificar las tonalidades emocionales presentes en grandes conjuntos de datos.

En esta sección, se explorará en detalle el análisis de sentimiento, desde sus fundamentos teóricos hasta las herramientas prácticas y métodos utilizados para extraer información valiosa de las expresiones lingüísticas. Desde la importancia del procesamiento de texto hasta la aplicación de técnicas avanzadas de aprendizaje automático, nos sumergimos en un análisis exhaustivo que destaca la relevancia y las aplicaciones prácticas de esta disciplina en el análisis de datos contemporáneo.

2.3.1.- Qué es el análisis de sentimiento y cómo funciona

Según(Nick Barney, s.f) el análisis de sentimientos, es una disciplina dentro del procesamiento del lenguaje natural (NLP) que tiene como objetivo identificar y comprender las emociones, actitudes y opiniones expresadas en un texto. Es una herramienta ampliamente utilizada por las organizaciones para evaluar y categorizar las opiniones de los usuarios sobre productos, servicios o ideas.

Para realizar el análisis de sentimientos, se emplean sistemas y herramientas que procesan el texto proveniente de diversas fuentes en línea, tales como correos electrónicos, publicaciones de blogs, reseñas en línea, tickets de atención al cliente, artículos de noticias, respuestas a encuestas, chats web, tweets, foros y comentarios.

¿Cómo funciona el análisis de sentimientos?

Estos son los pasos principales en el proceso de análisis de sentimientos:

Recopilación de datos: Se identifican y se recopilan los textos relevantes para el análisis, utilizando técnicas como el web scraping o interfaces de programación de aplicaciones (API) de recopilación de datos.

Limpieza de datos: Los datos recopilados se procesan y se eliminan elementos innecesarios o irrelevantes, como ruido o partes del discurso que no aportan información valiosa para el análisis de sentimientos.

Extracción de características: Se utilizan algoritmos de aprendizaje automático para extraer automáticamente características del texto que son indicativas de los sentimientos expresados. Esto puede incluir el uso de técnicas como la bolsa de palabras, que registra la frecuencia de aparición de palabras clave en el texto.

Elección de un modelo de aprendizaje automático: Se selecciona un modelo de aprendizaje automático adecuado para el análisis de sentimientos, el cual puede ser basado en reglas predefinidas o utilizar técnicas de aprendizaje profundo. Los modelos basados en reglas suelen ser utilizados en dominios específicos donde se requiere un alto grado de precisión y control humano, mientras que los modelos de aprendizaje profundo se enfocan en aprender patrones a partir de grandes conjuntos de datos.

Clasificación de sentimientos: Una vez que se ha elegido un modelo, se utiliza para asignar una puntuación o etiqueta de sentimiento al texto analizado, indicando si es positivo, negativo o neutral. Además, se pueden realizar análisis a nivel de documento, oración o incluso suboración para obtener un nivel de detalle más preciso.

2.3.1.1.- Tipos de análisis de sentimientos

Según (Husna Sayedi, 2021), los algoritmos de análisis de opinión generalmente se dividen en tres grupos diferentes: automático, basado en reglas e híbrido. Profundicemos en cada uno de los tres tipos diferentes de algoritmos de análisis de sentimientos.

Automático: En los algoritmos de análisis automático de sentimientos, se utilizan técnicas de aprendizaje automático para etiquetar datos de texto con diferentes sentimientos. Se aplica un modelo de clasificación de aprendizaje automático para determinar si el texto de entrada pertenece a un conjunto específico de clases de sentimientos, como positivo, negativo o neutral.

El proceso de análisis automático de sentimientos sigue un flujo específico. En primer lugar, los datos de entrada se introducen en un modelo donde se extraen características y se asocian con una etiqueta particular. Las técnicas de extracción

de características pueden incluir la vectorización de texto, el uso de bolsas de palabras o bolsas de n-gramas, incrustaciones de palabras y frecuencias de palabras. A continuación, los datos de entrenamiento se etiquetan con diferentes sentimientos para entrenar un modelo supervisado.

En el siguiente paso, se generan predicciones de clases o etiquetas para nuevos puntos de datos utilizando el modelo aprendido durante el entrenamiento. Los modelos de clasificación comunes utilizados en el análisis automático de sentimientos incluyen el clasificador Naive Bayes, la regresión logística, las máquinas de vectores de soporte, la regresión lineal y el aprendizaje profundo.

Basado en reglas: En los algoritmos de análisis de opiniones basados en reglas, el sistema etiqueta automáticamente los datos de entrada utilizando un conjunto de reglas predefinidas para identificar la polaridad de las opiniones de los usuarios. Estos algoritmos se apoyan en técnicas de procesamiento del lenguaje natural (PNL), como la lematización, el etiquetado de partes del discurso, el análisis léxico y la tokenización.

Un ejemplo de un modelo de análisis de sentimientos basado en reglas es el siguiente: se crean listas de palabras asociadas a diferentes sentimientos. Por ejemplo, las palabras "horrible", "peor" pueden considerarse negativas, mientras que las palabras "genial", "amor", "increíble" pueden considerarse positivas. Luego, se analiza una frase de entrada, contabilizando las palabras polarizadas. Este recuento se clasifica en diferentes categorías, como "positivo", "negativo" o "neutral", según el recuento más alto de palabras polarizadas asociadas a los sentimientos establecidos.

Híbrido: Un algoritmo de sentimiento híbrido combina las técnicas utilizadas en los modelos de análisis automático y basado en reglas. Este enfoque utiliza tanto técnicas de aprendizaje automático como PNL para crear un modelo de análisis de sentimientos más complejo y preciso. Aunque la construcción de estos algoritmos puede ser más compleja, los resultados suelen ser más precisos que al utilizar un enfoque automático o basado en reglas por separado.

A continuación, se presenta una visión más general del análisis de sentimientos, según lo planteado por (Medhat, Hassan y Korashy, 2014).

Machine Learning: El enfoque de Machine Learning abarca diversos modelos estadísticos capaces de aprender de las bases de datos con las que son alimentados. Mayormente, estos modelos son de tipo supervisado, lo que implica que la primera serie de datos está previamente clasificada y las series posteriores se etiquetan automáticamente. Los métodos supervisados reconocen patrones comunes en los textos que no son evidentes bajo lecturas convencionales. En

contraste, los métodos no supervisados se alimentan de datos no clasificados, y el modelo determina los patrones de etiquetas.

Análisis de Sentimiento basado en léxicos:

Este enfoque difiere del modelo de Machine Learning en que no tiene la capacidad de aprender de los datos con los que se alimenta. Se divide en dos tipos:

Basado en corpus: Esta metodología se aplica a grandes volúmenes de textos, como los presentes en redes sociales. Permite un análisis detallado de las opiniones de los usuarios, el reconocimiento de conceptos y la extracción de valoraciones de sentimientos mediante un corpus con documentos completamente valorizados manualmente.

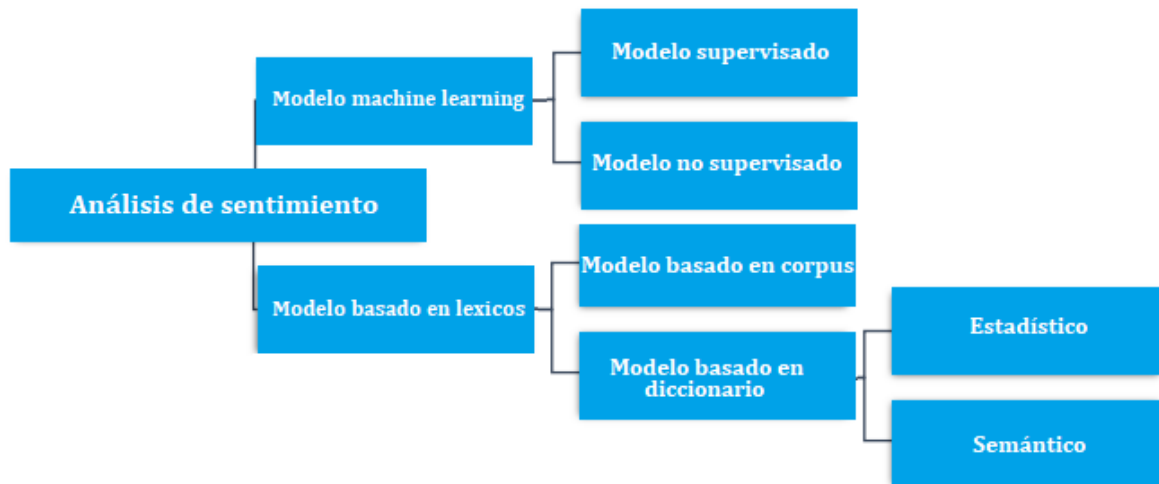
Basado en diccionarios: Este método utiliza un diccionario que contiene palabras clasificadas previamente según la polaridad de su sentimiento. La efectividad de esta técnica depende en gran medida del grado de coincidencia entre el diccionario y el texto a codificar, así como de la correcta interpretación de la polaridad de las palabras etiquetadas.

El enfoque semántico: proporciona valores de sentimiento directos y se fundamenta en principios que calculan la similitud entre palabras. Este método asigna valores de sentimiento similares a palabras que comparten proximidad semántica. Al seguir este principio, se logra capturar las relaciones semánticas entre las palabras, otorgando valores de sentimiento coherentes a términos semánticamente cercanos.

El enfoque estadístico: busca identificar patrones de coexistencia o términos iniciadores de opiniones mediante técnicas estadísticas. Este análisis se realiza a través de la observación de la concurrencia de adjetivos en un corpus de texto. Identificando patrones estadísticos, este enfoque puede derivar polaridades posteriores, ofreciendo una perspectiva valiosa sobre la opinión general expresada en el conjunto de datos analizado.

Figura 4

Modelos de Análisis de Sentimientos: Supervisados y no Supervisados



Nota. Elaboración propia.

2.3.1.2.- Aplicaciones del análisis de sentimientos

El análisis de sentimientos es una herramienta versátil que encuentra aplicaciones en diversas industrias, algunas de las cuales incluyen:

Monitoreo de redes sociales: Permite realizar un seguimiento y comprender los sentimientos de los usuarios en plataformas como Facebook, Instagram y Twitter, brindando información sobre la percepción general de una marca, producto o evento.

Gestión de la reputación de marca: Ayuda a evaluar la imagen y el impacto de una marca al analizar las opiniones y comentarios expresados por los usuarios. Esto permite a las empresas tomar medidas proactivas para mejorar su reputación y abordar cualquier problema o preocupación.

Mejora de productos y servicios: Proporciona información valiosa sobre las preferencias y necesidades de los clientes al analizar sus opiniones y comentarios. Esto permite a las empresas realizar ajustes y mejoras en sus productos y servicios para satisfacer mejor las demandas del mercado.

Análisis de campañas de marketing: Permite evaluar el impacto y la efectividad de las estrategias de marketing y publicidad al analizar las reacciones y respuestas de los usuarios. Esto proporciona información clave para optimizar futuras campañas y maximizar el retorno de la inversión en marketing.

Atención al cliente: Ayuda a clasificar y categorizar los mensajes y consultas de los clientes para una respuesta más eficiente y personalizada. Al analizar los sentimientos expresados por los clientes, las empresas pueden identificar rápidamente problemas y satisfacer las necesidades de los clientes de manera más efectiva.

Investigación de mercado: Proporciona información valiosa sobre las tendencias emergentes, las preferencias del consumidor y la información competitiva. Al analizar los sentimientos de los usuarios en línea, las empresas pueden obtener información valiosa para la toma de decisiones estratégicas en el mercado.

2.3.1.3.- Herramientas para el análisis de sentimientos

TextBlob es una biblioteca de procesamiento del lenguaje natural (NLP) para Python que simplifica muchas tareas relacionadas con el procesamiento de texto, incluyendo el análisis de sentimientos. Esta biblioteca se basa en la popular biblioteca NLTK y proporciona una interfaz sencilla y consistente para acceder a diversas operaciones lingüísticas.

Características y Funcionalidades:

Análisis de Sentimientos: TextBlob(TextBlob, 2020) es conocido por su capacidad para realizar el análisis de sentimientos de manera efectiva. La función de análisis de sentimientos clasifica el texto en tres categorías principales: positivo, negativo y neutro. Esto es particularmente útil en aplicaciones donde se busca comprender la opinión o la actitud expresada en un conjunto de textos.

Modelo Pre Entrenado: TextBlob utiliza un modelo pre entrenado para el análisis de sentimientos, lo que facilita su implementación sin necesidad de construir un modelo desde cero. Este modelo se entrena en grandes conjuntos de datos etiquetados para reconocer patrones y características asociadas a distintos tonos emocionales.

Facilidad de Uso: Una de las ventajas clave de TextBlob es su interfaz simple y fácil de usar. Esto permite a los usuarios realizar tareas de procesamiento de lenguaje natural, como análisis de sentimientos, sin requerir un conocimiento profundo de técnicas de aprendizaje automático.

Personalización: Aunque TextBlob viene con un modelo pre entrenado, brinda la posibilidad de personalizar o ajustar el modelo según las necesidades específicas del usuario. Esto puede ser beneficioso al adaptar el análisis de sentimientos a contextos particulares o dominios especializados.

VADER (Valence Aware Dictionary and sEntiment Reasoner)

VADER (Valence Aware Dictionary and sEntiment Reasoner) es una herramienta diseñada para abordar los desafíos inherentes al análisis de sentimientos en contenidos de redes sociales. El artículo "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text" de C.J. Hutto y Eric Gilbert presenta este modelo basado en reglas y lo compara con once puntos de referencia típicos, incluyendo LIWC, ANEW, el General Inquirer y técnicas orientadas al aprendizaje automático que utilizan algoritmos como Naive Bayes, Máxima Entropía y Máquina de Soporte Vectorial (SVM).

El modelo VADER se destaca por su simplicidad y eficacia en la tarea de análisis de sentimientos. La metodología utilizada implica la construcción y validación empírica de un conjunto de características léxicas específicamente adaptadas para contextos similares a los de microblogs. Estas características incluyen medidas de intensidad de sentimiento asociadas a términos lexicales, las cuales son fundamentales para evaluar el tono emocional en mensajes breves y contextos informales.

Además de las características léxicas, el modelo VADER incorpora cinco reglas generales que capturan convenciones gramaticales y sintácticas relacionadas con la expresión y énfasis de la intensidad del sentimiento. Este enfoque parsimonioso, que utiliza un conjunto mínimo de reglas, demuestra ser altamente efectivo.

Los resultados obtenidos al aplicar el modelo VADER(Hutto, Gilbert, 2014, p216) a la tarea de analizar el sentimiento de tweets son notables. El modelo supera a evaluadores humanos individuales, alcanzando una precisión de clasificación F1 de 0.96, y muestra una generalización favorable a través de diferentes contextos en comparación con los puntos de referencia utilizados en el estudio (clasificación F1 de 0.84). Estos hallazgos subrayan la robustez y la capacidad de adaptación del modelo VADER en el análisis de sentimientos en entornos dinámicos y diversos como las redes sociales.

Pysentimiento: Análisis de Sentimientos en Español

Pysentimiento es una biblioteca de procesamiento del lenguaje natural (NLP) diseñada específicamente para realizar análisis de sentimientos en textos en español. Su enfoque principal es evaluar y comprender las emociones expresadas en textos, categorizando el tono emocional del contenido en positivo, negativo o neutro.

Características y Funcionalidades:

Análisis de Sentimientos: Pysentimiento(pysentimiento, 2023) sobresale en la capacidad de realizar análisis de sentimientos efectivos en textos en español, proporcionando categorizaciones precisas del tono emocional, ya sea positivo, negativo o neutro.

Modelo Pre Entrenado: La biblioteca utiliza un modelo pre entrenado específicamente adaptado para el análisis de sentimientos en español. Este modelo ha sido entrenado en conjuntos de datos etiquetados para capturar las complejidades y matices del lenguaje en este idioma.

Facilidad de Uso: Pysentimiento se presenta con una interfaz fácil de usar, permitiendo a los usuarios realizar análisis de sentimientos sin requerir un conocimiento profundo de técnicas de aprendizaje automático.

Adaptabilidad: Aunque la biblioteca viene con un modelo pre entrenado, ofrece la posibilidad de personalización o ajuste según las necesidades específicas del usuario. Esto puede ser beneficioso al adaptar el análisis de sentimientos a contextos particulares o dominios especializados en español.

Integración con Transformers:

Además, pysentimiento se integra con Transformers, una biblioteca que proporciona miles de modelos pre entrenados para realizar tareas en diferentes modalidades, como texto, visión y audio.

Transformers (Hugging Face): Transformers proporciona miles de modelos pre entrenados para realizar tareas en diferentes modalidades, como texto, visión y audio. Estos modelos pueden aplicarse en texto para tareas como clasificación, extracción de información, respuesta a preguntas, resumen, traducción y generación de texto en más de 100 idiomas. La biblioteca ofrece APIs para descargar y utilizar fácilmente estos modelos pre entrenados, así como para ajustarlos en conjuntos de datos propios y compartirlos en la comunidad a través del modelo hub.

BETO (Spanish BERT): BETO es un modelo BERT entrenado en un extenso corpus español, similar en tamaño a BERT-Base y entrenado con la técnica Whole Word Masking.

RoBERTuito: Es un modelo de lenguaje previamente entrenado para texto de redes sociales en español. Entrenado en 500 millones de tweets siguiendo las pautas de RoBERTa, ofrece variantes con y sin caja, así como desacentuado. RoBERTuito. Destaca en tareas como detección de discurso de odio, análisis de sentimientos y emociones, y detección de ironía, superando a otros modelos en estas áreas.

Estos modelos, integrados en Pysentimiento, permiten realizar análisis de sentimientos más precisos, abordar matices emocionales y evaluar tonos expresados en emojis, proporcionando una visión más completa de la carga emocional en el lenguaje. Esta integración enriquece la capacidad de Pysentimiento para adaptarse y comprender las complejidades del análisis de sentimientos en español.

2.4.- Minería de datos

La minería de datos emerge como un pilar esencial en la vastedad de la investigación y el análisis de información. En este intrigante campo, se despliega como una herramienta crucial para destilar conocimiento valioso de extensas colecciones de datos. Su pertinencia se magnifica al permitir la revelación de patrones, identificación de tendencias y descubrimiento de relaciones ocultas, erigiendo así una base sólida sobre la cual se erige la toma de decisiones informadas. Este proceso de extracción de conocimiento se convierte en un faro que ilumina el camino hacia la comprensión y aprovechamiento óptimos de la información, desentrañando los secretos enterrados en el vasto océano de datos. En este segmento, sumérgete con nosotros mientras exploramos las metodologías clave para llevar a cabo esta fascinante tarea de descubrimiento y análisis.

2.4.1.- ¿Por qué minar datos?

Como se ha señalado previamente, en la actualidad se generan volúmenes masivos de datos a través de internet. Este incremento exponencial en la disponibilidad de datos se atribuye a la amplia informatización de nuestra sociedad y al rápido desarrollo de herramientas avanzadas para la recolección y almacenamiento de información. Empresas a nivel global generan conjuntos de datos de proporciones colosales, que incluyen desde transacciones de ventas y registros de negociación de acciones hasta descripciones de productos, promociones de ventas, perfiles empresariales, rendimiento corporativo y comentarios de clientes.

La minería de datos se presenta como una respuesta esencial para extraer conocimientos significativos de este vasto océano de datos. Este proceso permite a las organizaciones tomar decisiones informadas y estratégicas en un entorno cada vez más digitalizado. Como se destaca en el libro "Data Mining Concepts and Techniques"(Kan, Kamber, Pei, 2012, p.6) "minería es un término vívido que caracteriza el proceso que encuentra un pequeño conjunto de pepitas preciosas a partir de una gran cantidad de materia prima". Este enfoque no solo resalta la magnitud de los datos disponibles, sino también la capacidad de la minería de datos para identificar y extraer información valiosa de esta abrumadora cantidad de información.

Figura 5

Búsqueda de conocimiento en los datos.



Nota.(Kan, Kamber, Pei, 2012, p,6)

El proceso de descubrimiento de conocimiento como una secuencia iterativa de los siguientes pasos:

Limpieza de Datos: Este paso crucial implica la eliminación de ruido y datos inconsistentes, asegurando que el conjunto de datos sea confiable y esté libre de errores.

Integración de Datos: En casos en los que los datos provienen de múltiples fuentes, este proceso implica combinar y fusionar conjuntos de datos para crear una visión unificada y completa.

Selección de Datos: Pertinente a la tarea de análisis, este paso se centra en recuperar datos específicos de la base de datos que son relevantes para los objetivos de la minería.

Transformación de Datos: Esta fase implica la conversión y consolidación de datos en formas apropiadas para la minería, logradas mediante operaciones de resumen o agregación.

Minería de Datos: En el núcleo del proceso, se aplican métodos inteligentes para extraer patrones e ideas significativas del conjunto de datos.

Evaluación de Patrones: Para discernir patrones verdaderamente valiosos que representen conocimiento, esta etapa utiliza medidas de interés.

Presentación del Conocimiento: El paso final implica el uso de técnicas de visualización y representación del conocimiento para presentar el conocimiento extraído de manera comprensible e ilustrativa a los usuarios finales.

2.4.2 .- ¿Qué tecnologías se utilizan?

La minería de datos, como un ámbito altamente centrado en la aplicación práctica, ha incorporado una diversidad de técnicas provenientes de áreas tales como estadística, aprendizaje automático, reconocimiento de patrones, sistemas de bases de datos y almacenes de datos, recuperación de información.

La naturaleza interdisciplinaria inherente a la investigación y desarrollo en minería de datos desempeña un papel crucial en el éxito de este campo y en su amplio impacto en diversas aplicaciones. En esta sección, se presentan ejemplos de disciplinas variadas que ejercen una influencia significativa en la evolución de los métodos utilizados en la minería de datos. Bajo la vista de (Kan, Kamber, Pei, 2012, p,23) se verán estas tecnologías que se usan en la minería de datos.

Estadística

La estadística, centrada en la recopilación, análisis, interpretación y presentación de datos, está intrínsecamente vinculada a la minería de datos. Los modelos estadísticos, que describen el comportamiento mediante variables aleatorias y distribuciones de probabilidad, son esenciales en tareas como la caracterización y clasificación en minería de datos. Estos modelos también se utilizan para abordar desafíos como el modelado de ruido y datos faltantes en conjuntos grandes, facilitando su identificación y gestión.

La investigación estadística contribuye con herramientas de predicción, mientras que la estadística inferencial se emplea para hacer inferencias sobre procesos o poblaciones, siendo útil para verificar resultados de la minería de datos a través de pruebas de hipótesis. La aplicación de métodos estadísticos en minería de datos enfrenta desafíos, especialmente con grandes conjuntos distribuidos, requiriendo adaptación para reducir el costo computacional, siendo aún más complejo en aplicaciones en línea con flujos de datos rápidos y en tiempo real.

Aprendizaje Automático

El aprendizaje automático se centra en cómo las computadoras pueden mejorar su rendimiento mediante el aprendizaje a partir de datos, especialmente en el reconocimiento de patrones y la toma de decisiones inteligentes. Existen diferentes tipos de aprendizaje automático: supervisado (clasificación), no supervisado (agrupación), semi supervisado (uso de ejemplos etiquetados y no etiquetados) y activo (permitiendo la participación activa del usuario). Estos enfoques comparten similitudes con la minería de datos, donde la investigación se centra en la precisión del modelo, la eficiencia en grandes conjuntos de datos, la escalabilidad y la exploración de métodos alternativos.

Sistemas de bases de datos y almacenes de datos

La investigación en sistemas de bases de datos se enfoca en la creación, mantenimiento y utilización eficiente de bases de datos para organizaciones y usuarios finales. Estos sistemas, reconocidos por su alta escalabilidad en el procesamiento de grandes conjuntos de datos estructurados, han establecido principios en modelos de datos, lenguajes de consulta, procesamiento y optimización de consultas, almacenamiento e índices.

La minería de datos se beneficia de tecnologías de bases de datos escalables para manejar eficientemente grandes conjuntos de datos. Las tareas de minería de datos pueden ampliar las capacidades de los sistemas de bases de datos existentes, satisfaciendo los requisitos de análisis de datos sofisticados. Los sistemas de bases de datos actuales han integrado capacidades analíticas mediante el uso de almacenes de datos y minería de datos. Estos almacenes integran datos de diversas fuentes y tiempos, creando cubos de datos que no solo facilitan el procesamiento analítico en bases multidimensionales, sino que también impulsan la minería de datos multidimensional.

Recuperación de información.

La recuperación de información (IR) se dedica a buscar documentos o información, ya sea en texto o multimedia, con enfoques basados en modelos probabilísticos. A diferencia de los sistemas de bases de datos, IR considera datos no estructurados y utiliza consultas principalmente basadas en palabras clave. Los documentos se modelan como conjuntos de palabras, y la similitud entre ellos se mide a través de la similitud entre sus modelos de lenguaje.

La modelación de temas en conjuntos de documentos se realiza mediante distribuciones de probabilidad, permitiendo identificar temas principales en colecciones de documentos. La combinación de IR, minería de datos y técnicas de

recuperación facilita la identificación de temas en grandes cantidades de datos textuales y multimedia disponibles en línea, siendo esencial en entornos como bibliotecas digitales, gobiernos digitales y sistemas de información de salud.

Procesamiento de lenguaje natural.

Como ya se mencionó anteriormente, el Procesamiento de Lenguaje Natural (PLN) es una rama de la inteligencia artificial que se enfoca en la interacción entre las computadoras y el lenguaje humano. Como tecnología esencial en la minería de datos, el PLN permite a las máquinas comprender, analizar y generar información a partir de datos basados en texto. Utilizando algoritmos y técnicas lingüísticas, el PLN facilita la extracción de patrones significativos y conocimientos valiosos en conjuntos de datos de texto, mejorando así la capacidad de análisis y toma de decisiones en el ámbito de la minería de datos.

2.4.3.- Metodologías para la minería de datos

En el intrincado mundo de la minería de datos, las metodologías son los mapas y brújulas que guían a los investigadores y analistas en su travesía de descubrimiento. Estas metodologías no solo ofrecen estructura y enfoque, sino que también desempeñan un papel vital en la eficacia y fiabilidad del proceso.

En este segmento, nos sumergimos en las profundidades de metodologías destacadas como KDD, CRISP-DM y SEMMA. Desde la comprensión inicial hasta la implementación y comunicación de resultados, se explorarán las fases fundamentales que componen estas metodologías, desentrañando los secretos de cómo transforman montañas de datos en pepitas de sabiduría. Prepárate para embarcarte en un viaje de descubrimiento guiado por las metodologías que definen la minería de datos contemporánea.

2.4.3.1.- Introducción a las Metodologías

Como afirmó (Lerma,2016,como se citó en Gonzáles,2017 p.61), una metodología es el "estudio del buen uso de los métodos y las técnicas. No es suficiente conocerlos, es necesario utilizarlos como se debe, es decir, saber adaptarlos lo más rigurosamente posible, por una parte, al objeto preciso de la investigación o del estudio considerado y, por otra parte, a los objetivos propuestos" . Esta definición subraya la esencia de las metodologías en la investigación y destaca su papel esencial.

En el vasto campo de la investigación, las metodologías ejercen un rol primordial al proporcionar una sólida estructura y un enfoque sistemático para el abordaje y la resolución de problemas. Estas metodologías se convierten en herramientas esenciales que permiten a los investigadores canalizar sus esfuerzos de manera efectiva, asegurando que los resultados obtenidos sean tanto confiables como significativos.

La elección de la metodología adecuada representa un pilar fundamental en todo proyecto de investigación. Esta elección conlleva la selección meticulosa de un enfoque específico que orientará la recopilación y el análisis de datos, así como la interpretación de los resultados. Al optar por una metodología concreta, los investigadores evitan adentrarse en la nebulosa de la ambigüedad y la arbitrariedad. En su lugar, abrazan un proceso estructurado que les brinda la confianza necesaria para avanzar de manera sistemática hacia la consecución de sus objetivos.

La elección sabia de una metodología no solo añade rigor a la investigación, sino que también permite a los investigadores obtener una comprensión más profunda de los fenómenos estudiados. Facilita la estandarización de procedimientos, garantizando que los datos se recolectan de manera uniforme y que los análisis se realicen de acuerdo con criterios establecidos. Esto, a su vez, otorga validez y solidez a los hallazgos, lo que es esencial en la generación de nuevo conocimiento y la toma de decisiones informadas.

En esta línea, explorar y comprender las metodologías utilizadas en un campo particular de investigación es esencial para cualquier investigador serio. En este contexto, dos metodologías ampliamente reconocidas, el Knowledge Discovery in Databases (KDD), o en español, Descubrimiento de Conocimiento en Bases de Datos, y la Metodología CRISP-DM, que son las siglas de Cross-Industry Standard Process for Data Mining (CRISP-DM), en español, Proceso Estándar Inter-Industrias para Minería de Datos, se alzan como enfoques clave en el ámbito de la minería de datos y el análisis de información.

En las secciones subsiguientes, examinaremos en detalle estos enfoques, resaltando sus respectivas contribuciones y aplicaciones en el panorama actual de la investigación. Las metodologías, en última instancia, no sólo impulsan la investigación, sino que también proporcionan la brújula que guía a los investigadores a través del intrincado laberinto del conocimiento.

2.4.3.2.- Fases de la metodología KDD

En el contexto de la minería de datos y el análisis de información, el proceso conocido como KDD, o Descubrimiento de Conocimiento en Bases de Datos,

desempeña un papel fundamental. En líneas generales, el KDD representa el proceso global de extraer conocimiento valioso a partir de grandes conjuntos de datos, mientras que la minería de datos se concentra en una etapa particular dentro de este proceso. Este enfoque se describe de manera destacada en el artículo de (Fayyad, Piatetsky-Shapiro y Smyth, 1996):

"KDD se refiere al proceso general de descubrimiento de conocimiento útil a partir de datos, mientras que la minería de datos se refiere a una fase específica en este proceso. La minería de datos implica la aplicación de algoritmos especializados para extraer patrones de datos. Además de la minería de datos, el proceso KDD incorpora etapas adicionales, como la preparación de datos, la selección de datos, la limpieza de datos, la incorporación de conocimiento previo apropiado y la interpretación adecuada de los resultados de la minería, todas fundamentales para asegurar que el conocimiento derivado de los datos sea efectivamente valioso. La aplicación ciega de métodos de extracción de datos puede ser una actividad peligrosa, que fácilmente conduce al descubrimiento de patrones inválidos y sin sentido."

Este proceso aborda un enfoque sistemático y riguroso para el análisis de datos, lo que lo convierte en un componente esencial en proyectos de minería de datos y análisis de información.

KDD ha experimentado un proceso de evolución constante y sigue avanzando gracias a la sinergia entre diversos campos de investigación. Estos campos incluyen el aprendizaje automático, el reconocimiento de patrones, las bases de datos, las estadísticas, la inteligencia artificial, la

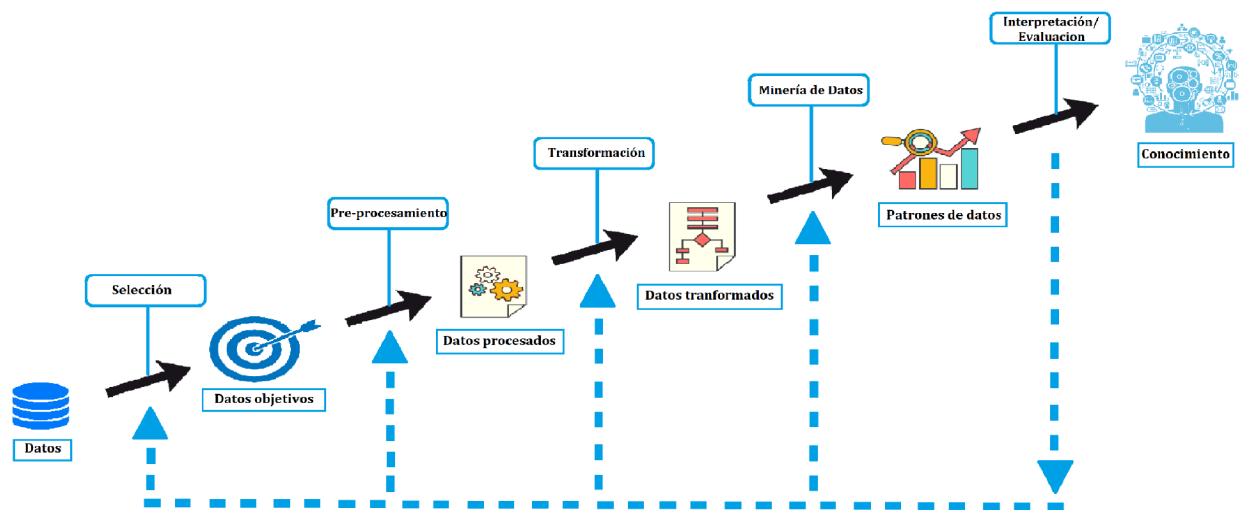
adquisición de conocimientos para sistemas expertos, la visualización de datos y la informática de alto rendimiento. Esta intersección de disciplinas ha permitido que KDD se enriquezca continuamente, adaptándose y aprovechando los avances en cada uno de estos dominios para mejorar su capacidad de descubrimiento de conocimiento a partir de datos.

Una pregunta natural es ¿En qué se diferencia KDD del reconocimiento de patrones o el aprendizaje automático y los otros campos relacionados? y es que según (Fayyad, Piatetsky-Shapiro, & Smyth, 1996, p.39):

"La respuesta es que estos campos proporcionan algunos de los métodos de extracción de datos que se utilizan en el paso de extracción de datos del proceso KDD. KDD se centra en el proceso general de descubrimiento de conocimiento a partir de datos, incluido cómo se almacenan y acceden a los datos, cómo se pueden escalar los algoritmos a conjuntos de datos masivos y aun así ejecutarse de manera eficiente, cómo se pueden interpretar y visualizar los resultados y cómo la relación hombre-máquina en general." Proceso el cual se puede observar en la figura 6.

Figura 6

Proceso de la metodología KDD



Nota. Elaboración propia.

El Descubrimiento de Conocimiento en Bases de Datos, conocido como KDD, es un proceso esencial en el campo de la minería de datos y el análisis de información. Se trata de una metodología interactiva y altamente iterativa que implica una serie de pasos cruciales. En este proceso, los investigadores y analistas trabajan en estrecha colaboración para extraer conocimiento valioso de grandes conjuntos de datos. A continuación, examinaremos en detalle los pasos fundamentales descritos en (Fayyad, et al., 1996, p. 42) que componen este proceso, cada uno de los cuales desempeña un papel crítico en la obtención de información significativa a partir de los datos.

Primero es desarrollar una comprensión del dominio de la aplicación y el conocimiento previo relevante e identificar el objetivo del proceso KDD desde el punto de vista del cliente.

El segundo es crear un conjunto de datos objetivo: seleccionar un conjunto de datos o centrarse en un subconjunto de variables o muestras de datos sobre las cuales se realizará el descubrimiento.

En tercer lugar, está la limpieza y el pre-procesamiento de datos. Las operaciones básicas incluyen eliminar el ruido, si corresponde, recopilar la información necesaria para modelar o tener en cuenta el ruido, decidir estrategias para manejar los campos de datos faltantes y tener en cuenta la información de secuencia temporal y los cambios conocidos.

El cuarto es la reducción y proyección de datos: encontrar características útiles para representar los datos según el objetivo de la tarea. Con métodos de transformación

o reducción de dimensionalidad, se puede reducir el número efectivo de variables consideradas o se pueden encontrar representaciones invariantes para los datos. El quinto es hacer coincidir los objetivos del proceso KDD (paso 1) con un método de extracción de datos particular. Por ejemplo, el resumen, la clasificación, la regresión, la agrupación, etc.

El sexto es el análisis exploratorio y la selección de modelos e hipótesis: elegir los algoritmos de extracción de datos y seleccionar los métodos que se utilizarán para buscar patrones de datos. Este proceso incluye decidir qué modelos y parámetros podrían ser apropiados (por ejemplo, los modelos de datos categóricos son diferentes a los modelos de vectores sobre reales) y hacer coincidir un método particular de extracción de datos con los criterios generales del proceso KDD (por ejemplo, el El usuario final podría estar más interesado en comprender el modelo que en sus capacidades predictivas).

El séptimo es la minería de datos: búsqueda de patrones de interés en una forma de representación particular o en un conjunto de tales representaciones, incluidas reglas o árboles de clasificación, regresión y agrupamiento. El usuario puede ayudar significativamente al método de extracción de datos si realiza correctamente los pasos anteriores.

El octavo es interpretar los patrones extraídos, posiblemente regresando a cualquiera de los pasos del 1 al 7 para realizar más iteraciones. Este paso también puede implicar la visualización de los patrones y modelos extraídos o la visualización de los datos dados los modelos extraídos.

El noveno es actuar sobre el conocimiento descubierto: utilizar el conocimiento directamente, incorporarlo a otro sistema para acciones futuras o simplemente documentar e informar a las partes interesadas. Este proceso también incluye verificar y resolver conflictos potenciales con conocimientos previamente creados (o extraídos).

2.4.3.3.- Fases de la metodología CRISP-DM

CRISP-DM es una metodología estándar utilizada en la minería de datos y el análisis de datos. Proporciona una estructura metodológica que guía a los profesionales a través de las distintas etapas del proceso de descubrimiento de conocimiento a partir de datos. A continuación, te proporciono una explicación e introducción a las principales etapas de CRISP-DM según (Rüdiger Wirth, Jochen Hipp, 2000):

Comprensión de negocio: Esta fase inicial se centra en comprender los objetivos y requisitos del proyecto desde una perspectiva empresarial y luego convertir este conocimiento en una definición del problema de minería de datos y un plan de proyecto preliminar diseñado para lograr los objetivos.

Comprensión de datos: La fase de comprensión de los datos comienza con una recopilación inicial de datos y continúa con actividades para familiarizarse con los datos, identificar problemas de calidad de los datos, descubrir las primeras ideas sobre los datos o detectar subconjuntos interesantes para formular hipótesis sobre información oculta. Existe un estrecho vínculo entre la comprensión empresarial y la comprensión de datos. La formulación del problema de la minería de datos y el plan del proyecto requieren al menos cierta comprensión de los datos disponibles.

Preparación de datos: La fase de preparación de datos cubre todas las actividades para construir el conjunto de datos final (datos que se introducirán en las herramientas de modelado) a partir de los datos iniciales sin procesar. Es probable que las tareas de preparación de datos se realicen varias veces y no en ningún orden prescrito. Las tareas incluyen selección de tablas, registros y atributos, limpieza de datos, construcción de nuevos atributos y transformación de datos para herramientas de modelado.

Modelado: En esta fase, se seleccionan y aplican varias técnicas de modelado y sus parámetros se calibran a valores óptimos. Normalmente, existen varias técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas requieren formatos de datos específicos. Existe un estrecho vínculo entre la preparación de datos y el modelado. A menudo, uno se da cuenta de problemas de datos mientras modela u obtiene ideas para construir nuevos datos.

Evaluación: En esta etapa del proyecto, ha creado uno o más modelos que parecen tener alta calidad, desde una perspectiva de análisis de datos. Antes de proceder a la implementación final del modelo, es importante evaluarlo más a fondo y revisar los pasos ejecutados para construirlo, para asegurarse de que logre adecuadamente los objetivos comerciales. Un objetivo clave es determinar si hay algún tema empresarial importante que no haya sido suficientemente considerado. Al final de esta fase, se debe tomar una decisión sobre el uso de los resultados de la minería de datos.

Despliegue: La creación del modelo generalmente no es el final del proyecto. Por lo general, los conocimientos adquiridos deberán organizarse y presentarse de manera que el cliente pueda utilizarlos. Dependiendo de los requisitos, la fase de implementación puede ser tan simple como generar un informe o tan compleja como implementar un proceso de minería de datos repetible. En muchos casos será el usuario, no el analista de datos, quien llevará a cabo los pasos de implementación.

En cualquier caso, es importante entender de antemano qué acciones serán necesarias para poder utilizar realmente los modelos creados.

2.4.3.4.- Fases de la metodología SEMMA

La metodología SEMMA (Sample, Explore, Modify, Model, Assess) fue propuesta por Statistical Analysis Systems Institute (SAS), el cual la define como el proceso de selección, exploración y modelado aplicado a cantidades significativas de datos almacenados que permitan el descubrimiento de patrones como herramientas de apoyo para el negocio. Según SAS más que una metodología para minería de datos, SEMMA es un conjunto de herramientas funcionales enfocadas hacia los aspectos de desarrollo propio de un modelo de minería.

SAS Institute define la metodología SEMMA como: El proceso de muestreo, exploración, modificación, modelado y evaluación (SEMMA) de grandes cantidades de datos para descubrir patrones previamente desconocidos que pueden utilizarse como una ventaja empresarial (Statistical Analysis Systems Institute [SAS], 2023).

SEMMA se compone de cinco fases representando las etapas de un proyecto de minería de datos las cuales según (Claudia L. Hernández G, María Ximena Dueñas R, 2009, p. 83) son:

Muestreo: En esta etapa se realiza la extracción de una muestra de los datos que permita representar características comunes de la población para posteriormente comenzar el análisis de los mismos. Con esta fase se logra facilitar los procesos de minado sobre los datos, reduciendo costes y tiempo para la organización.

Exploración: La exploración de datos a través de técnicas estadísticas permite realizar un seguimiento a los mismos logrando detectar, identificar y posteriormente eliminar datos que presenten anomalías o deficiencias en las fases siguientes hacia el descubrimiento de información.

Modificación: En esta fase se realiza una selección y transformación de los datos de acuerdo con las variables seleccionadas para el proceso de minado, la cual permitirá de acuerdo con éstas adaptar el enfoque de selección y diseño del modelo.

Modelado: En este punto de la metodología se hace uso de herramientas de software que permitan la utilización de técnicas y métodos propios de la minería de datos, las cuales tiendan hacia el descubrimiento de asociaciones o combinaciones entre los datos, logrando así la predicción de resultados con un alto nivel de confianza. Entre las técnicas más utilizadas para el modelado de datos, se

encuentran: métodos estadísticos, de agrupamiento, redes neuronales, árboles de decisión, lógica difusa, reglas de asociación, entre otros.

Evaluación: Uno de los pasos principales dentro de una metodología es la valoración de la solución. A partir del modelo obtenido en la fase anterior se realiza una evaluación de resultados para verificar el éxito del proyecto. Una buena práctica para comprobar la validez del modelo es seleccionar otra muestra de datos y aplicarlo para verificación de resultados, si este resulta óptimo se procede con el proceso de producción, en caso contrario se desarrollará otro modelo.

2.4.3.5.- Ventajas y limitaciones

En el vasto mundo de la investigación y la toma de decisiones, las metodologías desempeñan un papel central al proporcionar una estructura y un marco sistemático para abordar problemas y desafíos complejos. Estas metodologías, que abarcan desde enfoques científicos hasta técnicas prácticas, se convierten en herramientas esenciales para orientar las acciones y procesos, tanto en el ámbito académico como en el empresarial como en otros muchos más.

La elección de una metodología adecuada es un paso crucial en cualquier proyecto, ya que influirá en la calidad de los resultados y en la eficiencia de su consecución. Sin embargo, no existe una metodología universalmente perfecta; cada una tiene sus ventajas y limitaciones distintivas. Comprender estas ventajas y limitaciones es esencial para tomar decisiones informadas sobre cuál utilizar en un contexto particular.

En esta exploración sobre las ventajas y limitaciones de las metodologías, analizaremos tanto los beneficios que ofrecen como las restricciones que imponen en una variedad de campos, desde la investigación científica hasta el análisis de datos y la gestión de proyectos. Examinaremos cómo estas metodologías pueden aumentar la precisión y la confiabilidad de los resultados, al tiempo que podrían presentar desafíos en términos de flexibilidad, adaptabilidad y recursos necesarios.

Al comprender mejor las ventajas y limitaciones de las metodologías, los investigadores, profesionales y tomadores de decisiones podrán seleccionar y aplicar enfoques de manera más informada, maximizando así su capacidad para abordar problemas complejos y alcanzar objetivos con éxito. A lo largo de este análisis. En última instancia, la elección de una metodología adecuada y su aplicación cuidadosa son pasos esenciales en la búsqueda de resultados efectivos en proyectos de análisis de datos, minería de datos y ciencia de datos.

Por ejemplo, en una encuesta (KDnuggets, 2014) sobre metodologías utilizadas en proyectos de análisis de datos, se observa que la metodología CRISP-DM es ampliamente adoptada, seguida de cerca por otras como SEMMA y KDD en la quinta posición. Esta encuesta refleja la diversidad de enfoques que las personas eligen en función de sus necesidades y preferencias. Cada una de estas metodologías presenta sus propias fases, que van desde la comprensión del negocio hasta la implementación y la comunicación de resultados.

Cada fase desempeña un papel crucial en el proceso y puede influir significativamente en la calidad de los resultados obtenidos. Esta diversidad de opciones y la importancia de cada fase subrayan la necesidad de una cuidadosa consideración al elegir la metodología más adecuada para un proyecto específico. Como se ve reflejado en la siguiente tabla 5 sobre las fases de cada metodología.

Tabla 5

Fases de cada metodología de minería de datos

Fases	KDD	CRIPS-DM	SEMMA
Análisis y comprensión del negocio	Comprensión del dominio de la aplicación	Comprensión del negocio	
Selección y preparación de los datos	Crea el conjunto de datos limpieza y preprocesamiento de los datos, reducción y proyección de los datos	Entendimiento y preparación de los datos	Muestreo, comprensión y modificación
Modelado	Determinar la tarea de minería, el algoritmo y los datos que se desean minar	Modelado de los datos	Modelado de los datos
Evaluación	Interpretación	Evaluación	Valoración
Implementación	Utilización del nuevo conocimiento	Despliegue	

Nota. Elaboración propia.

En resumidas cuentas, la elección de la metodología dependerá de una variedad de factores, siendo uno de ellos el nivel de conocimiento que se posea sobre cada

metodología y su idoneidad para el proyecto en cuestión. Además, no podemos pasar por alto la naturaleza de los datos que se van a utilizar, ya que esto influirá en gran medida en la eficacia de la metodología seleccionada. En última instancia, tomar una decisión informada teniendo en cuenta estos factores clave es esencial para el éxito de cualquier proyecto de análisis de datos y minería de información.

2.4.3.6.- Consideraciones sobre las metodologías de minería de datos

La metodología de minería de datos, representada por KDD (Descubrimiento de Conocimiento en Bases de Datos), CRISP-DM (Metodología Estándar para Minería de Datos Cruzada), y SEMMA (Sample, Explore, Modify, Model, Assess), desempeña un papel crucial en la extracción de conocimiento valioso de grandes conjuntos de datos. Estas metodologías proporcionan estructuras sistemáticas que guían a los profesionales a lo largo de diferentes etapas, asegurando un proceso riguroso y coherente.

En el caso de KDD, su enfoque sistemático abarca desde la comprensión del dominio hasta la acción sobre el conocimiento descubierto.(Fayyad et al.1996) enfatizan la importancia de no aplicar métodos de extracción de datos de manera ciega, subrayando la necesidad de considerar la preparación, selección y limpieza de datos, así como la interpretación adecuada de los resultados. KDD se enriquece continuamente gracias a la sinergia entre diversas disciplinas, incluyendo aprendizaje automático, reconocimiento de patrones y estadísticas.

CRISP-DM, una metodología estándar, se organiza en fases clave que van desde la comprensión del negocio hasta el despliegue del modelo. Este enfoque garantiza que los objetivos comerciales estén alineados con el proceso de minería de datos, abordando la comprensión de negocio, la preparación de datos, el modelado y la evaluación de manera secuencial.

SEMMA, propuesto por SAS, destaca por sus cinco fases: muestreo, exploración, modificación, modelado y evaluación. Cada fase cumple una función específica, desde la extracción de una muestra representativa hasta la evaluación del modelo construido. La metodología SEMMA se presenta como un conjunto de herramientas funcionales que se centran en el desarrollo propio de un modelo de minería.

Al explorar las ventajas y limitaciones de estas metodologías, se reconoce que no existe una metodología universalmente perfecta. Cada una tiene sus propias características, y su elección dependerá de factores como el contexto del proyecto, la naturaleza de los datos y el nivel de conocimiento sobre cada metodología. La diversidad de enfoques, como evidencia la encuesta de KDnuggets, refleja la

necesidad de una consideración cuidadosa al seleccionar la metodología más adecuada para un proyecto específico.

En última instancia, la elección informada de una metodología y su aplicación cuidadosa son esenciales para alcanzar resultados efectivos en proyectos de análisis de datos, minería de datos y ciencia de datos. Cada fase desempeña un papel crucial, y la diversidad de opciones subraya la importancia de considerar los factores clave para garantizar el éxito del proyecto.

2.4.4.- Fases de la minería de datos

La minería de datos, más que una simple expedición, es un viaje estructurado a través de diversas fases cruciales. En este segmento, se dará a conocer cada etapa del proceso, desde el desarrollo de una comprensión profunda del dominio hasta la interpretación y aplicación del conocimiento extraído. Cada fase, meticulosamente diseñada, contribuye al éxito general del proyecto al abordar aspectos como la preparación de datos, la selección de modelos y la evaluación de resultados. Únete a nosotros mientras navegamos por estas fases, revelando cómo transformar datos en valiosas percepciones a lo largo de un viaje que redefine la forma en que interactuamos con la información.

2.4.4.1.- Selección de Datos

Antes de sumergirnos en las complejidades de la minería de datos, resulta imperativo realizar una exhaustiva preparación de los datos. Este paso crítico conlleva una inspección detallada de los atributos y valores presentes en los conjuntos de datos del mundo real, los cuales, con frecuencia, son ruidosos y de dimensiones considerables, derivando de una amalgama de fuentes heterogéneas. Este escenario subraya la necesidad de adquirir una comprensión profunda antes de emprender la extracción de conocimientos. En este contexto, el presente capítulo se enfoca en el proceso de familiarización con los datos, un requisito fundamental para la tarea inicial y crucial de preprocesamiento de datos.

Bajo la referencia del libro "Data Mining Concepts and Techniques" (Han, Kamber, Pei, 2012, p. 40), se definen diversos atributos que establecen las bases conceptuales para abordar la selección de datos.

2.4.4.1.1.- Sistemas de recolección de información de las redes sociales

Las redes sociales han emergido como fuentes inagotables de información en la era digital, proporcionando datos valiosos sobre las interacciones y opiniones de los usuarios. En este contexto, los sistemas de recolección de información juegan un

papel crucial al permitir la extracción efectiva de datos relevantes de diversas plataformas sociales.

Este capítulo explorará a fondo estos sistemas, centrándose en su importancia, técnicas de recolección de datos y las herramientas clave utilizadas en este proceso. Desde el web scraping hasta el uso de APIs y herramientas de escucha social, analizaremos las distintas metodologías que potencian la recopilación de información en el vasto paisaje de las redes sociales.

¿Qué son los sistemas de recolección de información?

Los sistemas de recolección de información de las redes sociales juegan un papel fundamental en la era actual, donde la generación de información en estas plataformas es abundante y valiosa para llevar a cabo análisis de big data. Estas herramientas y técnicas posibilitan y dan facilidad a la recopilación de datos relevantes de diversas plataformas de redes sociales, como Facebook, Twitter, Instagram, YouTube, LinkedIn y otros sitios web similares. Al utilizar estos sistemas, es posible obtener una visión más profunda de las interacciones, opiniones y comportamientos de los usuarios en tiempo real, lo que brinda una base sólida para el análisis y la toma de decisiones estratégicas.

Técnicas para la recolección de datos.

El web scraping es uno de los métodos más utilizados en los sistemas de recolección de información. Esta técnica consiste en extraer datos de sitios web mediante el uso de bots o programas automatizados. Con el web scraping, es posible obtener información precisa y específica de perfiles de usuarios, publicaciones, comentarios, seguidores y otros datos relevantes de las redes sociales.

El web scraping se destaca por su eficiencia y capacidad para recopilar grandes volúmenes de datos de manera rápida y automatizada. Al automatizar el proceso de extracción de datos, se ahorra tiempo y esfuerzo en comparación con la recopilación manual. Además, el web scraping permite obtener datos estructurados y listos para su análisis, lo que facilita la posterior aplicación de técnicas de análisis de datos y minería de texto.

Esta técnica es especialmente valiosa para la investigación de mercado, la monitorización de la competencia, la generación de bases de datos y otros análisis de big data. Al recopilar datos relevantes de las redes sociales a gran escala, se obtiene una visión más amplia y precisa de las tendencias, preferencias y comportamientos de los usuarios.

En el contexto del web scraping ético, se deben tener en consideración las siguientes pautas (Theo Vasilis, 2022):

Evitar sobrecargar el sitio web de destino: Es importante realizar el web scraping de manera responsable y evitar sobrecargar o impactar negativamente en el rendimiento del sitio web objetivo. Esto implica establecer un ritmo adecuado de extracción de datos para no causar interrupciones ni afectar la experiencia de otros usuarios.

Respetar los derechos de autor y políticas de uso: Al realizar web scraping, es fundamental reconocer y respetar los derechos de autor y las políticas de uso establecidas por los propietarios del sitio web y los creadores de los datos. Es necesario cumplir con los términos de servicio y las restricciones legales al acceder y utilizar la información recopilada.

Fomentar la web abierta: Se debe promover y respetar los principios de la web abierta, que incluyen el acceso libre a la información pública. Al realizar web scraping, es esencial utilizar esta técnica de manera responsable y ética, evitando violar la privacidad de las personas y respetando los límites establecidos por los sitios web que se visitan.

Evitar el monopolio de los datos: Es importante evitar el uso desleal o monopolístico de los datos extraídos mediante web scraping. No se debe utilizar la información recopilada para obtener ventajas injustas o perjudicar a otros competidores o usuarios. En su lugar, se debe promover la competencia justa y el acceso equitativo a los datos en beneficio de la comunidad en general.

Sin embargo, es fundamental tener en cuenta que el web scraping ético también implica consideraciones legales y éticas más amplias. Esto implica respetar los términos de servicio de los sitios web y obtener el permiso adecuado cuando sea necesario. Además, se deben seguir las leyes de protección de datos y derechos de autor para garantizar un uso responsable de la información obtenida.

Herramientas para la recolección de datos.

Existen diversas herramientas utilizadas en web scraping que pueden facilitar el proceso de extracción de datos. A continuación, se mencionan algunas de ellas:

Selenium: Es una herramienta que permite interactuar con páginas web, automatizando acciones como abrir una página, acceder a datos específicos, llenar formularios, hacer clic en elementos, cambiar de pestaña, entre otros. Es compatible con XML y lenguaje XPATH, lo que la hace especialmente útil para web scraping dinámico.

Requests: Esta librería en Python facilita el acceso a cualquier URL, ya sea que requiera autenticación o no. Es comúnmente utilizada para extraer contenido de páginas web y para descargar archivos que contienen enlaces específicos de descarga.

Scrapy: Es una de las librerías más populares para realizar web scraping. Proporciona una estructura y funcionalidades sólidas para hacer solicitudes al servidor y analizar la información obtenida, lo que permite la recolección eficiente de datos. Scrapy es especialmente útil para realizar crawling vertical y horizontal, recorriendo múltiples páginas relacionadas mediante paginación.

Beautiful Soup: Esta librería en Python se utiliza para analizar y recorrer el código HTML completo de una página web, permitiendo extraer información, valores e imágenes deseadas. Proporciona métodos sencillos para la manipulación de los elementos del DOM.

Mechanize: Es una librería utilizada principalmente para automatizar el llenado y envío de formularios, activar el uso de cookies y realizar acciones que podrían interrumpir la automatización de un proceso de extracción de datos. También permite navegar por la web al proporcionar funciones para obtener todos los enlaces y formularios presentes en una página.

Por otro lado, las APIs de redes sociales brindan la posibilidad de acceder y extraer datos de manera programática de diversas plataformas de redes sociales. Estas interfaces de programación de aplicaciones permiten obtener información estructurada y en tiempo real, que incluye datos demográficos, interacciones de usuarios, estadísticas de publicaciones y mucho más.

Al utilizar las APIs de redes sociales, los desarrolladores pueden acceder a datos específicos de perfiles, publicaciones, comentarios, seguidores y otras métricas relevantes. Esto les permite analizar tendencias, realizar investigaciones de mercado, identificar patrones de comportamiento y tomar decisiones informadas.

Las APIs de redes sociales también permiten realizar consultas avanzadas, como la búsqueda de contenido basada en palabras clave, hashtags o ubicaciones geográficas. Esto brinda a los usuarios la capacidad de personalizar y refinar la información que desean extraer de las redes sociales.

Además, al utilizar las APIs, los datos obtenidos se actualizan en tiempo real, lo que permite realizar un seguimiento continuo de las interacciones y métricas de las redes sociales. Esto es especialmente valioso para empresas y profesionales del marketing que desean mantenerse al tanto de las tendencias y la actividad en las redes sociales.

Algunos ejemplos de APIs de redes sociales son los siguientes:

Twitter API: Permite acceder a una amplia variedad de datos relacionados con Twitter, como tweets, perfiles de usuario, tendencias y más. También proporciona funcionalidades para realizar acciones en la plataforma, como publicar tweets, seguir usuarios y buscar contenido específico.

Facebook Graph API: Brinda acceso a datos y funcionalidades relacionadas con la plataforma de Facebook. Esto incluye perfiles de usuario, páginas, publicaciones, comentarios, eventos y más. Con esta API, es posible obtener información detallada sobre las interacciones de los usuarios y realizar acciones específicas, como publicar en perfiles o páginas.

Instagram API: Permite acceder a una variedad de datos relacionados con Instagram, como información de perfiles, publicaciones, seguidores, comentarios y más. Además de obtener datos, la API permite interactuar con la plataforma, como dar "me gusta" a publicaciones, seguir usuarios y realizar búsquedas basadas en hashtags y ubicaciones.

Por último, tenemos Las herramientas de escucha social son sistemas que permiten monitorear y recopilar información de las redes sociales de manera efectiva. Estas herramientas utilizan diversos métodos para recopilar datos relevantes, como la búsqueda de palabras clave, hashtags, menciones y comentarios. Algunas de las funcionalidades y características específicas que se pueden encontrar en estas herramientas son:

Búsqueda de palabras clave: Estas herramientas permiten realizar búsquedas específicas basadas en palabras clave relevantes para tu empresa o marca. Puedes rastrear menciones de tu marca, productos o servicios, así como también identificar temas y tendencias relacionadas con tu industria.

Seguimiento de hashtags: Las herramientas de escucha social te permiten realizar un seguimiento de hashtags relevantes en las redes sociales. Esto es útil para identificar conversaciones específicas en torno a un tema determinado, participar en discusiones relevantes y comprender mejor las tendencias en tu industria.

Monitoreo de menciones y comentarios: Estas herramientas te brindan la capacidad de monitorear y analizar las menciones y comentarios que recibes en las redes sociales. Esto te permite estar al tanto de las interacciones de los usuarios, responder a preguntas o comentarios y gestionar la reputación en línea de tu marca. Algunos ejemplos son:

Hootsuite: Hootsuite es una herramienta popular de gestión de redes sociales que también ofrece funciones de escucha social. Permite realizar búsquedas basadas en

palabras clave, hashtags y ubicaciones para monitorear las conversaciones relevantes en las redes sociales. Además, ofrece análisis y reportes detallados.

Brandwatch: Es una plataforma de escucha social que te permite monitorear y analizar menciones, comentarios y conversaciones en las redes sociales. Ofrece características avanzadas de filtrado y segmentación para obtener información precisa sobre tu marca y competidores.

Sprout Social: Es una herramienta de gestión de redes sociales que incluye funciones de escucha social. Puedes realizar búsquedas de palabras clave, hashtags y menciones para identificar las conversaciones relevantes. También te permite responder a los comentarios directamente desde la plataforma.

Apify: Es una plataforma que combina web scraping y automatización para extraer datos de las redes sociales y otras fuentes web. Con Apify, puedes crear "actores" personalizados para recopilar información específica de las redes sociales, como perfiles de usuario, publicaciones, comentarios y más. También puedes programar tareas de extracción periódicas y recibir los datos recopilados en diferentes formatos.

2.4.4.2.- Preprocesamiento de Datos

Las bases de datos del mundo real, dada su envergadura típicamente monumental que a menudo almacenan grandes cantidades de gigabytes y su probable origen diversa de fuentes heterogéneas, son altamente propensas a la presencia de datos ruidosos, faltantes e inconsistentes. La calidad de los resultados en minería de datos se ve directamente afectada por la calidad de los datos subyacentes. Surge entonces la interrogante: ¿Cómo se pueden preprocesar los datos para mejorar tanto la calidad de los datos como, en consecuencia, los resultados de la minería? ¿De qué manera puede optimizarse el proceso de minería mediante un preprocesamiento eficaz?.

Diversas técnicas de preprocesamiento de datos están disponibles para abordar estas interrogantes. La limpieza de datos se erige como una estrategia clave para eliminar el ruido y corregir inconsistencias en los conjuntos de datos. La integración de datos, por su parte, amalgama información proveniente de múltiples fuentes en un almacén de datos coherente, como un data warehouse. La reducción de datos es una técnica que reduce el tamaño de los datos mediante la agregación, eliminación de funciones redundantes o agrupación.

Además, las transformaciones de datos, como la normalización, escalan los datos para situarlos dentro de un rango más manejable, por ejemplo, de 0.0 a 1.0. Este ajuste contribuye a mejorar la precisión y eficiencia de los algoritmos de minería que emplean mediciones de distancia. Importante destacar que estas técnicas no son

mutuamente excluyentes; más bien, pueden trabajar en conjunto de manera sinérgica.

Un ejemplo ilustrativo pueden ser los siguientes:

Limpieza de datos: puede implicar transformaciones para corregir datos erróneos, como la conversión de todas las entradas de un campo de fecha a un formato común.

Eliminación de Valores Atípicos: La limpieza de datos puede incluir la identificación y eliminación de valores atípicos que podrían distorsionar los resultados. Por ejemplo, en un conjunto de datos de ingresos, la presencia de un valor extremadamente alto podría ser un error o una entrada inusual que afecta negativamente a los análisis.

Manejo de Datos Faltantes: Otra faceta crucial de la limpieza de datos implica tratar con valores faltantes. Puedes optar por eliminar registros con datos faltantes, estimar los valores ausentes a través de técnicas de imputación o incluso utilizar algoritmos más avanzados para predecir y completar esos valores.

Corrección de Errores de Escritura: A veces, los datos pueden contener errores tipográficos o de escritura. La limpieza de datos podría implicar la estandarización de entradas, como corregir nombres mal escritos o normalizar las categorías para garantizar consistencia.

Unificación de Formatos de Texto: La consistencia en los formatos de texto es esencial. La limpieza de datos puede abordar la unificación de formatos, como estandarizar el uso de mayúsculas y minúsculas en categorías o corregir abreviaturas para garantizar uniformidad en la representación de los datos.

Normalización de Datos Numéricos: Cuando se trabaja con atributos numéricos, la limpieza podría involucrar la normalización para llevar todos los datos a una escala común. Esto es especialmente útil en algoritmos sensibles a las diferencias de magnitud entre atributos.

Validación de Datos con Reglas de Negocio: La limpieza de datos también puede abordar la validación según reglas de negocio específicas. Por ejemplo, en un conjunto de datos de ventas, podrías verificar que las fechas de entrega no preceden a las fechas de pedido, identificando y corrigiendo posibles inconsistencias.

Conversión de Unidades: En conjuntos de datos que involucran medidas, la limpieza podría incluir la conversión de unidades para garantizar coherencia. Por ejemplo, convertir todas las medidas de peso a la misma unidad (kilogramos o libras).

Antes de abordar la pregunta fundamental sobre la necesidad de preprocesar los datos, resulta esencial dirigir la mirada hacia los planteamientos de (Han, Kamber, Pei, 2012, p.84), quienes resaltan la importancia de identificar problemáticas recurrentes al analizar conjuntos de datos del mundo real. En este contexto, los datos exhiben notorias características problemáticas, tales como inexactitudes, incompletitudes e inconsistencias. Se evidencian escenarios donde varios atributos carecen de valores, y se detectan errores, valores inusuales e inconsistencias en las transacciones registradas.

La calidad deficiente de los datos, marcada por la presencia de inexactitudes, incompletitudes e inconsistencias, subraya la relevancia ineludible del preprocesamiento de datos. Este paso crítico tiene como objetivo abordar y corregir estas problemáticas antes de aplicar las técnicas de minería de datos. Entre las razones subyacentes que impulsan la necesidad de preprocesamiento se encuentran errores humanos o informáticos, la presencia de datos faltantes encubiertos, limitaciones tecnológicas, inconsistencias en las convenciones de nomenclatura y formatos inconsistentes, entre otros desafíos inherentes.

En última instancia, el preprocesamiento de datos emerge como un componente esencial para asegurar la confiabilidad y significado de los resultados en diversos contextos y aplicaciones de minería de datos.

2.4.4.2.1.-Sistemas para la limpieza de datos

En el entorno del análisis de datos, la calidad y la integridad de los conjuntos de datos desempeñan un papel esencial en la obtención de resultados precisos y confiables. La presencia de errores, inconsistencias y valores atípicos puede afectar significativamente la validez de las conclusiones derivadas del análisis. Por esta razón, la limpieza de datos emerge como un componente crítico en el proceso de preparación de datos para su posterior análisis.

En esta sección, se explorarán los sistemas diseñados específicamente para la limpieza de datos. Desde la identificación y manejo de valores nulos hasta la corrección de errores tipográficos, estos sistemas ofrecen herramientas y técnicas para abordar los desafíos asociados con la calidad y la consistencia de los datos. Examinaremos detalladamente estas soluciones, destacando su relevancia en la mejora de la integridad de los datos utilizados en proyectos de análisis de gran envergadura.

¿Qué es la limpieza de datos?

La limpieza de datos es un proceso esencial en el ámbito del análisis de datos, donde la calidad y la integridad de la información son cruciales para obtener

resultados precisos y confiables. Este proceso se centra en identificar y corregir errores, inconsistencias y valores atípicos en los conjuntos de datos, garantizando que la información utilizada en el análisis sea precisa y coherente.

Uno de los conceptos fundamentales en la limpieza de datos es la identificación de valores nulos o faltantes. Estos pueden surgir por diversas razones, como errores en la recopilación de datos o la ausencia de información. La tarea principal es decidir cómo abordar estos valores, ya sea eliminando los, imputando valores basados en patrones o utilizando métodos más avanzados.

Según en un artículo de Amazon titulado "¿En qué consiste la limpieza de los datos?"(Amazon, sf) definió lo que es la limpieza de datos: "La limpieza de datos es un proceso esencial para preparar los datos sin procesar para las aplicaciones de machine learning (ML) y de inteligencia empresarial (BI). Es posible que los datos sin procesar contengan numerosos errores, que pueden afectar a la precisión de los modelos de ML y dar lugar a predicciones incorrectas y a un impacto empresarial negativo. "

Por otro lado,(Chapman, 2005) define la limpieza de datos como "Un proceso utilizado para determinar datos inexactos, incompletos o irrazonables y luego mejorar la calidad mediante la corrección de errores y omisiones detectados."

Los científicos de datos destacaron la preparación de datos como una de las actividades que más tiempo consume. En el artículo "A glimpse inside the mind of a data scientist"(IBM, 2016), se realizó la siguiente pregunta a varios científicos de datos para conocer cuáles eran los mayores obstáculos a la hora de llevar a cabo un análisis de big data ¿Qué tareas consumen la mayor parte de tu tiempo?:

Por un lado (Jack Burgess,2016), desarrollador de herramientas de operaciones de seguridad, Telstra dice lo siguiente: "La preparación de datos tiende a ser una de mis actividades que consumen más tiempo. Es una tarea extremadamente crítica, pero pesada. También deja menos tiempo para analizar los datos y ofrecer nuevos conocimientos a los responsables de la toma de decisiones."

Por otro lado, (Michael Schmidt, 2016), científico de datos y fundador de Nutonian, dice lo siguiente:"Obtener acceso a los datos y ponerlos en el formato adecuado puede llevar mucho tiempo. Los problemas de disponibilidad pueden variar desde no tener suficiente volumen o variedad de datos hasta tener datos extremadamente inconsistentes o "sucios", donde el esfuerzo de limpieza, filtrado o reparación es tan monumental que aumenta el riesgo más allá de lo tolerable".

El informe "Data Science Report" de (CrownFlower,2016) revela que, contrariamente a la percepción popular, la mayor parte del tiempo de los científicos de datos se dedica a limpiar y organizar datos. Este proceso, conocido como "disputa de datos",

abarca tareas desde la verificación de listas hasta la depuración de bases de datos. Cerca del 60% de los científicos de datos encuestados expresaron que dedicaban demasiado tiempo a esta fase del trabajo, destacando la considerable importancia de abordar datos desordenados en el flujo de trabajo típico del científico de datos.

Figura 6

Labores que llevan más tiempo en un análisis big data.



Nota. (CrownFlower,2016).

La limpieza de datos es un componente esencial en el proceso de preparación de datos para el análisis, y diversas herramientas y técnicas se utilizan para abordar los desafíos asociados con la calidad y la consistencia de los conjuntos de datos. Estas herramientas y técnicas buscan identificar y corregir errores, inconsistencias y valores atípicos que podrían afectar la integridad y la precisión de los datos utilizados en el análisis.

Transformación de Datos.

Esta sección aborda métodos clave de transformación de datos, un paso esencial en el preprocesamiento para mejorar la eficiencia del proceso de minería y facilitar la comprensión de los patrones identificados. Además, se explora la discretización de datos como una forma específica de transformación.

Estrategias de Transformación de Datos:

En la transformación de datos, se aplican cambios para consolidar la información de manera adecuada para la minería. Las estrategias comunes para la transformación de datos descritos por (Han, Kamber, Pei, 2012, p.111) abarcan:

Suavizado: Orientado a eliminar el ruido presente en los datos, emplea técnicas como agrupación, regresión y agrupación.

Construcción de Atributos: Se centra en la creación y adición de nuevos atributos a partir del conjunto existente, fortaleciendo el proceso de minería.

Agregación: Implica la aplicación de operaciones de resumen o agregación a los datos. Por ejemplo, los datos de ventas diarias pueden consolidarse para calcular totales mensuales y anuales. Este paso es esencial para construir un cubo de datos en análisis de datos de varios niveles de abstracción.

Normalización: Escala los datos de atributos para que se sitúen dentro de un rango más pequeño, como -1,0 a 1,0 o 0,0 a 1,0.

Discretización: Sustituye los valores brutos de atributos numéricos (por ejemplo, edad) con etiquetas de intervalo (por ejemplo, 0–10, 11–20, etc.) o conceptuales (por ejemplo, joven, adulto, mayor). Estas etiquetas pueden organizarse recursivamente en conceptos de nivel superior, generando así una jerarquía de conceptos para el atributo numérico.

Generación de Jerarquía de Conceptos para Datos Nominales: Permite la generalización de atributos nominales, como la calle, hacia conceptos de nivel superior, por ejemplo, ciudad o país. Muchas jerarquías de atributos nominales pueden definirse automáticamente en el nivel de definición del esquema de la base de datos.

Herramientas para la limpieza y normalización de datos.

Las herramientas de limpieza de datos varían desde simples funciones de software hasta plataformas especializadas que ofrecen soluciones más avanzadas. Algunas de las técnicas comunes incluyen la identificación y manejo de valores nulos, la corrección de errores tipográficos, la estandarización de formatos, y la detección y tratamiento de valores atípicos.

Pandas: Una Biblioteca Fundamental para el Análisis de Datos en Python

La biblioteca Pandas, en desarrollo desde 2008, surge con el objetivo de cerrar la brecha entre las herramientas de análisis de datos disponibles en Python, un lenguaje de computación científica y sistemas de propósito general, y las numerosas plataformas de computación estadística y lenguajes de bases de datos de dominios específicos.

Más allá de proporcionar funcionalidades equivalentes, Pandas se destaca por implementar características únicas, como la alineación automática de datos y la indexación jerárquica, que no se encuentran tan estrechamente integradas en otras bibliotecas o entornos informáticos. Inicialmente concebida para aplicaciones de análisis de datos financieros, Pandas aspira a hacer de Python un entorno de

computación estadística más atractivo y práctico tanto para profesionales académicos como de la industria.

El nombre de la biblioteca, Pandas, proviene de "datos de panel", un término común en estadística y econometría para conjuntos de datos multidimensionales. Esta elección refleja la versatilidad de Pandas al abordar conjuntos de datos complejos y variados.

El artículo de (McKinney, 2011) destaca que Pandas no solo ofrece una funcionalidad equivalente, sino que también introduce características distintivas, marcando así su contribución única al panorama de las herramientas de análisis de datos. Además, el autor subraya que el desarrollo de Pandas no se limitó solo a aplicaciones financieras; se espera que proporcione un entorno de computación estadística más accesible para profesionales académicos e industriales.

Aunque esta descripción proporciona una visión general de algunas características destacadas de Pandas, es importante tener en cuenta que el artículo de McKinney no agota todas las funcionalidades de la biblioteca. Pandas ofrece una amplia gama de capacidades para el análisis de datos, convirtiéndola en una herramienta esencial en el arsenal de cualquier científico de datos o analista.

NumPy: La Piedra Angular para la Computación Científica en Python

NumPy (NumPy, 2023), una biblioteca fundamental para la computación científica en Python, ha desempeñado un papel crucial desde su desarrollo. La biblioteca tiene como objetivo principal proporcionar un soporte eficiente para operaciones en arreglos multidimensionales y matrices, brindando a Python la capacidad de competir con lenguajes especializados en ámbitos científicos y de ingeniería.

Según la documentación de NumPy, la biblioteca ofrece:

- **Funciones Matemáticas Eficientes:** NumPy proporciona funciones matemáticas rápidas y eficientes, lo que la convierte en una elección predilecta para la manipulación de datos numéricos en Python.
- **Operaciones con Arreglos:** NumPy permite realizar operaciones matemáticas en matrices enteras sin la necesidad de utilizar bucles, mejorando la eficiencia y velocidad del código.
- **Manipulación de Datos:** Facilita la manipulación de datos, incluida la transposición, indexación, corte y concatenación, simplificando las tareas de limpieza y preparación de datos.

La documentación oficial de NumPy proporciona una guía detallada de sus capacidades y funciones. El énfasis en la eficiencia y la capacidad de trabajar con grandes conjuntos de datos hace de NumPy una herramienta esencial para científicos de datos, ingenieros y profesionales en campos relacionados.

SciPy: Herramientas Científicas Avanzadas para Python

SciPy, una extensión robusta construida sobre NumPy, eleva las capacidades de Python en el ámbito científico e ingenieril. Al proporcionar bibliotecas adicionales, SciPy se convierte en una herramienta esencial que abarca optimización, álgebra lineal, integración, interpolación, funciones especiales y más.

Esta biblioteca no solo complementa NumPy, sino que también amplía significativamente las funcionalidades disponibles, convirtiéndose en una opción indispensable para profesionales que buscan soluciones avanzadas en sus investigaciones y proyectos (Virtanen, Gommers, Oliphant, et al, 2020).

scikit-learn: Aprendizaje de Máquina en Python

scikit-learn, también conocido como sklearn, emerge como una biblioteca integral para el aprendizaje de máquina en Python. Diseñada para ser simple y eficiente, ofrece herramientas esenciales para la minería y análisis de datos, así como para la construcción, evaluación e implementación de modelos de aprendizaje automático.

Entre sus características destacadas descritas en “Scikit-learn: Machine Learning in Python” (Pedregosa, Varoquaux, Gramfort, A., et al, 2011) se incluyen:

Variedad de Algoritmos: scikit-learn integra una amplia gama de algoritmos de aprendizaje supervisado y no supervisado, desde máquinas de soporte vectorial hasta bosques aleatorios, proporcionando flexibilidad para adaptarse a diversos escenarios.

Herramientas para Evaluación: Ofrece funciones para evaluar el rendimiento de modelos, incluyendo métricas de precisión, recall, y curvas ROC, facilitando la selección del modelo más adecuado.

Preprocesamiento de Datos: scikit-learn incluye herramientas para el preprocesamiento de datos, como la normalización y la reducción de dimensionalidad, optimizando la preparación de datos para el análisis.

Además, la comunidad activa y la documentación detallada hacen de scikit-learn una elección confiable para aquellos que buscan implementar soluciones de aprendizaje de máquina con Python.

NLTK (Natural Language Toolkit): Herramienta para el Procesamiento del Lenguaje Natural

NLTK es una biblioteca en Python diseñada para trabajar con el procesamiento del lenguaje natural (NLP). Su objetivo principal es proporcionar herramientas y recursos para el análisis de texto y la construcción de aplicaciones basadas en NLP. Desarrollado por Steven Bird y Edward Loper, NLTK se ha convertido en una opción esencial para investigadores y desarrolladores en el campo del procesamiento del lenguaje natural.

Funcionalidades Clave destacadas en el libro "Natural Language Processing with Python"(Bird, Klein, Loper, 2009, p.15).

Simplicidad: Proporcionar un marco intuitivo junto con componentes básicos sustanciales, dando usuarios un conocimiento práctico de PNL sin empantanarse en tediosos limpieza generalmente asociada con el procesamiento de datos lingüísticos anotados

Consistencia: Proporcionar un marco uniforme con interfaces y estructuras de datos consistentes y nombres de métodos fácilmente adivinados.

Extensibilidad: Proporcionar una estructura en la que se puedan acomodar fácilmente nuevos módulos de software, incluidas implementaciones alternativas y enfoques competitivos para la misma tarea.

Modularidad: Proporcionar componentes que se puedan utilizar de forma independiente sin necesidad de comprender el resto del conjunto de herramientas.

2.4.4.3.- Minería de Datos

Introducción a las Técnicas de Minería de Datos:

La minería de datos, como componente esencial del proceso analítico, despliega diversas técnicas para revelar patrones, tendencias y relaciones ocultas en conjuntos de datos. La importancia de esta fase radica en su capacidad para transformar datos crudos en información valiosa y, en última instancia, en conocimientos prácticos.

Importancia de la Minería de Datos:

Descubrimiento de Patrones:

Permite identificar relaciones complejas y patrones no inmediatamente evidentes en los datos, proporcionando una comprensión más profunda de los fenómenos estudiados.

Toma de Decisiones Informada:

Facilita la toma de decisiones basada en evidencia al extraer información relevante, mejorando la eficiencia y precisión en la toma de decisiones.

Optimización de Procesos:

Permite identificar áreas de mejora en procesos empresariales y operativos mediante el análisis de datos, posibilitando la optimización continua.

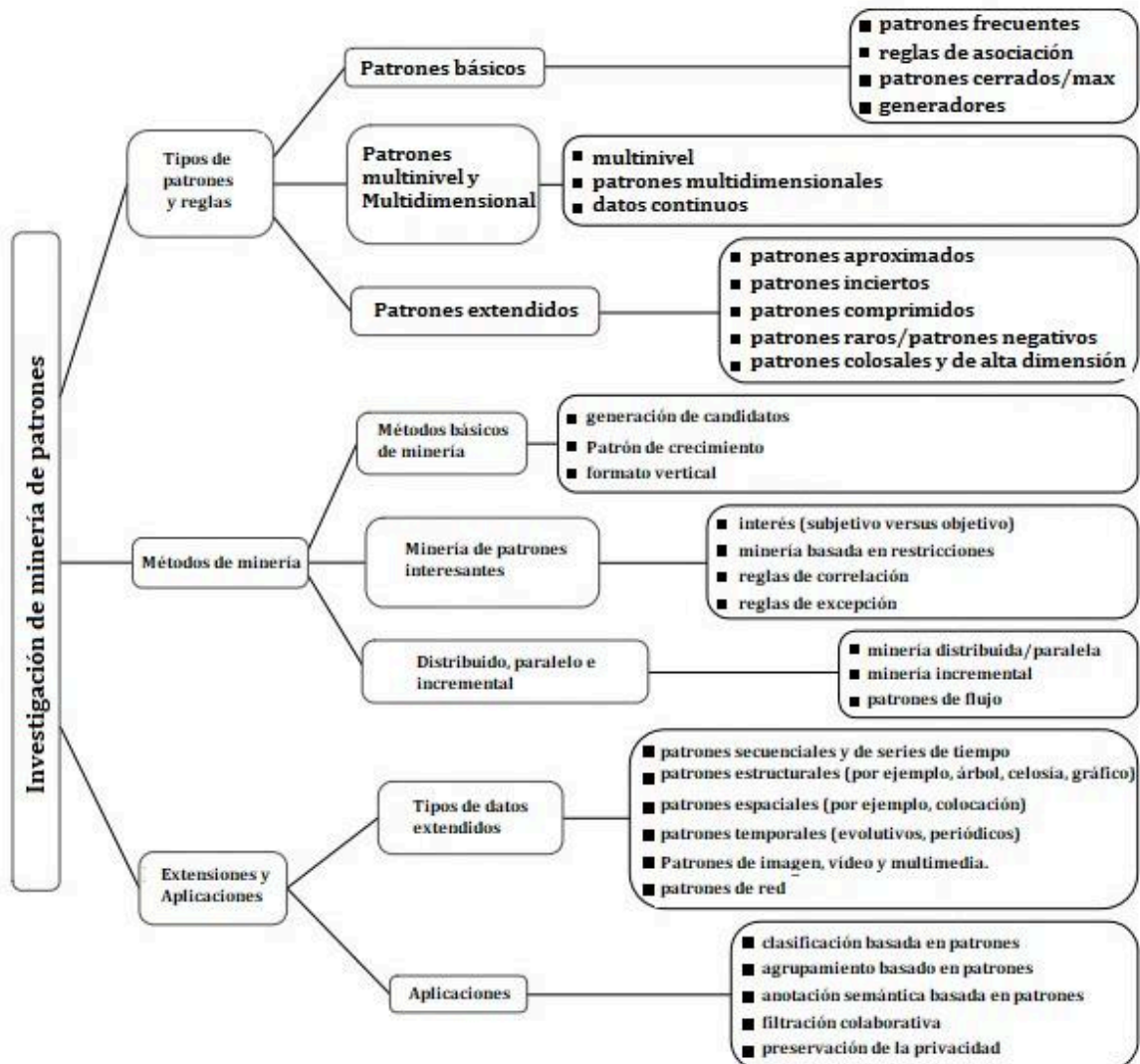
Predicciones y Pronósticos:

Utiliza modelos predictivos para anticipar tendencias futuras, facilitando la planificación estratégica y la previsión de desafíos.

La trascendencia de la minería de datos se manifiesta en su capacidad para facilitar el descubrimiento de patrones complejos, proporcionando una comprensión más profunda de los fenómenos subyacentes. Además, no solo se limita a la revelación de patrones; posibilita también la identificación de áreas de mejora en procesos empresariales y operativos, contribuyendo así a la optimización continua. La minería de datos va más allá, ofreciendo herramientas predictivas que utilizan modelos para anticipar tendencias futuras, facilitando la planificación estratégica y la previsión de desafíos. Ahora, adentrémonos en la investigación de Minería de Patrones.

Figura 8

Descifrando Patrones: De lo Básico a lo Avanzado en Minería de Datos



Nota. (Han, Kamber, Pei, 2012, p. 280)

Patrones Básicos

Patrones Frecuentes: Conjuntos de elementos que ocurren con frecuencia, como conjuntos de elementos frecuentes, subsecuencias y subestructuras frecuentes, revelando asociaciones y correlaciones.

Reglas de Asociación: Implicaciones del tipo $A \Rightarrow B$, definidas con soporte y midiendo la frecuencia conjunta de A y B en las transacciones.

Patrones Cerrados/Máximos: Conjuntos que no pueden extenderse manteniendo la misma frecuencia de soporte.

Generadores: Conjuntos mínimos que, al combinarse, generan conjuntos de elementos frecuentes.

Patrones Multinivel y Multidimensionales

Multinivel: Explora patrones en diferentes niveles de abstracción, permitiendo descubrimientos detallados o abstractos según sea necesario.

Patrones Multidimensionales: Examina patrones en múltiples dimensiones, incluyendo patrones de alta dimensión.

Datos Continuos: Información sin límites discretos, manejada mediante técnicas como discretización o análisis estadístico.

Patrones Extendidos

Patrones Aproximados: Extraen un conjunto comprimido o aproximado de patrones frecuentes para gestionar la cantidad masiva de patrones descubiertos.

Patrones Inciertos: Patrones donde la certeza o frecuencia no es exacta, con grados de incertidumbre.

Patrones Comprimidos: Minería de patrones comprimidos mediante la agrupación de patrones.

Patrones Raros/Negativos: Busca patrones raros o que reflejan correlaciones negativas entre elementos.

Patrones de Alta Dimensión y Colosales: Trata la minería de datos de alta dimensión y patrones colosales.

Métodos de Minería

Métodos Básicos

Generación de Candidatos: Algoritmo Apriori: Búsqueda por niveles que genera y poda candidatos para encontrar conjuntos de elementos frecuentes, utilizando anti monotonicidad y árboles hash.

Crecimiento de Patrones

Crecimiento de Patrones: Extrae conjuntos frecuentes sin generar candidatos, utilizando un árbol FP para comprimir la base de datos y mejorar la eficiencia.

Formato Vertical (Eclat, CHARM, ...):

Minería en Formato Vertical (Eclat): Transforma datos de transacciones y extrae conjuntos mediante intersecciones basadas en Apriori y optimizaciones adicionales.

Minería de Patrones Interesantes

Interesantes (Subjetivo vs. Objetivo): Explora la noción de interés, diferenciando entre criterios subjetivos y objetivos.

Minería Basada en Restricciones: Estrategia que incorpora restricciones del usuario para limitar el espacio de búsqueda, incluyendo tipo de conocimiento, datos, dimensión/nivel, interés y reglas.

Reglas de Correlación: Describen relaciones de correlación entre conjuntos de elementos en datos.

Reglas de Excepción: Destacan excepciones o patrones inusuales en los datos.

Distribuido, Paralelo e Incremental

Minería Distribuida/Paralela: Métodos que permiten la minería en entornos distribuidos o paralelos para mejorar la eficiencia en grandes conjuntos de datos.

Minería Incremental: Enfoque que permite la actualización continua de patrones a medida que se agregan nuevos datos.

Patrones de Flujo: Métodos diseñados para manejar flujos continuos de datos, extrayendo patrones en tiempo real.

Extensiones y Aplicaciones

Tipos de Datos Extendidos

Patrones Secuenciales y de Series Temporales: Representan secuencias o comportamientos a lo largo del tiempo.

Patrones Estructurales (Árboles, Celosías, Gráficos): Describen la estructura o relación entre elementos en forma de árboles, celosías, gráficos, entre otros.

Patrones Espaciales (Colocación): Revelan relaciones y ubicaciones espaciales entre elementos.

Patrones Temporales (Evolutivos, Periódicos): Capturan cambios a lo largo del tiempo, ya sea evolutivo o periódico.

Patrones de Imagen, Vídeo y Multimedia: Específicos en datos visuales o multimedia, como imágenes, videos y otros formatos multimedia.

Patrones de Red: Emergen de datos de red, describiendo conexiones y relaciones entre nodos.

Aplicaciones

Clasificación Basada en Patrones: Utiliza minería de patrones frecuentes como paso intermedio para mejorar la comprensión y el análisis de datos, conocido como clasificación basada en patrones.

Agrupamiento Basado en Patrones: Eficaz al agrupar datos de alta dimensionalidad, mejorando la comprensión y permitiendo la anotación semántica o el análisis contextual.

Anotación Semántica Basada en Patrones: Emplea patrones para agregar información semántica a los datos, mejorando su interpretación y utilidad.

Filtrado Colaborativo: Utiliza patrones para filtrar y recomendar información basándose en la colaboración entre usuarios con preferencias similares.

Preservación de la Privacidad: Implementa técnicas de minería de patrones que aseguran la privacidad de los datos, permitiendo análisis sin revelar información sensible.

2.4.4.4.- Presentación de Resultados

¿Por qué es crucial visualizar los datos?

La visualización de datos desempeña un papel fundamental en múltiples fases del análisis de datos, abarcando desde la limpieza inicial hasta la presentación final de hallazgos. Aunque a veces pasada por alto en los textos académicos, esta práctica adquiere relevancia real en investigaciones concretas. Un ejemplo claro se observa en el análisis de tiempos de finalización de maratones, donde la visualización revela patrones unilaterales, subrayando la constante presencia de esta etapa en la investigación de datos (marastats, 2019, citado en Unwin, 2020, p.2).

La frase "Una imagen vale más que mil palabras" enfatiza que una visualización, por sí sola, no puede reemplazar la necesidad de un contexto y una comprensión detallada para ser verdaderamente efectiva. Se resalta la noción de que, al igual que un libro completo, la presentación visual se vuelve memorable y comprensible cuando se respalda con información detallada. Además, se subraya la importancia de la sinergia entre texto y gráficos, destacando que ambos elementos se complementan y deben analizarse conjuntamente para obtener una comprensión más completa. Se enfatiza la necesidad de cuestionar y analizar críticamente los gráficos, incluso aquellos elogiados, para identificar posibles inexactitudes o imprecisiones. En resumen, se propone un enfoque crítico y reflexivo hacia la interpretación de gráficos en el contexto de la visualización de datos.

2.4.4.4.1.-Visualización de datos

Aunque pueda parecer paradójico, el éxito en la visualización de datos no se inicia con la propia visualización. En lugar de lanzarse directamente a la creación de gráficos o presentaciones de datos, es crucial dedicar tiempo y atención a comprender el contexto que impulsa la necesidad de comunicación. Este capítulo se enfoca en explorar los elementos esenciales del contexto y presenta estrategias que facilitarán la preparación para el éxito al comunicar visualmente con datos.

Distinción entre Análisis Exploratorio y Explicativo

Antes de sumergirnos en los detalles específicos del contexto, es imperativo establecer una distinción clara entre el análisis exploratorio y el explicativo. El análisis exploratorio se lleva a cabo para comprender los datos y descubrir aspectos notables o interesantes dignos de destacar para la audiencia. En términos prácticos, realizar un análisis exploratorio es como buscar perlas en ostras.

A menudo, es necesario explorar múltiples hipótesis o analizar los datos desde diversas perspectivas para encontrar información valiosa. Sin embargo, al comunicar nuestro análisis a la audiencia, es esencial trasladarse al ámbito explicativo. En este contexto, se tiene un mensaje específico para transmitir, una historia que contar, generalmente centrada en esa información valiosa.

Un error común radica en presentar un análisis exploratorio (mostrar simplemente los datos crudos) cuando se debería optar por un análisis explicativo (tomarse el tiempo para transformar los datos en información comprensible para la audiencia). Aunque es comprensible querer destacar todo el trabajo realizado y la solidez del análisis, resistir este impulso es crucial.

2.4.4.4.2.- Análisis Explicativo: Comprendiendo la Audiencia, Objetivos y Métodos

Cuando se aborda el análisis explicativo, es crucial reflexionar y tener claridad antes de visualizar datos o crear contenido. En primer lugar, ¿con quién se está comunicando? Comprender a fondo la audiencia y cómo la perciben es esencial para identificar puntos comunes que garanticen la atención de su mensaje. En segundo lugar, ¿qué desea que la audiencia conozca o haga? Debe tener en mente cómo quiere que la audiencia actúe, considerando cómo se comunicará con ellos y el tono general que desea establecer.

Solo después de responder de manera concisa a estas dos preguntas, uno está preparado para abordar la tercera: ¿Cómo puede utilizar los datos para expresar su punto de vista? Vamos a explorar en más detalle el contexto de quién, qué y cómo, a través de la explicación dada en el libro “Storytelling with data: a data visualization guide for business professionals”(Wiley, 2015, p.20).

Quién

Tu audiencia: Cuanto más específica sea la identificación de la audiencia, más eficaz será la comunicación. Evitar audiencias generales y focalizarse en grupos específicos permite adaptar el mensaje de manera más efectiva. La identificación de quienes toman decisiones contribuye a limitar la audiencia y mejora la resonancia del mensaje.

Tú: Considerar la relación con la audiencia y cómo se espera que se perciba. La confianza y la credibilidad son factores cruciales que influyen en la estructuración de la comunicación, el uso de datos y la narrativa general.

Qué

Acción: Definir lo que se espera que la audiencia sepa o haga. Es fundamental comunicar información relevante y justificar por qué debería importar a la audiencia. Establecer claramente el propósito de la comunicación es esencial para guiar la presentación de datos.

Cómo

Mecanismo: Determinar cómo se comunicará con la audiencia. El método de comunicación, ya sea una presentación en vivo o un documento escrito, tiene implicaciones en el control sobre la asimilación de la información y el nivel de detalle necesario. La elección del mecanismo dependerá de la audiencia y los objetivos comunicativos.

Finalmente, solo después de articular claramente quién es la audiencia y qué se espera que conozcan o hagan, se debe abordar la pregunta final: ¿Qué datos están disponibles para respaldar y expresar el punto de vista? Los datos se convierten en evidencia que respalda la historia construida y contada, y se explorará más sobre cómo presentar visualmente estos datos en los capítulos siguientes.

2.4.4.4.3.- Elegir una Imagen Efectiva

La elección de una imagen efectiva en visualización de datos es esencial para transmitir información de manera clara y comprensible. En este contexto, la variedad de gráficos disponibles desempeña un papel crucial, ya que diferentes tipos de visualizaciones son más adecuados para resaltar ciertos patrones o características en los datos. Desde gráficos de barras que comparan cantidades hasta diagramas de dispersión que exploran relaciones, cada tipo de gráfico tiene su propósito y aplicación específicos.

En esta sección bajo la definición de (Wiley, 2015, p.20), se verá cómo seleccionar sabiamente entre estos diversos tipos de gráficos para asegurar una representación precisa y efectiva de la información que se desea comunicar. La comprensión de las fortalezas y debilidades de cada formato gráfico permitirá a los comunicadores de datos elegir la imagen más apropiada para contar la historia que se encuentra en los datos.

texto

Cuando necesitas comunicar uno o dos números, es más efectivo utilizar un texto simple resaltando los números directamente, evitando el uso innecesario de gráficos. La fuerza de los números puede perderse en una representación visual excesiva.. En lugar de complicar con gráficos, una frase sencilla puede transmitir la información de manera clara. En la siguiente figura 9 se detalla como es visualmente un gráfico de texto.

Figura 9

Gráfico de texto



Nota.(Wiley, 2015, p.39).

Tablas

Las tablas son eficaces para la comunicación verbal, ya que se leen visualmente entre filas y columnas. Son ideales para audiencias diversas que buscan información específica. Las tablas son útiles al comunicar múltiples unidades de medida. El diseño de una tabla debe ser discreto para permitir que los datos sean el foco principal. Se sugiere utilizar bordes claros o incluso prescindir de ellos para evitar competir por la atención con los datos. En la siguiente figura 10 se muestra un ejemplo.

Figura 10

Tabla

Light borders				Minimal borders			
Group	Metric A	Metric B	Metric C	Group	Metric A	Metric B	Metric C
Group 1	\$X.X	Y%	Z,ZZZ	Group 1	\$X.X	Y%	Z,ZZZ
Group 2	\$X.X	Y%	Z,ZZZ	Group 2	\$X.X	Y%	Z,ZZZ
Group 3	\$X.X	Y%	Z,ZZZ	Group 3	\$X.X	Y%	Z,ZZZ
Group 4	\$X.X	Y%	Z,ZZZ	Group 4	\$X.X	Y%	Z,ZZZ
Group 5	\$X.X	Y%	Z,ZZZ	Group 5	\$X.X	Y%	Z,ZZZ

Nota.(Wiley, 2015, p.41).

Mapa de calor

Un mapa de calor es una estrategia que combina detalles tabulares con elementos visuales. Se utiliza para visualizar datos en formato de tabla, reemplazando o complementando los números con celdas de colores. Estas celdas de colores transmiten la magnitud relativa de los números, ofreciendo una representación visual efectiva de la información. En la siguiente figura se muestra un ejemplo de un mapa de calor.

Figura 11

Mapa de calor.

Heatmap			
LOW-HIGH			
	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

Nota.(Wiley, 2015, p.42).

Gráficos

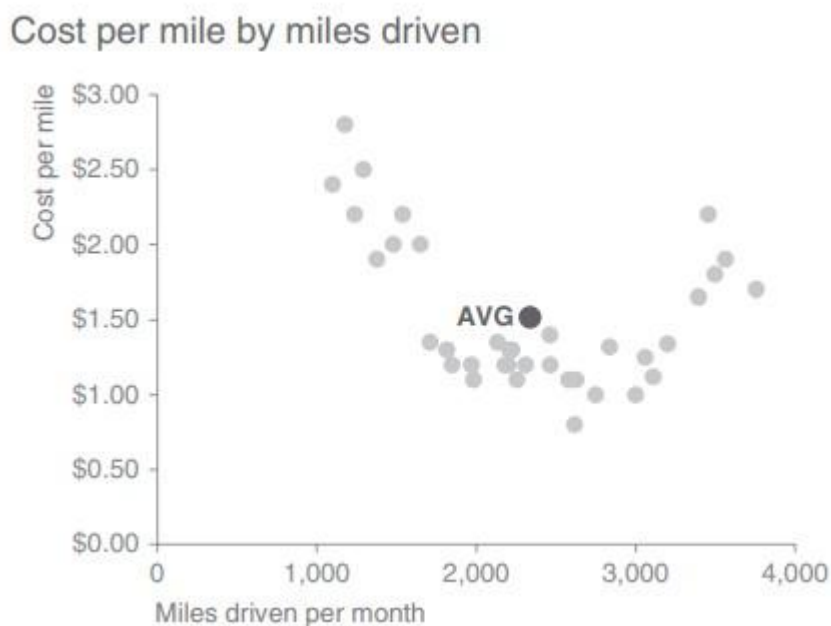
Los gráficos interactúan con nuestro sistema visual, procesando información más rápidamente que las tablas. Un gráfico bien diseñado suele transmitir información de manera más eficiente que una tabla. Aunque existen varios tipos de gráficos, unos pocos satisfacen la mayoría de las necesidades diarias.

Gráfico de dispersión

Los gráficos de dispersión son útiles para mostrar la relación entre dos variables al codificar datos en un eje x y un eje y, permitiendo identificar si hay una relación y su naturaleza. Aunque más comunes en campos científicos, también tienen aplicaciones en entornos empresariales. En la siguiente figura se muestra un ejemplo de un gráfico de dispersión.

Figura 12

Gráfico de dispersión



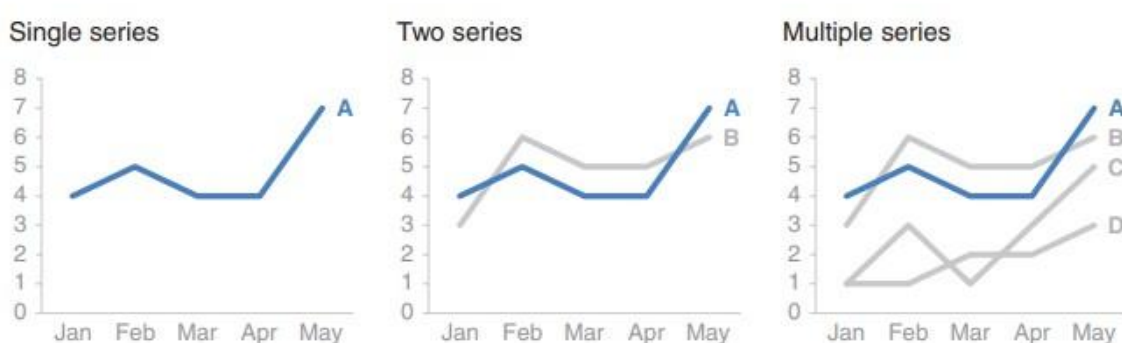
Nota.(Wiley, 2015, p.44).

Gráfico de Líneas

Los gráficos de líneas se utilizan con mayor frecuencia para trazar datos continuos. Debido a que los puntos están conectados físicamente a través de la línea, implica una conexión entre los puntos que puede no tener sentido para datos categóricos (un conjunto de datos ordenados o divididos en diferentes categorías). A menudo, nuestros datos continuos están en alguna unidad de tiempo: días, meses, trimestres o años. En la siguiente figura se muestra un ejemplo de gráfico de líneas.

Figura 13

Gráfico de líneas



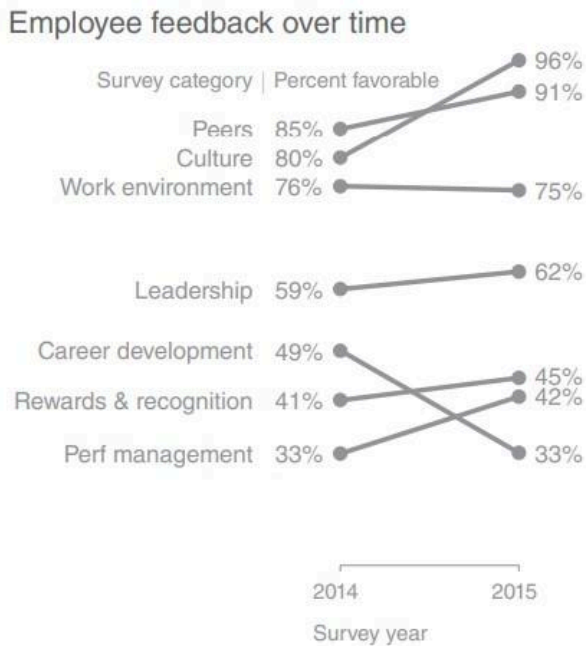
Nota. (Wiley, 2015, p.46).

Gráfico de Pendiente

Los gráficos de pendiente pueden ser útiles cuando tiene dos períodos de tiempo o puntos de comparación y desea mostrar rápidamente aumentos y disminuciones relativas o diferenciadas en varias categorías entre los dos puntos de datos.

Figura 14

Gráfico de pendiente



Nota.(Wiley, 2015, p.48).

Barras

A veces se evitan los gráficos de barras porque son comunes. Esto es un error. Más bien, se deben aprovechar los gráficos de barras porque son comunes, lo que significa una curva de aprendizaje menor para su audiencia. En lugar de utilizar su capacidad cerebral para tratar de comprender cómo leer el gráfico, su audiencia lo dedica a descubrir qué información extraer de lo visual.

Figura 15

Gráfico de barras



Nota.(Wiley, 2015, p.50).

Para concluir, es innegable que la visualización de datos desempeña un papel esencial en el entorno actual, donde los datos son una parte fundamental para el progreso y las mejoras continuas en diversos ámbitos como el empresarial, educativo y de la salud, centrándose en la gestión y análisis de datos. Algunas razones fundamentales que respaldan esta afirmación incluyen:

Rápida Asimilación de Información: La visualización permite una absorción veloz de la información, ya que una imagen supera en eficacia a mil líneas de datos. Con el crecimiento constante de los datos, la visualización facilita la identificación de tendencias y maneja de manera eficiente el flujo continuo de nueva información.

Comprensión de Próximos Pasos: A través de las tendencias visuales, se logra una comprensión más simple de los próximos pasos, ahorrando tiempo y energía que, de lo contrario, se dedicaría al análisis detallado de datos. Observar el panorama general en lugar de cada pieza individual ahorra horas de trabajo.

Conexión de Puntos: La visualización de datos no solo exhibe patrones y tendencias, sino que también destaca correlaciones y relaciones importantes pero sutiles entre las condiciones del negocio.

Mantenimiento del Interés de la Audiencia: La presentación atractiva de gráficos captura rápidamente la atención de la audiencia, crucial en un mundo con períodos de atención cortos. Mantener el interés resulta vital al compartir ideas de manera efectiva.

Reducción de la Dependencia de Científicos de Datos: La visualización hace que los datos sean accesibles y comprensibles para equipos más allá del departamento de TI. Los equipos de finanzas, ventas y marketing no solo muestran interés en los datos, sino que también tienen la capacidad de buscar respuestas por sí mismos.

Facilitación de la Compartición de Datos: Las visualizaciones se distribuyen fácilmente entre equipos, siendo más receptivos a visualizaciones atractivas que a masivas hojas de cálculo de Excel.

Identificación Rápida de Datos Atípicos: La visualización revela rápidamente valores atípicos, esenciales para entender su impacto en los datos. La identificación rápida permite comprender su relevancia y decidir si deben ser ignorados o no.

Ayuda en la Retención de Puntos Importantes: Los elementos visuales facilitan la retención de conceptos cruciales. Recordar y memorizar información es más efectivo cuando se tiene un gráfico como referencia, en lugar de depender solo de palabras o elementos de línea.

Facilita la Acción Posterior a un Descubrimiento: La visualización permite descubrir información y tomar decisiones de manera más rápida. Esto facilita la revisión y actualización eficiente de estrategias, contribuyendo al logro del éxito con menos errores y mayor velocidad.

2.4.4.5.- Toma de Decisiones

La visualización de datos desempeña un papel crucial en la toma de decisiones y análisis de tendencias en el contexto empresarial moderno. En palabras de W. Edwards Deming, "Sin datos, solo eres una persona cualquiera con una opinión", subrayando la importancia de respaldar las decisiones con información objetiva. La capacidad de interpretar visualmente datos complejos se ha vuelto imperativa en un entorno empresarial cada vez más centrado en datos.

La visualización no solo simplifica la comprensión de datos, sino que también agiliza el proceso de toma de decisiones al proporcionar representaciones gráficas impactantes de la realidad empresarial. Hans Rosling, experto en salud global y estadísticas, compara la visualización de datos con un "telescopio de nuestro tiempo", permitiendo ver aspectos que de otra manera pasarían desapercibidos.

En el ámbito de la minería de datos para el análisis de sentimientos en redes sociales, la visualización se convierte en un catalizador para transformar datos en conocimiento práctico. Permite identificar patrones, intereses y tendencias entre los usuarios, ofreciendo información valiosa para comprender las dinámicas sociales. Además, la capacidad predictiva de la visualización puede ser utilizada para anticipar tendencias futuras, como las carreras más solicitadas, proporcionando una guía valiosa para la toma de decisiones fundamentadas y el éxito en la implementación de estrategias basadas en datos.

Capítulo 3: Enfoque metodológico

Este capítulo se explorará cómo se utilizó la metodología KDD (Descubrimiento de Conocimiento en Bases de Datos).

Se explorarán las consideraciones esenciales para la implementación de KDD en el contexto del análisis de sentimientos y se distinguirá las diversas etapas del proceso, abordando desde el tratamiento y procesamiento de datos hasta la selección de modelos y algoritmos de minería de datos. Este capítulo sirve como hoja de ruta metodológica que asegura un enfoque sistemático y riguroso en el desarrollo y ejecución de la investigación.

3.1.- Descripción de la Metodología KDD para el Análisis de Sentimientos

La Metodología de Descubrimiento de Conocimiento (KDD) se posiciona como el marco metodológico central en este proyecto de análisis de sentimientos de los comentarios en las redes sociales de la Universidad de Playa Ancha. Esta metodología representa un proceso sistemático que dirigirá cada fase de la investigación, desde la identificación de objetivos hasta la interpretación de los resultados.

Esta sección introductoria proporciona una visión general, considerando que en el apartado "Metodologías para la minería de datos" se han detallado las metodologías para el desarrollo de proyectos de big data. Aquí, se analizará la esencia de KDD, destacando su estructura, objetivos y la importancia crucial que desempeñará en la extracción y análisis de datos de redes sociales. La aplicación de KDD no solo garantiza la rigurosidad científica del estudio, sino que también potencia la eficacia en la identificación de patrones y tendencias en los datos, allanando el camino para una interpretación profunda y fundamentada.

A continuación, detallaremos las etapas específicas que se llevarán a cabo bajo la metodología KDD y el análisis de sentimiento:

3.2.- Comprensión del contexto

En el marco del análisis de sentimientos en las redes sociales de la Universidad de Playa Ancha, la comprensión del contexto se erige como una fase crucial. Este proceso no solo se limita a un mero escrutinio superficial, sino que se sumerge en las capas más significativas de la interacción digital. El objetivo principal es trascender más allá de simples análisis para descubrir los temas de mayor relevancia y el sentir general de la comunidad universitaria.

Contexto Universitario y Desafíos en la Comunicación:

En el entorno actual de la educación superior, la efectividad de la comunicación y la participación en las redes sociales se convierte en una piedra angular para el éxito institucional. La Universidad de Playa Ancha enfrenta desafíos en la optimización de sus canales digitales y la promoción de una participación activa. La percepción pública de la institución impacta directamente en su imagen y calidad académica.

Importancia de una Estrategia Efectiva: La falta de una estrategia sólida puede limitar el alcance e influencia de la UPLA, dificultando la generación de un diálogo constructivo. La efectividad de la comunicación no solo implica transmitir información, sino también escuchar activamente y responder a las necesidades de la comunidad. Este enfoque es esencial para cumplir con la misión de las universidades estatales, que va más allá de la transmisión de conocimiento.

Colaboración Territorial y Desarrollo Regional: La comunicación efectiva no solo beneficia a la universidad sino que también enriquece la formación de los estudiantes y contribuye al desarrollo regional. La conexión con el entorno territorial permite a la UPLA adaptar sus programas a las demandas específicas de la región. La colaboración con actores locales genera un impacto significativo en el desarrollo regional y fortalece el compromiso de los estudiantes con la comunidad.

Enfoque de Análisis de Big Data y Sentimientos: Para abordar estos desafíos, se propone una metodología basada en el análisis de big data y el análisis de sentimientos. Estos enfoques permiten recopilar, procesar y analizar exhaustivamente datos generados en redes sociales. El análisis de big data proporciona una visión completa, identificando tendencias y temas emergentes. El análisis de sentimientos, por su parte, evalúa el tono emocional de las interacciones, proporcionando una comprensión más profunda de la percepción pública.

Objetivo del Estudio: El objetivo principal de este estudio es evaluar la efectividad de los canales de comunicación y participación en las redes sociales de la UPLA. Se busca identificar áreas de mejora mediante la aplicación de enfoques avanzados que permitan no solo recopilar datos, sino también comprender la percepción y

emociones de la comunidad universitaria. Esta evaluación integral busca fortalecer la conexión entre la universidad y quienes interactúan en sus plataformas digitales.

3.3.- Creación de un Conjunto de Datos Objetivo

Radiografía Detallada de las Redes Sociales

La creación del conjunto de datos objetivo se erige como el cimiento sobre el cual se construirá el análisis de sentimientos en las redes sociales de la Universidad de Playa Ancha. Con un enfoque específico en las plataformas más influyentes, como Facebook, Instagram y Youtube, se busca capturar de manera exhaustiva la esencia de las interacciones digitales.

Selección Estratégica de Plataformas: La elección de Facebook, Instagram y Youtube como plataformas primarias se fundamenta en su relevancia y alcance significativo en el entorno universitario. Estas plataformas no solo son las más utilizadas, sino que también ofrecen una diversidad de formatos y medios para interactuar con la audiencia.

Parámetros Clave para el Conjunto de Datos: El conjunto de datos objetivo se compone de elementos cruciales que proporcionarán una visión holística de la interacción digital:

1. Título de la publicación: El objetivo es poder identificar los temas de mayor interés por la comunidad. El análisis de los títulos revelará los temas que generan más participación y atención entre otros elementos a buscar.
2. Comentario y Respuestas: El objetivo es poder evaluar la calidad y tono emocional de la interacción, como también poder analizar los comentarios permite entender las opiniones y emociones expresadas por la comunidad.
3. Me Gusta y Compartidos: El objetivo es poder medir la popularidad y alcance de las publicaciones como la cantidad de me gusta y compartidos que refleja la resonancia de la información compartida.
4. Hora de Publicación: El objetivo es poder identificar las horas con mayor interacción como también conocer los momentos más activos proporcionando una perspectiva sobre el comportamiento de la audiencia.
5. Comentarios en Respuestas: El objetivo es poder destacar comentarios con mayor interacción como también comentarios que generan más respuestas indicando temas que provocan debate y mayor participación.
6. Metodología de Recopilación: La recopilación de datos se realizará de manera sistemática y regular, asegurando la representatividad de las interacciones a lo largo del tiempo. Se emplearán técnicas avanzadas para la extracción eficiente de datos en estas plataformas, garantizando la captura integral de la información relevante.

7. Impacto en el Análisis de Sentimientos: Este conjunto de datos detallado no solo facilitará la evaluación de la percepción pública, sino que también permitirá comprender los factores que impulsan la participación. La intersección entre la cantidad de interacciones y el análisis de sentimientos proporcionará una visión completa de la dinámica digital en la Universidad de Playa Ancha.

3.4.- Limpieza y Preprocesamiento de Datos

La limpieza y preprocesamiento de datos constituyen una etapa esencial en el análisis de sentimientos en las redes sociales de la Universidad de Playa Ancha. Este proceso se concibe como una suerte de destilación digital, donde la materia prima de la información cruda se somete a un refinamiento meticuloso. La necesidad de esta fase se fundamenta en la diversidad y complejidad de los datos recopilados, donde la falta de definición y la variabilidad en la calidad deben ser abordadas para garantizar una comprensión robusta y precisa.

Proceso Riguroso de Limpieza: La limpieza de datos se erige como el primer bastión, donde se emprende una búsqueda y eliminación de anomalías, inconsistencias y datos sin definición clara. Este proceso asegura que la información sea coherente, confiable y libre de artefactos que puedan sesgar los análisis subsiguientes.

Estandarización para la Comprensión: La diversidad de formatos, estilos y expresiones en las interacciones digitales exige una estandarización cuidadosa. Esta fase busca homogeneizar la información, estableciendo un formato cohesivo que facilite la comparación y comprensión de los datos. Así, desde abreviaturas hasta emojis, se traducen y estandarizan para garantizar una representación fiel de la comunicación digital.

Gestión de Datos Faltantes: La realidad digital a menudo presenta lagunas en la información, ya sea por comentarios incompletos o ausencia de datos en ciertos campos. Estrategias específicas se implementan para lidiar con estos vacíos, considerando la relevancia de cada dato en el contexto del análisis. La información temporal y secuencial también se aborda con técnicas que preservan su coherencia.

Procesamiento para Análisis de Sentimientos: La preparación de los datos no solo busca homogeneidad, sino también optimización para el análisis de sentimientos. Se aplican técnicas de procesamiento del lenguaje natural (NLP) para tokenizar y eliminar ruido lingüístico. Este paso sienta las bases para la interpretación adecuada de emociones y opiniones expresadas en los comentarios.

Validación Continua: La limpieza y preprocesamiento no son eventos aislados; son procesos iterativos que se ajustan según las demandas específicas de los datos y los objetivos del análisis. La validación continua asegura que la calidad de los datos perdure en todas las fases del proyecto.

Impacto en la Efectividad del Análisis:

Esta fase no solo es un requisito técnico, sino un impulsor crítico de la calidad de los análisis subsiguientes. Un conjunto de datos limpio y bien preparado potencia la precisión y confiabilidad del análisis de sentimientos, permitiendo que las conclusiones extraídas reflejen fielmente la dinámica emocional y opinativa en las redes sociales de la universidad.

3.5.- Coincidencia de Objetivos con Método de Extracción de Datos

Navegando las Emociones a través de la Clasificación

La selección cuidadosa del método de extracción de datos se convierte en una brújula vital para orientar el análisis de sentimientos en las redes sociales de la Universidad de Playa Ancha. En este viaje exploratorio de datos, la coincidencia de objetivos con la metodología adoptada se cristaliza en la elección deliberada de la clasificación como el medio principal para desentrañar las complejidades emocionales expresadas en los comentarios digitales.

Enfoque de Clasificación: El análisis de sentimientos demanda una comprensión avanzada de las emociones en los comentarios, y la clasificación emerge como la herramienta idónea para esta tarea. Este método categoriza los comentarios en tres clases fundamentales: Positivo, Negativo y Neutro. Esta tríada emocional proporciona una lente precisa para discernir las distintas tonalidades de las interacciones digitales.

Justificación de la Clasificación: La elección de la clasificación se respalda en la necesidad de evaluar el tono emocional de los comentarios de manera sistemática y cuantitativa. El enfoque trifásico permite capturar la riqueza de las respuestas emocionales, proporcionando una visión estructurada que facilita la interpretación y toma de decisiones.

Pysentimiento como Aliado Tecnológico: La implementación de Pysentimiento como herramienta de evaluación del tono emocional agrega un componente tecnológico crucial. Este paquete de Python se convierte en el oráculo que, mediante análisis de texto, asigna etiquetas emocionales a los comentarios, contribuyendo al proceso de clasificación de manera eficiente y escalable.

Desafíos y Ajustes Continuos: La clasificación no es un proceso estático; enfrenta el desafío de la diversidad lingüística y las expresiones contextuales. A medida que se navega por este océano de datos, se realizarán ajustes continuos en el método de clasificación para adaptarse a las particularidades del lenguaje digital y reflejar con precisión las emociones expresadas.

Impacto en la Profundidad del Análisis: La alineación de objetivos con la clasificación no solo simplifica la tarea de explorar las emociones en los comentarios, sino que también amplifica la profundidad del análisis. Al abordar las respuestas emocionales de manera estructurada, se desbloquea una comprensión más completa de la percepción y actitudes hacia la universidad en el paisaje digital.

Guiando el Viaje: La clasificación no solo actúa como un método, sino como un guía en este viaje de descubrimiento de conocimiento. Alineando los objetivos con la clasificación, se establece un curso claro hacia la comprensión profunda de las interacciones emocionales que dan forma a la narrativa digital de la Universidad de Playa Ancha.

3.6.- Minería de Datos

Explorando la Materia Prima Emocional

La minería de datos en el contexto de la Universidad de Playa Ancha se erige como una expedición en busca de patrones significativos, una búsqueda implacable que explorara la materia prima emocional tejida en los comentarios digitales. Este proceso de extracción de patrones se convierte en el pilar donde convergen las palabras clave, estructuras gramaticales y relaciones entre variables, revelando así los matices de los sentimientos expresados.

Exploración Profunda de Patrones: En la gran cantidad de datos generados en las redes sociales, la minería se transforma en una herramienta para la exploración profunda. Se busca más allá de las superficies de las interacciones digitales, desenterrando patrones que trascienden la mera secuencia de palabras y se sumergen en la esencia emocional de los comentarios.

Palabras Clave como Hilo Conductor: La identificación de patrones de palabras clave se posiciona como un hilo conductor en esta travesía. Las palabras, más allá de su función lingüística, se convierten en señales reveladoras de emociones. La minería meticulosa de estas palabras clave facilita la construcción de un atlas emocional, cartografiando los puntos de mayor resonancia en la comunidad digital.

Estructuras Gramaticales y Relaciones Entre Variables: La exploración no se limita al léxico; se extiende a las estructuras gramaticales y las relaciones entre variables. La forma en que se construyen las oraciones, las conexiones lógicas entre ideas y la interdependencia de diferentes variables emergen como elementos cruciales. Estos detalles, a menudo subyacentes, ofrecen una visión más rica de la complejidad emocional expresada en el ciberespacio universitario.

Relevancia para la Identificación de Sentimientos: Cada patrón extraído no es simplemente un descubrimiento aislado; tiene relevancia en la identificación de sentimientos. La conexión entre patrones específicos y las categorías emocionales (positivo, negativo, neutro) se convierte en la piedra angular para asignar significado a la expresión digital.

Herramientas Tecnológicas como Cinceles Virtuosos: En esta tarea ardua, las herramientas tecnológicas, en particular aquellas diseñadas para análisis de texto, actúan como cinceles virtuosos. Algoritmos avanzados, incluyendo aquellos basados en inteligencia artificial, permiten una extracción eficiente de patrones, acelerando el proceso y amplificando la capacidad de descubrimiento.

Iteración y Adaptación Continua: La minería de datos no es una empresa estática. La iteración continua, la adaptación a nuevas tendencias lingüísticas y la evolución de patrones a lo largo del tiempo son elementos esenciales. Este enfoque dinámico asegura que la identificación de sentimientos no quede estancada en un momento específico, sino que evolucione con la comunidad digital.

Iluminando la Riqueza Emocional: En última instancia, la extracción de patrones no solo ilumina la riqueza emocional de los comentarios, sino que también proporciona un conjunto de datos estratificado que se convierte en la materia prima para análisis posteriores. Es en estos patrones identificados donde reside el tesoro de percepciones y actitudes que dan forma a la narrativa digital de la Universidad de Playa Ancha. La minería de datos no es simplemente un medio, sino un proceso transformador que desvela la esencia emocional inmersa en las interacciones digitales.

3.7.- Planificación

La Carta Gantt, una herramienta de gestión de proyectos ampliamente reconocida, desempeña un papel crucial en la planificación y ejecución efectiva de proyectos académicos, como la elaboración de un proyecto de título. En este contexto, la presentación de la Carta Gantt se convierte en una herramienta valiosa para comunicar la estructura temporal y las fases clave del proyecto a todas las partes interesadas.

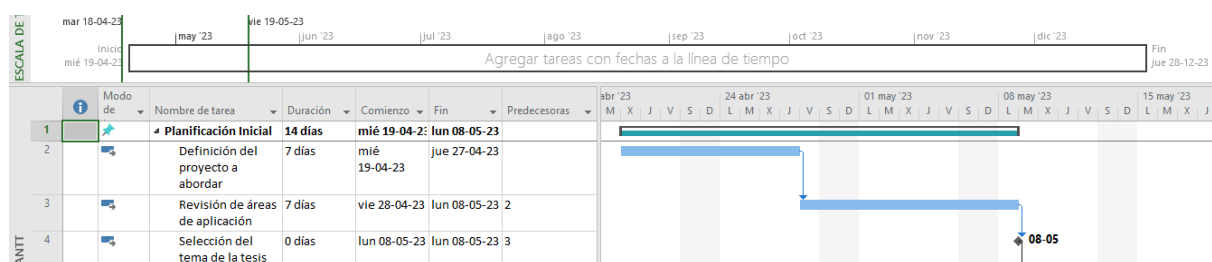
En esta sección, se explorará la manera de abordar la presentación de la Carta Gantt en el marco de la planificación del proyecto de título. Nos centraremos en detallar cada fase del proyecto, desde la definición inicial hasta la entrega final, destacando las tareas específicas y los hitos clave que conforman el camino hacia el logro exitoso de los objetivos de investigación.

Al desglosar las actividades planificadas y sus respectivos plazos, no solo se proporcionará una visión clara de la progresión del proyecto, sino que también se facilitará la identificación de posibles desafíos y la gestión eficiente de los recursos disponibles. La presentación de la Carta Gantt no solo se convierte en una herramienta de programación, sino también en un medio para fomentar la transparencia, la colaboración y la toma de decisiones informada a lo largo del ciclo de vida del proyecto.

A continuación, se detallarán las tareas clave junto con sus descripciones correspondientes.

Figura 16

Carta Gantt: Planificación inicial



Nota. Elaboración propia.

Tarea planificación inicial

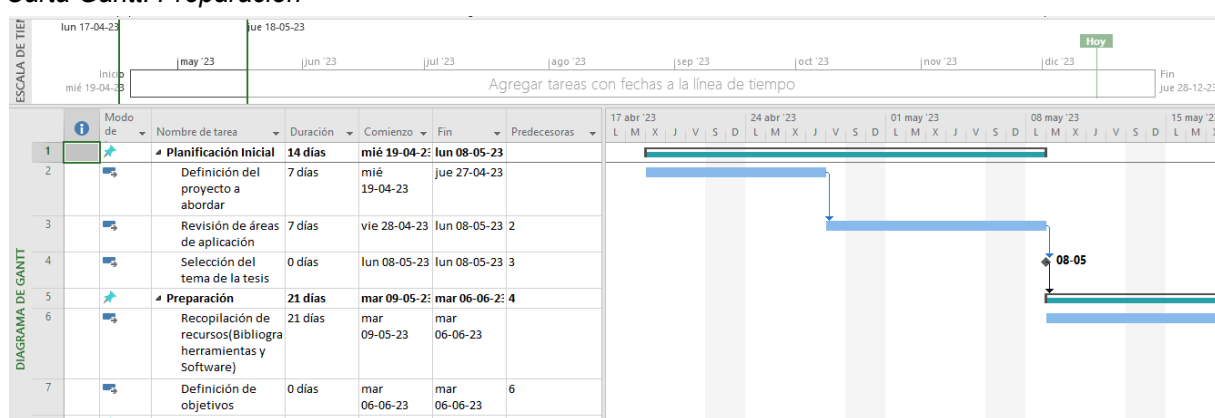
- Definición del proyecto a abordar: En esta fase, se llevará a cabo una identificación detallada de las áreas de interés relacionadas con el ámbito de

estudio. Se explorarán diversos campos para definir claramente el proyecto a abordar en el proyecto de título.

- Revisión de áreas de aplicación: Se llevará a cabo una revisión exhaustiva de las áreas de aplicación relevantes para asegurar que el proyecto de título tenga una orientación específica y esté alineado con las necesidades actuales.
- Selección del tema del proyecto de título: Este paso implica la elección y aprobación del tema de investigación por parte del profesor guía, asegurando su relevancia y viabilidad para la realización del proyecto de título.

Figura 17

Carta Gantt: Preparación



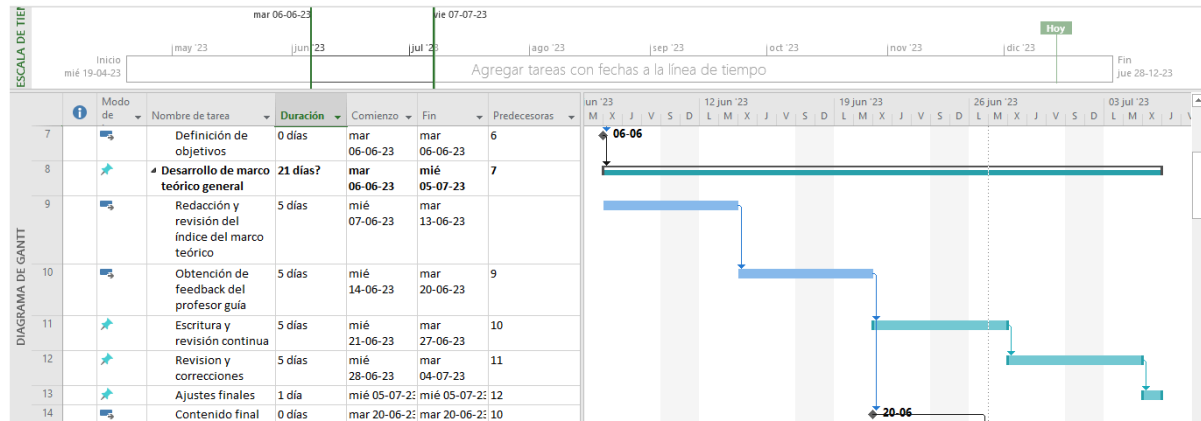
Nota. Elaboración propia.

Tarea preparación

- Recopilación de recursos (Bibliografía, herramientas y Software): Se llevará a cabo la recolección de fuentes bibliográficas, herramientas y software esenciales para respaldar la investigación. Este paso permitirá evaluar la disponibilidad de herramientas y métodos para realizar el análisis de sentimientos con big data, asegurando una base robusta para el desarrollo subsiguiente.
- Definición de objetivos: En esta fase, se detallarán y evaluarán los objetivos del proyecto, asegurando su factibilidad y alineación con el análisis de sentimientos en las redes sociales de la universidad.

Figura 18

Carta Gantt: Desarrollo de marco teórico general



Nota. Elaboración propia.

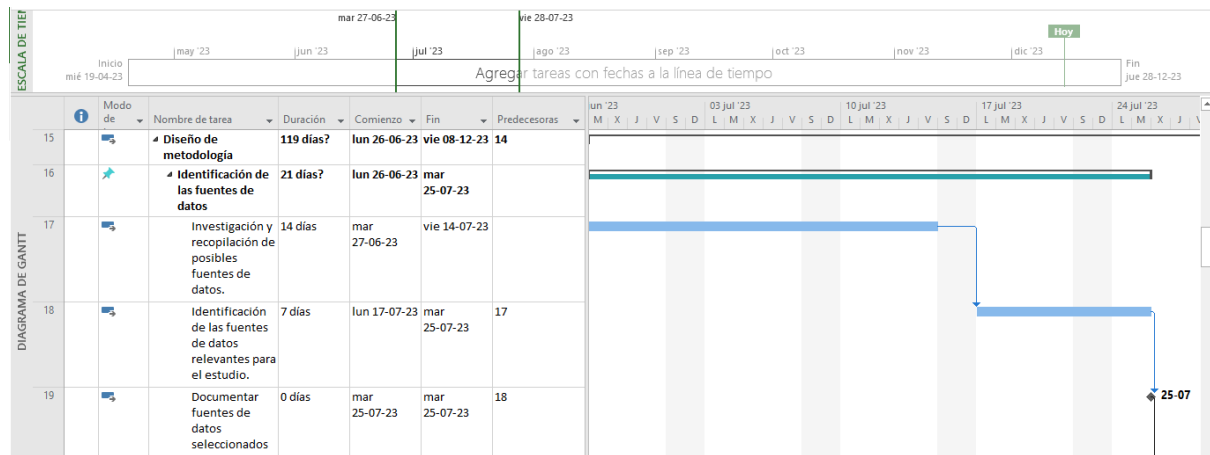
Tarea desarrollo de marco teórico general

- Redacción y revisión del índice del marco teórico: Se procederá a la estructuración y redacción del marco teórico, seguido de una revisión detallada del índice. Este paso tiene como objetivo asegurar una presentación lógica y coherente de los conceptos fundamentales.
- Obtención de feedback del profesor guía: Se buscará activamente la retroalimentación del profesor guía, buscando mejorar y perfeccionar el marco teórico a través de sus comentarios y sugerencias.
- Escritura y revisión continua: El proceso de escritura será continuo, con revisiones periódicas para garantizar la coherencia del contenido y su alineación con los estándares académicos establecidos.
- Revisión y correcciones: Se llevarán a cabo revisiones exhaustivas, enfocadas en la corrección de errores gramaticales, mejora de la claridad y aseguramiento de la cohesión del marco teórico en su totalidad.
- Ajustes finales: Últimos ajustes y refinamientos se aplicarán para perfeccionar el marco teórico, preparándolo de manera óptima para la siguiente fase del proyecto.
- Contenido final: Se consolidará y finalizará el contenido del marco teórico, asegurando su completitud y coherencia antes de su inclusión definitiva en el proyecto de título.

Tarea diseño de metodología

Figura 19

Carta Gantt: Identificación de las fuentes de datos

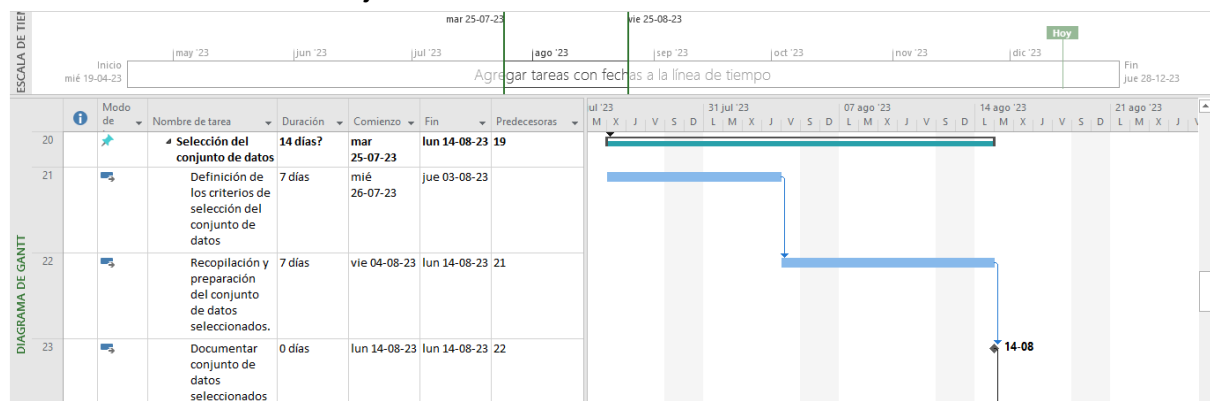


Nota. Elaboración propia.

Identificación de las fuentes de datos: Realización de una investigación exhaustiva para identificar fuentes de datos relevantes para la investigación. Identificación de las fuentes de datos relevantes para el estudio.

Figura 20

Carta Gantt: Selección del conjunto de datos

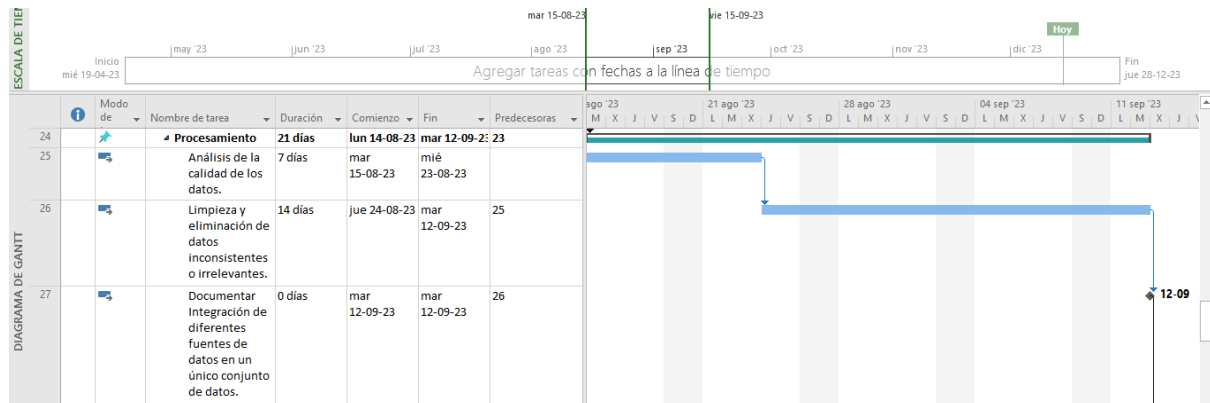


Nota. Elaboración propia.

Selección del conjunto de datos Definición de criterios claros para la selección del conjunto de datos que se utilizará en la investigación.

Figura 21

Carta Gantt: Procesamiento

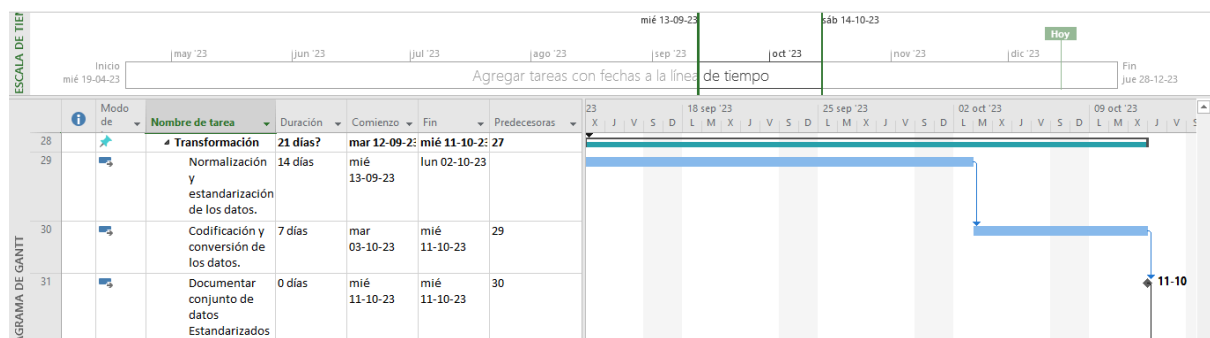


Nota. Elaboración propia.

Procesamiento: Preparación de los datos para su análisis. Análisis de la calidad de los datos. Limpieza y eliminación de datos inconsistentes o irrelevantes.

Figura 22

Carta Gantt: Transformación

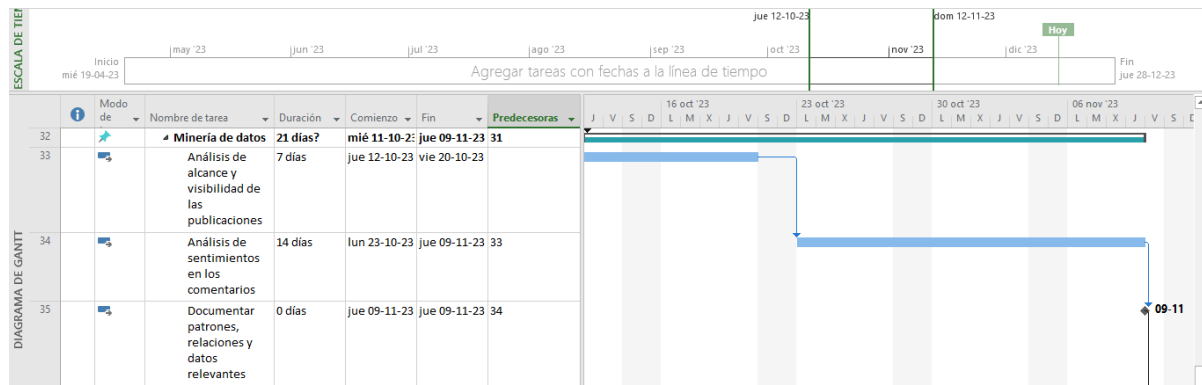


Nota. Elaboración propia.

Transformación: Preparación de los datos para el análisis mediante la normalización y estandarización. Normalización y estandarización de los datos.

Figura 23

Carta Gantt: Minería de datos



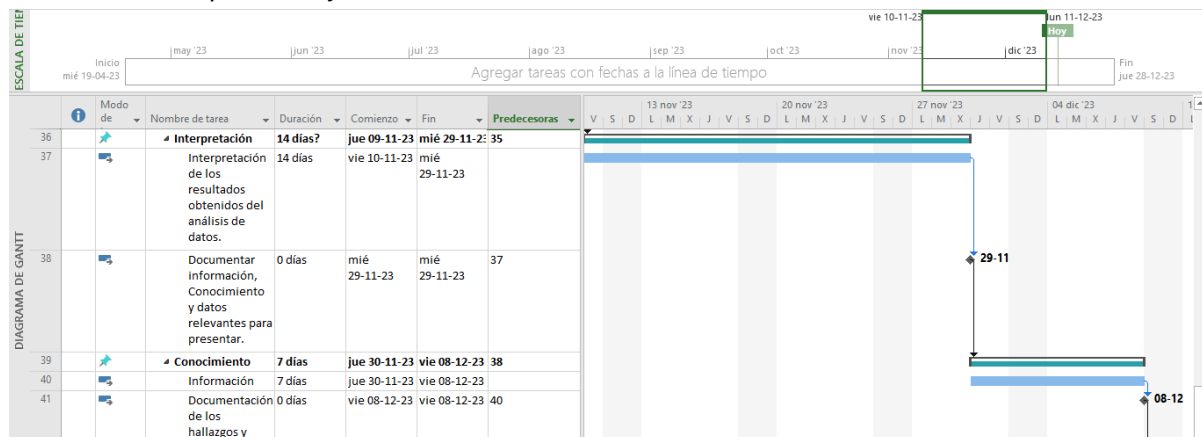
Nota. Elaboración propia.

Minería de datos: Análisis de datos para descubrir patrones y tendencias.

Actividades Principales: Análisis de alcance y visibilidad de las publicaciones y contenidos compartidos por la UPLA en redes sociales. Análisis de sentimientos en los comentarios de las publicaciones de la UPLA en redes sociales. entre otros patrones que puedan surgir.

Figura 24

Carta Gantt: Interpretación y conocimiento



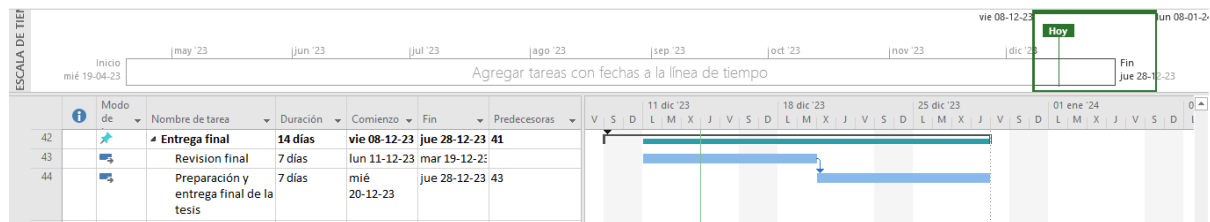
Nota. Elaboración propia.

Interpretación: Dar sentido a los resultados obtenidos en la minería de datos. Interpretación de los resultados para extraer conclusiones significativas.

Conocimiento: Consolidación de hallazgos en una presentación coherente. Presentación clara y comprensible de la información.

Figura 25

Carta Gantt: Entrega final



Nota. Elaboración propia.

Tarea entrega final: Revisión final, se llevará a cabo una revisión completa del trabajo para corregir cualquier error y mejorar la presentación. Revisión exhaustiva del contenido en busca de errores gramaticales, ortográficos y de formato. Mejora de la claridad y coherencia del documento.

Preparación de la versión final del proyecto de título para su presentación y entrega. Finalización de los ajustes sugeridos durante la revisión final.

Capítulo 4: Presentación de resultados

4.1.- Implementación

En este apartado, se presenta una exhaustiva descripción de los algoritmos y programas empleados para llevar a cabo el análisis de sentimientos en las publicaciones de las redes sociales de la Universidad de Playa Ancha. Se destaca la prioridad otorgada al diccionario léxico-emocional, seleccionado por su eficacia comprobada en estudios previos y su perfecta integración con la herramienta pysentimiento, esencial para la ejecución de este análisis.

El conjunto de herramientas seleccionadas para la implementación del análisis de sentimientos en las publicaciones de las redes sociales de la Universidad de Playa Ancha se compone de varias bibliotecas y recursos especializados. Además de las descripciones detalladas de cada herramienta.

4.1.1.- Proceso Detallado del Entorno de Desarrollo

Configuración del Entorno:

En la creación de este entorno de desarrollo, se emplearon las siguientes especificaciones de hardware:

- Procesador: Intel Core i3-7100
- Memoria RAM: 16 GB
- Sistema Operativo: Windows 10 Home, 64 bits

Estas características proporcionan la capacidad y estabilidad necesarias para llevar a cabo eficientemente todas las fases del análisis de sentimientos, desde la extracción de datos hasta la visualización de resultados. La elección específica de este hardware se basó en la consideración de los requisitos computacionales del proyecto y en la optimización del rendimiento durante todo el proceso de desarrollo.

El desarrollo del análisis de sentimientos se ejecuta en un entorno Jupyter Notebook que utiliza Python versión 3.8.5. Este entorno integral abarca todas las fases del proceso, desde la extracción de comentarios hasta la visualización de resultados, incluyendo almacenamiento, normalización y minería de datos.

Integración de Herramientas:

pandas versión 1.1.3: Esta biblioteca es esencial para la manipulación y análisis eficiente de datos estructurados. Su versatilidad permite realizar operaciones complejas en conjuntos de datos, facilitando la preparación y organización de la información recopilada.

selenium versión 4.1.0: Selenium es una herramienta fundamental en el proceso de web scraping, permitiendo la automatización del navegador web. En este contexto, se utiliza para extraer comentarios valiosos de las redes sociales de la Universidad de Playa Ancha, asegurando un acceso eficiente y estructurado a la información deseada. utilizado para la extracción de comentarios de Facebook e Instagram.

facebook-scraper versión 0.2.59: Diseñado específicamente para la extracción de datos de la plataforma Facebook, este scraper proporciona una solución eficaz y especializada. Su implementación permite acceder y recopilar información relevante de manera ética y eficiente.

fake-useragent versión 1.1.3: Esta herramienta esencial proporciona encabezados de usuario ficticios, desempeñando un papel crucial para evitar restricciones durante el proceso de extracción de datos. Garantiza un acceso sin inconvenientes y ético a la información, respetando las políticas de privacidad de las plataformas visitadas.

google-api-core versión 2.11.1: Estas bibliotecas son necesarias para la interacción efectiva con la API de Google. Facilitan la obtención de datos relevantes y la integración de información específica que contribuye al análisis de sentimientos. Utilizado para la extracción de comentarios de YouTube.

pysentimiento versión 0.7.1: Esta librería esencial está dedicada al análisis de sentimientos. Ofrece herramientas y métodos robustos para evaluar y comprender las emociones expresadas en los comentarios recopilados, proporcionando una capa fundamental para la interpretación de resultados.

matplotlib versión 3.3.2 y seaborn versión 0.11.0: Estas librerías son utilizadas para la visualización clara y efectiva de resultados. Proporcionan herramientas gráficas poderosas que permiten representar de manera comprensible los patrones y las tendencias identificadas durante el análisis de sentimientos.

4.1.2.- Selección de plataformas de redes sociales para el análisis de sentimientos

La elección de utilizar una amalgama de medios de comunicación social, como Facebook, Instagram y YouTube, se fundamenta en la riqueza y diversidad de formatos que ofrecen. Cada plataforma destaca por sus características distintivas, proporcionando un abanico de posibilidades para la recopilación de datos.

Facebook, con su versatilidad en publicaciones de texto y multimedia, permite explorar expresiones sentimentales de manera detallada y variada. Instagram, centrado en imágenes y breves descripciones, aporta una perspectiva visual y concisa. Por último, YouTube, como plataforma de video, añade una dimensión única al análisis al capturar las interacciones en formato audiovisual. Esta variedad no solo enriquece la investigación, sino que también facilita una comprensión más completa de las dinámicas emocionales.

Además, la elección de estas plataformas se respalda en su amplia audiencia dentro de la comunidad universitaria. Al ser utilizadas de manera extendida y diversificada por personas de diversas edades, intereses y preferencias de contenido, garantizan una muestra representativa y completa de las interacciones sentimentales en la Universidad de Playa Ancha. La presencia activa en estas redes sociales contribuye significativamente a la validez y aplicabilidad de los resultados obtenidos.

Es esencial destacar la relevancia intrínseca de estas plataformas para la Comunidad Universitaria. No solo son populares a nivel general, sino que también sirven como canales de comunicación específicamente relevantes para la comunidad académica. En estos espacios, se comparten y discuten eventos, noticias y actividades relacionadas con la universidad, convirtiéndose en un reflejo fiel de las experiencias y percepciones de los miembros de la comunidad académica.

Asimismo, la elección se basa en la diversidad de contenidos que ofrecen estas plataformas. Facebook alberga publicaciones textuales y multimedia, Instagram se centra en imágenes y contenido escrito breve, mientras que YouTube proporciona un espacio para videos informativos y creativos. Esta variedad de formatos asegura la captura de una amplia gama de expresiones sentimentales, enriqueciendo la comprensión de las dinámicas emocionales.

La combinación estratégica de estas plataformas permite un análisis multidimensional de las interacciones sentimentales. La participación en comentarios, reacciones, compartidos y visualizaciones de videos brinda una visión completa de cómo la comunidad interactúa y responde a diferentes tipos de contenido. Esto proporciona una comprensión más profunda y holística de las

dinámicas emocionales dentro de la Universidad de Playa Ancha. La elección de estas redes sociales se fundamenta no solo en su popularidad, sino también en su capacidad para ofrecer una visión completa y matizada de las expresiones sentimentales, asegurando la robustez y relevancia de la investigación.

Además, el proceso de extracción de datos se inició el 15 de noviembre y se prolongó hasta completar la recolección de 500 publicaciones en cada plataforma de redes sociales. Este análisis implica recorrer minuciosamente las publicaciones del pasado para alcanzar la meta establecida. Durante esta búsqueda exhaustiva, se logró recopilar información pertinente desde agosto hasta noviembre, abarcando así la segunda mitad del año 2023. Cada punto temporal capturado constituye y proporciona una visión de la dinámica digital en ese periodo. Este enfoque temporal tiene como objetivo descubrir, con precisión, las variaciones y tendencias que han definido el entorno digital en este intervalo específico.

4.1.3.- Recopilación, preparación y obtención de los datos de las redes sociales relevantes para el análisis

En esta sección, se presenta en detalle el proceso exhaustivo de recopilación, preparación y obtención de datos de las redes sociales seleccionadas para llevar a cabo un análisis de sentimientos efectivo.

Para obtener datos de Facebook, Instagram y YouTube, se implementaron técnicas de webscraping y acceso a las APIs respectivas. La ejecución precisa de estas técnicas aseguró la obtención de datos relevantes, respetando rigurosamente las políticas y términos de servicio de cada plataforma. Cabe destacar que, como límite máximo, se extrajeron 500 publicaciones de las redes sociales con el objetivo de no saturar las páginas para evitar el bloqueo o desconexión de estas.

4.1.4.- Datos de interés de las redes sociales

Facebook

La selección de estos datos para el análisis Big Data con enfoque en el análisis de sentimientos sobre las emociones de los comentarios en publicaciones de la universidad en Facebook está justificada por su relevancia y capacidad para proporcionar información valiosa sobre la dinámica y la percepción de la comunidad. De esta forma los datos de interés son los siguientes:

Número de la publicación: Se realiza un conteo de las publicaciones extraídas en la plataforma. Este dato proporciona información sobre la frecuencia y volumen de contenido compartido.

Título: El título de la publicación es una información crucial para el análisis temático. Proporciona contexto sobre el contenido específico de la publicación, permitiendo una categorización adecuada.

Texto del Comentario: El texto del comentario es un elemento fundamental para el análisis general de sentimientos. Contiene la esencia de la interacción y permite evaluar las opiniones y emociones expresadas por los usuarios.

Me gusta: La cuantificación de los "Me gusta" proporciona una medida de la aprobación de la publicación. Este dato es esencial para comprender la recepción positiva del contenido.

Compartir: La evaluación de la viralidad a través de los compartidos indica la difusión y alcance potencial de ciertas publicaciones. Permite identificar contenido que ha resonado significativamente en la audiencia.

Fecha y Hora: La marca temporal de la publicación del comentario proporciona información sobre cuándo ocurrió la interacción. Es crucial para entender la temporalidad y la relevancia en el tiempo de las interacciones.

Reacciones al Comentario: La exploración de interacciones expresadas mediante emojis ofrece una visión más rica de las respuestas emocionales de la audiencia.

Respuestas: La identificación de comentarios generadores de interacción proporciona información sobre la participación activa de la audiencia, destacando aquellos comentarios que generan respuestas.

Instagram

Por otro lado, la inclusión de datos de Instagram en nuestro análisis se fundamenta en la riqueza informativa que esta red social proporciona. Instagram, siendo una plataforma visualmente centrada, nos ofrece una variedad de datos esenciales para comprender la interacción de los usuarios. De esta forma los datos de interés son los siguientes:

ID de Publicación: El ID de publicación es un identificador único para cada publicación en Instagram. Facilita la organización y referencia de datos.

Título: El título de la publicación en Instagram proporciona información sobre el contenido específico de la imagen o video compartido.

Comentario: El comentario asociado a la publicación refleja la interacción y comentarios de los usuarios.

Me gusta: La cuantificación de los "Me gusta" en Instagram proporciona una medida de la aceptación y popularidad de la publicación.

Fecha: La fecha de la publicación indica cuándo se compartió la imagen o video.

YouTube

La integración de datos de YouTube en nuestro análisis se justifica por la robusta fuente de información que esta plataforma representa. Desde el distintivo ID del video hasta el número de visitas, cada atributo seleccionado desempeña un papel crucial en la comprensión de la recepción y el impacto de los videos compartidos. YouTube, siendo una plataforma ampliamente utilizada para el contenido visual, nos brinda elementos esenciales para evaluar la popularidad, la interacción de los usuarios y la difusión de los videos. De esta forma los datos de interés son los siguientes:

ID del Video: El ID del video en YouTube es un identificador único que facilita la referencia y organización de los datos.

Título del Vídeo: El título del vídeo ofrece información sobre el contenido específico del video.

Autor: El autor del video indica quién creó y compartió el contenido en la plataforma.

Fecha de Publicación: La fecha de publicación del video proporciona información sobre la antigüedad del contenido.

Me gusta: La cuantificación de los "me gusta" en YouTube indica la aceptación y aprobación del contenido.

Compartir: La información sobre compartidos refleja la difusión del video en la plataforma.

Visitas: El número de visitas ofrece una medida de la popularidad y alcance del video.

Comentario: El comentario asociado al video refleja la interacción y comentarios de los usuarios en la plataforma.

4.1.5.- Almacenamiento, limpieza y preprocesamiento de datos para eliminar ruido y redundancia

En esta sección, se mostrará el crucial proceso de preparación de datos, que desempeña un papel fundamental en el análisis de información proveniente de múltiples plataformas de redes sociales. La gestión eficiente de datos es esencial para garantizar la calidad, confiabilidad y utilidad de los resultados obtenidos en investigaciones basadas en información digital.

Objetivos:

1. Almacenamiento: Presentaremos el manejo inicial de los datos recolectados desde Facebook, Instagram y YouTube, destacando el uso de herramientas como Pandas para cargar la información en estructuras de datos accesibles y manipulables.

2. Limpieza y Eliminación de Ruido: Las técnicas y estrategias empleadas para identificar y eliminar datos no deseados o irrelevantes, conocidos como ruido. Este proceso es esencial para obtener conjuntos de datos más coherentes y fiables.

3. Preprocesamiento: Abordaremos la estandarización y normalización de los datos, asegurándonos de que las diferentes fuentes de información sigan un formato uniforme. Además, nos enfocaremos en la detección y eliminación de redundancias, contribuyendo a la eficiencia y claridad de la información.

Importancia del Preprocesamiento: La calidad de los resultados en análisis de datos depende en gran medida de la calidad de los datos utilizados. La eliminación de ruido y la normalización adecuada son pasos cruciales para garantizar que las conclusiones extraídas sean precisas y confiables.

A lo largo de esta sección, se destacarán los métodos específicos utilizados en cada etapa del proceso, mostrando cómo estos contribuyen a la mejora de la integridad y utilidad de los datos.

4.1.6.- Técnicas de procesamiento de texto y análisis de sentimiento para formar un conjunto de datos más robusto

La creciente presencia de información en plataformas digitales, especialmente en redes sociales y medios en línea, ha generado un vasto conjunto de datos que presenta tanto desafíos como oportunidades para su análisis. En este contexto, las técnicas de procesamiento de texto y análisis de sentimiento emergen como

herramientas esenciales para extraer conocimiento valioso a partir de grandes volúmenes de información textual.

La investigación y aplicación de técnicas de procesamiento de texto se centran en la comprensión y manipulación de datos escritos en forma de texto. Este campo abarca diversas disciplinas, desde la tokenización y lematización hasta el modelado de temas y la generación de resúmenes. La capacidad de procesar y entender el contenido textual se ha vuelto crucial en diversos ámbitos, desde la investigación académica hasta el desarrollo de aplicaciones prácticas en el mundo empresarial.

Por otro lado, el análisis de sentimiento se posiciona como una rama clave en el procesamiento de texto al buscar identificar y clasificar las emociones, opiniones y actitudes expresadas en el lenguaje escrito. En un entorno donde la retroalimentación del usuario y la percepción pública son esenciales, el análisis de sentimiento se convierte en una herramienta invaluable para comprender la recepción y el impacto de contenidos específicos.

En este contexto, el presente análisis se enfoca en la aplicación de técnicas avanzadas de procesamiento de texto y análisis de sentimiento para extraer características relevantes de conjuntos de datos específicos. El objetivo principal es desentrañar patrones, identificar tendencias y comprender la dinámica de la información textual en un contexto particular. A través de la implementación de estas técnicas, se pretende ofrecer una perspectiva más profunda y significativa de los datos, permitiendo una toma de decisiones más informada y una comprensión más completa de los aspectos clave que subyacen en el contenido analizado.

En las secciones siguientes, se examinarán las metodologías y enfoques específicos utilizados para llevar a cabo el procesamiento de texto y análisis de sentimiento, así como los resultados obtenidos y su relevancia en el contexto de la investigación propuesta.

Las herramientas mencionadas son bibliotecas de Python ampliamente utilizadas en el ámbito de procesamiento de texto y análisis de datos. A continuación se describirán brevemente algunas de las herramientas utilizadas en el análisis:

1. Gensim: Gensim es una biblioteca de modelado de temas y procesamiento de texto en Python. Se utiliza comúnmente para realizar tareas como la creación de modelos LDA (Latent Dirichlet Allocation) para descubrimiento de temas en conjuntos de datos de texto.

2. NLTK (Natural Language Toolkit): NLTK es una biblioteca de procesamiento del lenguaje natural que proporciona herramientas para trabajar con datos de texto en lenguaje humano. Incluye módulos para tokenización, lematización, análisis de sentimientos y otras tareas de procesamiento de texto.

3. WordCloud: WordCloud es una biblioteca para la creación de nubes de palabras en Python. Se utiliza para visualizar las palabras más frecuentes en un conjunto de datos de texto, donde el tamaño de cada palabra en la nube representa su frecuencia relativa.

4. Matplotlib: Matplotlib es una biblioteca de visualización de datos en Python. Se utiliza para crear gráficos y visualizaciones, y en este caso, se utiliza para mostrar las nubes de palabras generadas por WordCloud.

Estas bibliotecas se integran eficazmente en el entorno de programación de Python y proporcionan herramientas esenciales para realizar análisis de texto y generar visualizaciones significativas. Al instalarlas con el comando ``pip install``, se asegura que estén disponibles en el entorno de desarrollo.

4.2.-Desarrollo

En esta sección, se abordará en detalle el código esencial para llevar a cabo, de principio a fin, el análisis de sentimientos. Desde la fase de extracción, pasando por la normalización y estandarización, hasta la generación de gráficos necesarios para visualizar de manera efectiva los hallazgos y la información valiosa descubierta.

4.2.1.- Programación de extracción de los comentarios de las redes sociales

Facebook: La "Programación de extracción de comentarios de Facebook" implica el desarrollo de un conjunto de instrucciones informáticas para automatizar la obtención de comentarios en la plataforma social. Los siguientes imports son componentes esenciales de esta programación.

import time: Módulo que proporciona funciones relacionadas con el tiempo, útil para introducir pausas y gestionar el tiempo en la ejecución del programa.

import pandas as pd: Biblioteca para manipulación y análisis de datos. Se utiliza comúnmente para estructurar y trabajar con conjuntos de datos tabulares.

import os: Módulo que ofrece una interfaz con el sistema operativo, facilitando operaciones relacionadas con archivos y directorios.

from selenium import webdriver: Selenium es una herramienta para la automatización de navegadores web. `webdriver` crea una instancia del navegador y `By` proporciona estrategias de localización de elementos en la página web.

from selenium.webdriver.chrome.service import Service: Un servicio de ChromeDriver para iniciar el navegador Chrome.

from selenium.webdriver.support.ui import WebDriverWait: Utilizados para esperar a que ciertas condiciones se cumplan antes de continuar con la ejecución del programa.

import facebook_scraper as fs: Una biblioteca que facilita la extracción de datos de Facebook, incluyendo comentarios y publicaciones, utilizando técnicas de web scraping.

from fake_useragent import UserAgent: Genera objetos User-Agent que pueden ser utilizados para simular diferentes navegadores y sistemas operativos durante la extracción de datos, ayudando a evitar bloqueos o restricciones.

Figura 26

Librerías necesarias para extraer datos

```
In [3]: #FACEBOOK EXTRACCIÓN DE COMENTARIOS Y TITULOS DE LAS PUBLICACIONES v3 (oficial)
import time
import pandas as pd
import os
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.chrome.service import Service
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
import facebook_scraper as fs
from fake_useragent import UserAgent

ua = UserAgent()
headers = {'User-Agent': ua.random}
```

Nota. Elaboración propia.

Descripción de las funciones para la extracción de comentarios de Facebook.

Función de extracción de enlaces de facebook.

La implementación eficiente de una estrategia de extracción de comentarios en la plataforma Facebook requiere la utilización de funciones específicas que permitan recopilar datos de manera efectiva y precisa. A continuación, se presenta una descripción detallada de las funciones diseñadas para este propósito, delineando sus roles y contribuciones en el proceso de adquisición de información relevante desde la red social. Estas funciones se han creado con el objetivo de proporcionar un enfoque estructurado y automatizado para la obtención sistemática de comentarios, facilitando así un análisis significativo y completo de la interacción en la página de Facebook de interés.

Esta función, denominada "extraer_enlaces_y_ids", está diseñada para obtener enlaces e IDs de las publicaciones en una página de Facebook. A continuación, se describe paso a paso lo que hace la función:

Figura 27

Código de extracción de datos en Facebook

```
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3
# Función para extraer enlaces y IDs de las publicaciones
def extraer_enlaces_y_ids(cantidad_deseada):
    # Crea un objeto Service y pasa la ruta del controlador como argumento
    service = Service("C:/Users/sebastian/Desktop/chromedriver-win64/chromedriver.exe")

    # Inicializa el navegador usando el objeto Service
    driver = webdriver.Chrome(service=service)

    # Abre la página web que deseas explorar
    driver.get("https://www.facebook.com/uplacomunica") # Reemplaza con la URL de tu elección

    # Espera un tiempo para que se cargue la página completamente
    time.sleep(5) # Espera 5 segundos

    ids_recuperados = [] # Lista para almacenar los IDs recuperados
    links_recuperados = [] # Lista para almacenar los enlaces recuperados

    # Bucle para recuperar enlaces y IDs hasta alcanzar la cantidad deseada
    while len(ids_recuperados) < cantidad_deseada:
        time.sleep(2) # Espera 2 segundos
        driver.execute_script("window.scrollTo(0, document.body.scrollHeight);")

        enlaces_con_clase = driver.find_elements(By.XPATH, "//a[contains(@class, 'x1i10hfl xjbqbw x6umtig x1b1mbwd xaqea5y xav7g")

        # Itera a través de los elementos y obtén directamente el atributo href (obtención de los links de las publicaciones)
        for enlace_elemento in enlaces_con_clase:
            url = enlace_elemento.get_attribute("href")
            if url and url not in links_recuperados:
                # Almacena el enlace completo en la lista de enlaces recuperados
                links_recuperados.append(url)

                # Extraer el ID del enlace y almacenarlo en la lista de IDs
                post_id = url.split('/')[-1]
                ids_recuperados.append(post_id)

            # Si has recopilado suficientes enlaces e IDs, sal del bucle
            if len(ids_recuperados) >= cantidad_deseada:
                break

    # Cierra el navegador
    driver.quit()

    # Retorna las listas de IDs y enlaces de las publicaciones
    return ids_recuperados[:cantidad_deseada]
```

Nota. Elaboración propia.

- **Inicialización del Navegador:** Se crea un objeto `Service` y se inicializa el navegador Chrome utilizando el controlador (`chromedriver.exe`) en una ubicación específica.
- **Acceso a la Página Web:** Se abre la página web de Facebook especificada en la función (`https://www.facebook.com/uplacomunica`).
- **Espera de Carga de Página:** Se introduce un tiempo de espera de 5 segundos (`time.sleep(5)`) para permitir la carga completa de la página.
- **Recuperación de Enlaces e IDs:** Se inicia un bucle que se ejecuta hasta alcanzar la cantidad deseada de enlaces e IDs (`cantidad\_deseada`). Se realiza un desplazamiento hacia abajo en la página para cargar más publicaciones mediante `driver.execute\_script("window.scrollTo(0, document.body.scrollHeight);")`. Se identifican enlaces con una clase específica mediante un XPath y se almacenan en `enlaces\_con\_clase`.

- **Iteración a través de los enlaces:** Se itera a través de los elementos de enlace y se obtiene el atributo `href`, que contiene el enlace completo de la publicación en Facebook.
- **Almacenamiento de enlaces e IDs:** Se verifica si el enlace no ha sido previamente recuperado, y si es así, se almacena tanto el enlace como el ID de la publicación en listas respectivas (`links\_recuperados` e `ids\_recuperados`). Se extrae el ID de la publicación a partir del enlace y se almacena en la lista de IDs.
- **Finalización del Bucle:** El bucle se detiene una vez que se recopila la cantidad deseada de enlaces e IDs.
- **Cierre del Navegador:** Se cierra el navegador para liberar recursos.
- **Retorno de Resultados:** La función retorna las listas de IDs y enlaces de las publicaciones, limitadas a la cantidad deseada.

Esta función utiliza Selenium y técnicas de web scraping para extraer enlaces e IDs de las publicaciones en una página de Facebook, proporcionando así una herramienta útil para la recopilación de datos específicos.

Función de extracción de comentarios de facebook parte uno.

El siguiente fragmento de código define una función denominada "extraer_comentarios_de_publicaciones" que tiene como propósito principal la obtención de comentarios de publicaciones en Facebook. Aquí está una descripción paso a paso de lo que hace:

Código de extracción de datos en Facebook uno

Nota. Elaboración propia.

- **Inicialización y Carga de Datos Existentes:** La función toma dos parámetros: 'ids_recuperados', que son los IDs de las publicaciones a procesar, y 'datos_existente', que representa datos previamente recolectados (si los hay). Imprime en la consola información sobre los datos existentes antes de comenzar el procesamiento.
- **Inicialización de Listas y Conjuntos:** 'comentarios_data': Lista destinada a almacenar datos de comentarios y respuestas. 'comentarios_unicos': Conjunto que almacena comentarios únicos para evitar duplicados.
- **Carga de Comentarios y Post IDs Existentes:** Si hay datos existentes, carga comentarios y post IDs existentes en conjuntos para evitar duplicaciones. Imprime información sobre los datos existentes de comentarios.
- **Inicialización del Contador:** 'total_publicaciones_procesadas': Inicializa el contador con el número total de publicaciones procesadas hasta el momento.
- **Creación de Diccionario para Títulos con "N/A":** 'titulos_con_na_a_indices': Un diccionario que mapea títulos con

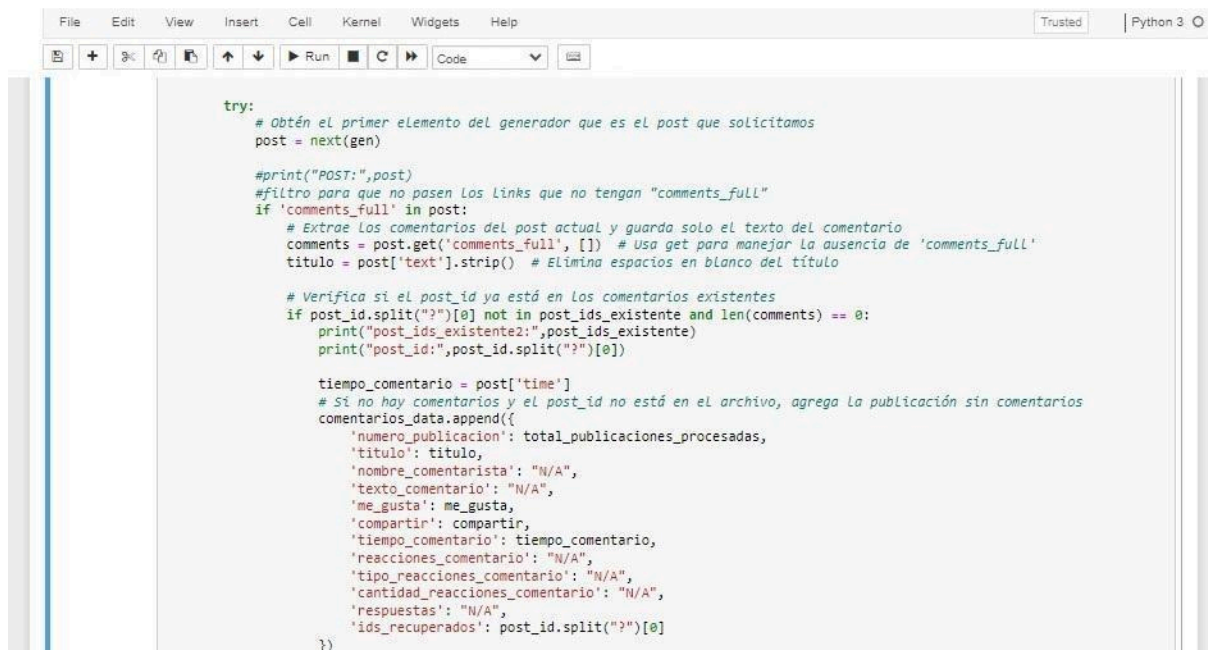
comentarios "N/A" a índices en el DataFrame. Este diccionario se utiliza para manejar publicaciones sin comentarios.

- **Iteración a través de Publicaciones Recuperadas:** Utiliza un bucle para iterar a través de los IDs de publicaciones recuperados (`ids\_recuperados`). Muestra información sobre el progreso de la iteración.
- **Obtención de comentarios con facebook_scraper:** Utiliza `facebook\_scraper` para obtener comentarios de la publicación actual. La función `fs.get\_posts` se encarga de este proceso, especificando la URL del post y opciones como el número máximo de comentarios a recuperar.

Función de extracción de comentarios de facebook parte dos.

Figura 29

Código de extracción de datos en Facebook dos



```
try:
    # Obtén el primer elemento del generador que es el post que solicitamos
    post = next(gen)

    #print("POST:",post)
    #filtro para que no pasen los links que no tengan "comments_full"
    if 'comments_full' in post:
        # Extrae los comentarios del post actual y guarda solo el texto del comentario
        comments = post.get('comments_full', []) # Usa get para manejar la ausencia de 'comments_full'
        titulo = post['text'].strip() # Elimina espacios en blanco del título

        # Verifica si el post_id ya está en los comentarios existentes
        if post_id.split("?")[0] not in post_ids_existente and len(comments) == 0:
            print("post_ids_existente2:",post_ids_existente)
            print("post_id:",post_id.split("?")[0])

        tiempo_comentario = post['time']
        # Si no hay comentarios y el post_id no está en el archivo, agrega la publicación sin comentarios
        comentarios_data.append({
            'numero_publicacion': total_publicaciones_procesadas,
            'titulo': titulo,
            'nombre_comentarista': "N/A",
            'texto_comentario': "N/A",
            'me_gusta': me_gusta,
            'compartir': compartir,
            'tiempo_comentario': tiempo_comentario,
            'reacciones_comentario': "N/A",
            'tipo_reacciones_comentario': "N/A",
            'cantidad_reacciones_comentario': "N/A",
            'respuestas': "N/A",
            'ids_recuperados': post_id.split("?")[0]
        })
    })
```

Nota. Elaboración propia.

Este fragmento de código está dentro de un bloque try y se encarga de procesar la información de cada publicación obtenida mediante la biblioteca facebook_scraper. Aquí hay una descripción de lo que hace:

Intento de Obtener el Primer Elemento del Generador:

Utiliza next(gen) para obtener el primer elemento del generador gen, que representa la información de la publicación de Facebook solicitada.

Filtro de Comentarios Completos:

Verifica si la clave 'comments_full' está presente en la información de la publicación. Si no lo está, significa que la información sobre comentarios completos no está disponible y se omite el procesamiento de esa publicación.

Extracción de Comentarios y Título:

Si la clave está presente, extrae la lista de comentarios de la publicación y el título de la publicación.

El título se obtiene del atributo 'text' de la publicación y se elimina cualquier espacio en blanco adicional.

Verificación de Existencia de Comentarios y Post ID:

Se verifica si el post_id (extraído de la URL) no está presente en el conjunto de post_ids_existente y si la cantidad de comentarios es cero.

Procesamiento de Publicaciones Sin Comentarios:

Si la verificación es exitosa, significa que la publicación no tiene comentarios y no está presente en el conjunto de post IDs existentes.

Agrega la información de la publicación (sin comentarios) a la lista comentarios_data, con valores "N/A" para campos específicos.

Incremento del Contador:

Incrementa el contador total_publicaciones_procesadas para mantener un seguimiento del número total de publicaciones procesadas.

Este fragmento se centra en procesar las publicaciones obtenidas, extrayendo información relevante como comentarios, título y otros detalles. Además, maneja específicamente el caso de publicaciones que no tienen comentarios y no están presentes en el conjunto de datos existentes para evitar duplicados.

Función de extracción de comentarios de facebook parte tres.

Figura 30

Código de extracción de datos en Facebook tres

```
Cell Kernel Widgets Help Trusted Python 3
Run Code
total_publicaciones_procesadas += 1 # Incrementa el contador por cada publicación
#post_ids_existente.add(post_id) # Agrega el post_id a los post_ids_existente
elif len(comments) > 0 and titulo in titulos_con_na_a_indices:
    # Obtiene el índice del título en el DataFrame
    indice = titulos_a_indices[titulo]
    # Elimina la fila correspondiente al título del DataFrame
    datos_existente = datos_existente.drop(indice)
else:
    for idx_comment, comment in enumerate(comments, start=1):
        nombre_comentarista = comment.get('commenter_name', 'Nombre no encontrado')
        texto_comentario = comment.get('comment_text', 'Comentario no encontrado')
        tiempo_comentario = comment.get('comment_time', 'Tiempo no encontrado')
        reacciones_comentario = comment.get('comment_reactors', 'Reacciones no encontradas')
        tipo_reacciones_comentario = comment.get('comment_reactions', 'Reacciones no encontradas')
        cantidad_reacciones_comentario = comment.get('comment_reaction_count', 'Número de reacciones no encontradas')
        respuestas = comment.get('replies', 'Respuestas no encontradas')
        me_gusta = comment.get('likes', 'Me gusta no encontrados')
        compartir = comment.get('shares', 'Compartir no encontrados')

        # Verifica si el comentario ya existe en los comentarios únicos
        if texto_comentario not in comentarios_unicos:
            # Agrega los datos del comentario a la lista comentarios_data junto con el identificador del comentario
            comentarios_data.append({
                'numero_publicacion': total_publicaciones_procesadas,
                'titulo': titulo,
                'nombre_comentarista': nombre_comentarista,
                'texto_comentario': texto_comentario,
                'me_gusta': me_gusta,
                'compartir': compartir,
                'tiempo_comentario': tiempo_comentario,
                'reacciones_comentario': reacciones_comentario,
                'tipo_reacciones_comentario': tipo_reacciones_comentario,
                'cantidad_reacciones_comentario': cantidad_reacciones_comentario,
                'respuestas': [],
                'ids_recuperados': post_id.split("?")[0]
            })
            # Agrega el comentario al conjunto de comentarios únicos
            comentarios_unicos.add(texto_comentario)

    total_publicaciones_procesadas += 1 # Incrementa el contador por cada publicación
```

Nota. Elaboración propia.

Este fragmento de código maneja diferentes escenarios según las condiciones y la información disponible para cada publicación obtenida. Aquí está la descripción de lo que hace:

- **Incremento del Contador y Actualización de Conjunto:** `total_publicaciones_procesadas += 1` Incrementa el contador por cada publicación procesada. Se utiliza para llevar un registro del número total de publicaciones procesadas.
- **Manejo de Publicaciones sin Comentarios:** Si la verificación inicial indica que la publicación no tiene comentarios y no está en el conjunto de ``post_ids_existente``, se agrega el ``post_id`` al conjunto de ``post_ids_existente`` y se incrementa el contador de publicaciones procesadas.
- **Manejo de Publicaciones con Comentarios Existentes:** Si la publicación tiene comentarios (``len(comments) > 0``) y el título de la publicación está en el conjunto de ``titulos_con_na_a_indices``, significa que esta publicación

anteriormente se registró como "N/A" (sin comentarios). Obtiene el índice correspondiente al título en el DataFrame de `datos\_existente` y elimina la fila correspondiente al título. Esto se hace para actualizar la información existente con los comentarios recién obtenidos.

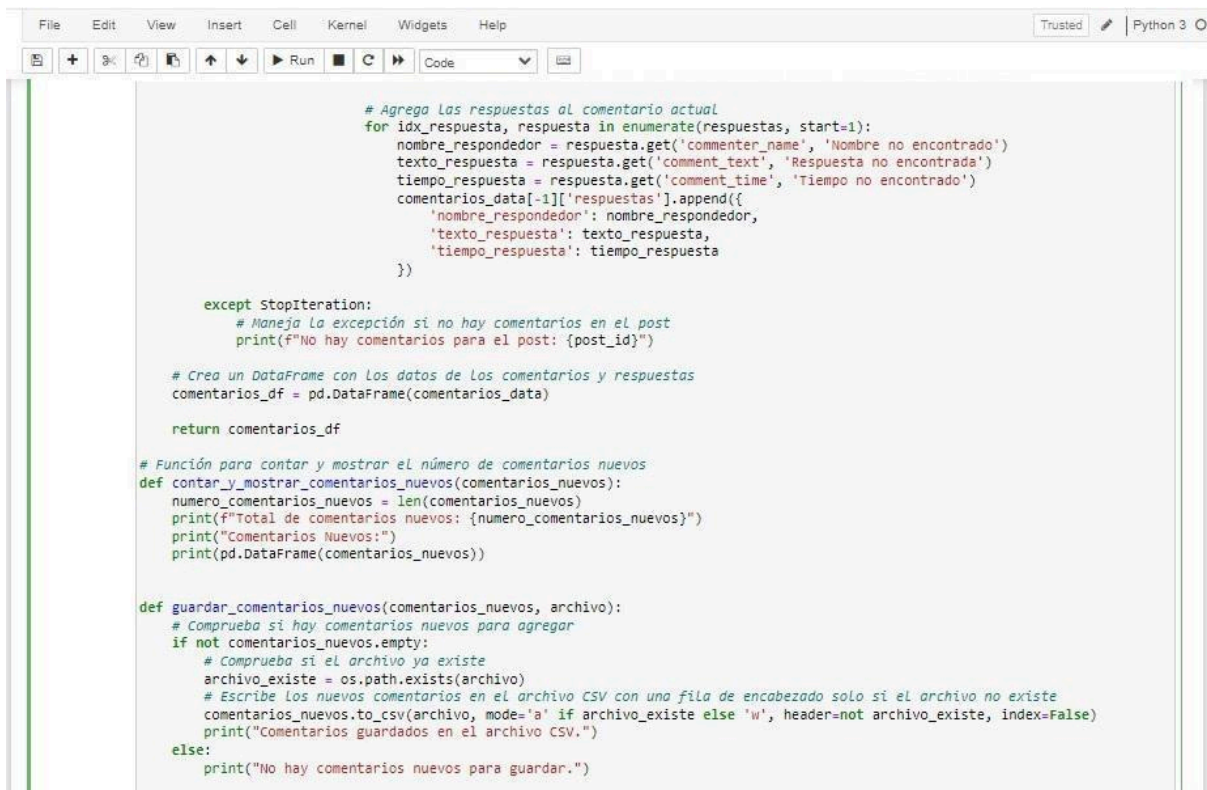
- **Procesamiento de Comentarios Individuales:** Si la publicación tiene comentarios (`len(comments) > 0`), procesa cada comentario individualmente mediante un bucle `for`. Obtiene información como el nombre del comentarista, el texto del comentario, el tiempo del comentario, las reacciones, el tipo de reacciones, la cantidad de reacciones, las respuestas, la cantidad de "me gusta" y la cantidad de veces que se compartió.

Este fragmento de código gestiona la actualización de datos para publicaciones sin comentarios, elimina registros antiguos de publicaciones con comentarios y procesa cada comentario individualmente, asegurándose de evitar duplicados y mantener un seguimiento preciso del número total de publicaciones procesadas.

Función de extracción de comentarios de facebook parte cuatro.

Figura 31

Código de extracción de datos en Facebook cuatro

A screenshot of a Jupyter Notebook interface. The top bar shows tabs for File, Edit, View, Insert, Cell, Kernel, Widgets, and Help. On the right, it says 'Trusted' and 'Python 3'. Below the tabs is a toolbar with icons for running, saving, and other actions. The main area contains a code cell with the following Python code:

```
# Agrega Las respuestas al comentario actual
for idx_respuesta, respuesta in enumerate(respuestas, start=1):
    nombre_respondedor = respuesta.get('commenter_name', 'Nombre no encontrado')
    texto_respuesta = respuesta.get('comment_text', 'Respuesta no encontrada')
    tiempo_respuesta = respuesta.get('comment_time', 'Tiempo no encontrado')
    comentarios_data[-1]['respuestas'].append({
        'nombre_respondedor': nombre_respondedor,
        'texto_respuesta': texto_respuesta,
        'tiempo_respuesta': tiempo_respuesta
    })

except StopIteration:
    # Maneja la excepción si no hay comentarios en el post
    print(f"No hay comentarios para el post: {post_id}")

# Crea un DataFrame con Los datos de Los comentarios y respuestas
comentarios_df = pd.DataFrame(comentarios_data)

return comentarios_df

# Función para contar y mostrar el número de comentarios nuevos
def contar_y_mostrar_comentarios_nuevos(comentarios_nuevos):
    numero_comentarios_nuevos = len(comentarios_nuevos)
    print(f"Total de comentarios nuevos: {numero_comentarios_nuevos}")
    print("Comentarios Nuevos:")
    print(pd.DataFrame(comentarios_nuevos))

def guardar_comentarios_nuevos(comentarios_nuevos, archivo):
    # Comprueba si hay comentarios nuevos para agregar
    if not comentarios_nuevos.empty:
        # Comprueba si el archivo ya existe
        archivo_existe = os.path.exists(archivo)
        # Escribe Los nuevos comentarios en el archivo CSV con una fila de encabezado solo si el archivo no existe
        comentarios_nuevos.to_csv(archivo, mode='a' if archivo_existe else 'w', header=not archivo_existe, index=False)
        print("Comentarios guardados en el archivo CSV.")
    else:
        print("No hay comentarios nuevos para guardar.")
```

Nota. Elaboración propia.

Este conjunto de funciones tiene como objetivo principal procesar y organizar los comentarios y respuestas de las publicaciones de Facebook. Aquí está la descripción de lo que hace el código:

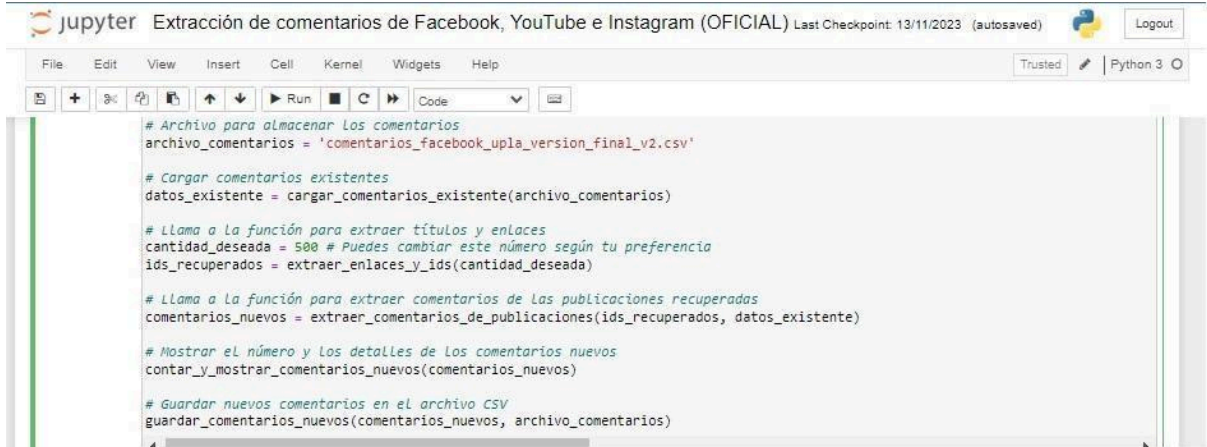
- **Agrega Respuestas al Comentario Actual:** Después de procesar los comentarios individuales de una publicación, este bloque se encarga de agregar las respuestas a este comentario. Utiliza un bucle `for` para iterar a través de las respuestas y extraer información como el nombre del respondedor, el texto de la respuesta y el tiempo de la respuesta. Agrega esta información a la lista de respuestas dentro de la estructura de datos de comentarios.
- **Manejo de la Excepción StopIteration:** Captura la excepción `StopIteration`, que se produce cuando no hay comentarios en el post. En este caso, imprime un mensaje indicando que no hay comentarios para el post.
- **Creación del DataFrame de Comentarios:** Después de procesar todas las publicaciones y comentarios, crea un DataFrame (`comentarios\_df`) utilizando la biblioteca `pandas` con los datos recopilados de comentarios y respuestas.
- **Función Adicional para Contar y Mostrar Comentarios Nuevos:** Se define una función adicional `contar\_y\_mostrar\_comentarios\_nuevos` que toma un DataFrame de comentarios y muestra el número total de comentarios nuevos y una vista del DataFrame.
- **Función para Guardar Comentarios Nuevos en un Archivo CSV:** La función `guardar\_comentarios\_nuevos` verifica si hay comentarios nuevos antes de intentar guardarlos en un archivo CSV. Comprueba si el archivo ya existe y determina si debe escribir una nueva fila de encabezado. Utiliza el método `to\_csv` de `pandas` para escribir los nuevos comentarios en el archivo CSV.

Estas funciones en conjunto ofrecen una manera organizada de procesar, mostrar y guardar comentarios y respuestas de publicaciones de Facebook. La modularidad y claridad en el código facilitan su mantenimiento y comprensión.

Función de extracción de comentarios de facebook parte cinco.

Figura 32

Código de extracción de datos en Facebook cinco



```
jupyter Extracción de comentarios de Facebook, YouTube e Instagram (OFICIAL) Last Checkpoint: 13/11/2023 (autosaved) Logout
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3
# Archivo para almacenar los comentarios
archivo_comentarios = 'comentarios_facebook_upla_version_final_v2.csv'

# Cargar comentarios existentes
datos_existente = cargar_comentarios_existente(archivo_comentarios)

# Llama a la función para extraer títulos y enlaces
cantidad_deseada = 500 # Puedes cambiar este número según tu preferencia
ids_recuperados = extraer_enlaces_y_ids(cantidad_deseada)

# Llama a la función para extraer comentarios de las publicaciones recuperadas
comentarios_nuevos = extraer_comentarios_de_publicaciones(ids_recuperados, datos_existente)

# Mostrar el número y los detalles de los comentarios nuevos
contar_y_mostrar_comentarios_nuevos(comentarios_nuevos)

# Guardar nuevos comentarios en el archivo CSV
guardar_comentarios_nuevos(comentarios_nuevos, archivo_comentarios)
```

Nota. Elaboración propia.

Este fragmento de código realiza varias operaciones relacionadas con la extracción, procesamiento y almacenamiento de comentarios de publicaciones de Facebook. Aquí está una descripción paso a paso:

- **Definición del Archivo de Almacenamiento de Comentarios:** ``archivo_comentarios = 'comentarios_facebook_upla_version_final_v2.csv':` Establece el nombre del archivo CSV donde se almacenarán los comentarios. Puedes cambiar el nombre según tus preferencias.
- **Carga de Comentarios Existente:** ``datos_existente = cargar_comentarios_existente(archivo_comentarios):` Utiliza una función (``cargar_comentarios_existente``) para cargar los comentarios existentes desde el archivo CSV especificado.
- **Extracción de Títulos y Enlaces:** ``cantidad_deseada = 500`:` Establece la cantidad deseada de comentarios a extraer (puedes ajustar este número). ``ids_recuperados = extraer_enlaces_y_ids(cantidad_deseada):` Llama a la función ``extraer_enlaces_y_ids`` para obtener los IDs y enlaces de las publicaciones.
- **Extracción de Comentarios de Publicaciones Recuperadas:** ``comentarios_nuevos = extraer_comentarios_de_publicaciones(ids_recuperados, datos_existente):` Llama a la función ``extraer_comentarios_de_publicaciones`` para obtener los comentarios de las publicaciones recuperadas, considerando los comentarios existentes.

- **Mostrar el Número y Detalles de Comentarios Nuevos:**
`contar\_y\_mostrar\_comentarios\_nuevos(comentarios\_nuevos)`: Utiliza una función (`contar\_y\_mostrar\_comentarios\_nuevos`) para mostrar el número total de comentarios nuevos y una vista de los detalles en un DataFrame.
- **Guardar Nuevos Comentarios en el Archivo CSV:**
`guardar\_comentarios\_nuevos(comentarios\_nuevos, archivo\_comentarios)`: Utiliza una función (`guardar\_comentarios\_nuevos`) para guardar los nuevos comentarios en el archivo CSV especificado.

Este código establece el archivo de almacenamiento, carga los comentarios existentes, extrae títulos y enlaces, obtiene los comentarios de las publicaciones recuperadas, muestra detalles y finalmente guarda los nuevos comentarios en el archivo CSV. Es una secuencia completa de operaciones para gestionar los comentarios de las publicaciones de Facebook.

Descripción de las funciones para la extracción de comentarios de Instagram parte uno.

El import de las bibliotecas para la extracción de datos en instagram fueron las mismas que para Facebook.

Figura 33

Código de extracción de datos en Instagram uno

```
def inicializar_driver():
    servicio = Service('C:/Users/sebastian/Desktop/chromedriver-win64/chromedriver.exe')
    driver = webdriver.Chrome(service=servicio)
    driver.wait = WebDriverWait(driver, 5)
    return driver

def iniciar_sesion_en_instagram(driver, nombre_usuario, contrasena):
    driver.get("https://www.instagram.com/accounts/login/")
    campo_usuario = WebDriverWait(driver, 10).until(EC.presence_of_element_located((By.NAME, 'username')))
    campo_contrasena = driver.find_element(By.NAME, 'password')
    campo_usuario.send_keys(nombre_usuario)
    campo_contrasena.send_keys(contrasena)
    campo_contrasena.send_keys(Keys.RETURN)
    WebDriverWait(driver, 10).until(EC.url_changes("https://www.instagram.com/accounts/login/"))

def hacer_scroll_hasta_abajo(driver):
    # Esta función hace scroll hasta el final de la página
    driver.execute_script("window.scrollTo(0, document.body.scrollHeight);")

def extraer_ids_de_publicaciones(driver, url_perfil, cantidad_publicaciones):
    driver.get(url_perfil)
    time.sleep(5)

    ids_publicaciones = []

    while len(ids_publicaciones) < cantidad_publicaciones:
        # Encontrar todos los elementos <a> dentro del cuerpo de la página usando XPath
        enlaces_publicaciones = driver.find_elements(By.XPATH, '//a[contains(@href, "/p/")]')

        # Extraer los atributos href de los enlaces y obtener los IDs de las publicaciones
        hrefs_publicaciones = [enlace.get_attribute('href') for enlace in enlaces_publicaciones]
        nuevos_ids_publicaciones = [href.split('/')[-2] for href in hrefs_publicaciones if href.split('/')[-2] not in ids_publicaciones]

        # Agregar nuevos IDs de publicaciones a la lista principal
        ids_publicaciones.extend(nuevos_ids_publicaciones)

        # Hacer scroll para cargar más publicaciones
        hacer_scroll_hasta_abajo(driver)

        # Esperar un breve momento para que carguen los nuevos elementos
        time.sleep(2)

    # Devolver solo la cantidad especificada de IDs de publicaciones
    return ids_publicaciones[:cantidad_publicaciones]
```

Nota. Elaboración propia.

Aquí está la explicación de cada función:

- **Inicializar_driver():** Esta función inicializa y retorna una instancia del navegador Chrome mediante Selenium WebDriver. Utiliza un servicio de ChromeDriver (`Service`) y establece un tiempo de espera de 5 segundos (`WebDriverWait`) para las operaciones del navegador. Retorna el objeto del controlador del navegador ya inicializado.
- **Iniciar_sesion_en_instagram(driver, nombre_usuario, contraseña):**
 - Inicia sesión en Instagram utilizando las credenciales proporcionadas.
 - Navega a la página de inicio de sesión de Instagram y espera a que los campos de usuario y contraseña estén presentes.
 - Introduce el nombre de usuario y la contraseña en los campos correspondientes.
 - Envía la tecla "RETURN" para iniciar sesión.
 - Espera hasta que la URL cambia, indicando que se ha completado la acción de inicio de sesión.
- **Hacer_scroll_hasta_abajo(driver):**
 - Esta función realiza un desplazamiento hacia abajo en la página web, simulando la acción de hacer scroll hasta el final de la página.
 - Utiliza el método `execute_script` de Selenium para ejecutar JavaScript y realizar el scroll hasta la parte inferior de la página.
- **Extraer_ids_de_publicaciones(driver,url_perfil, cantidad_publicaciones):**
 - Extrae los IDs de las publicaciones desde un perfil de Instagram.
 - Navega a la URL del perfil proporcionada.
 - Itera hasta que se hayan extraído la cantidad especificada de IDs de publicaciones.
 - Encuentra todos los elementos `<a>` que contienen `/p/` en el atributo href (indicando publicaciones).
 - Extrae los IDs de las publicaciones a partir de los enlaces y los añade a la lista `ids_publicaciones`.
 - Realiza un scroll hacia abajo para cargar más publicaciones.
 - Espera brevemente para permitir la carga de nuevos elementos.
 - Retorna una lista con la cantidad especificada de IDs de publicaciones.

Estas funciones están diseñadas para interactuar con la interfaz web de Instagram utilizando Selenium WebDriver, permitiendo la navegación, inicio de sesión y extracción de IDs de publicaciones desde un perfil específico.

Observación: Es esencial la habilidad para identificar con precisión el elemento que se desea extraer de una página utilizando XPath, un lenguaje de navegación para

Descripción de las funciones para la extracción de comentarios de Instagram parte dos.

Código de extracción de datos en Instagram dos

Nota. Elaboración propia.

124

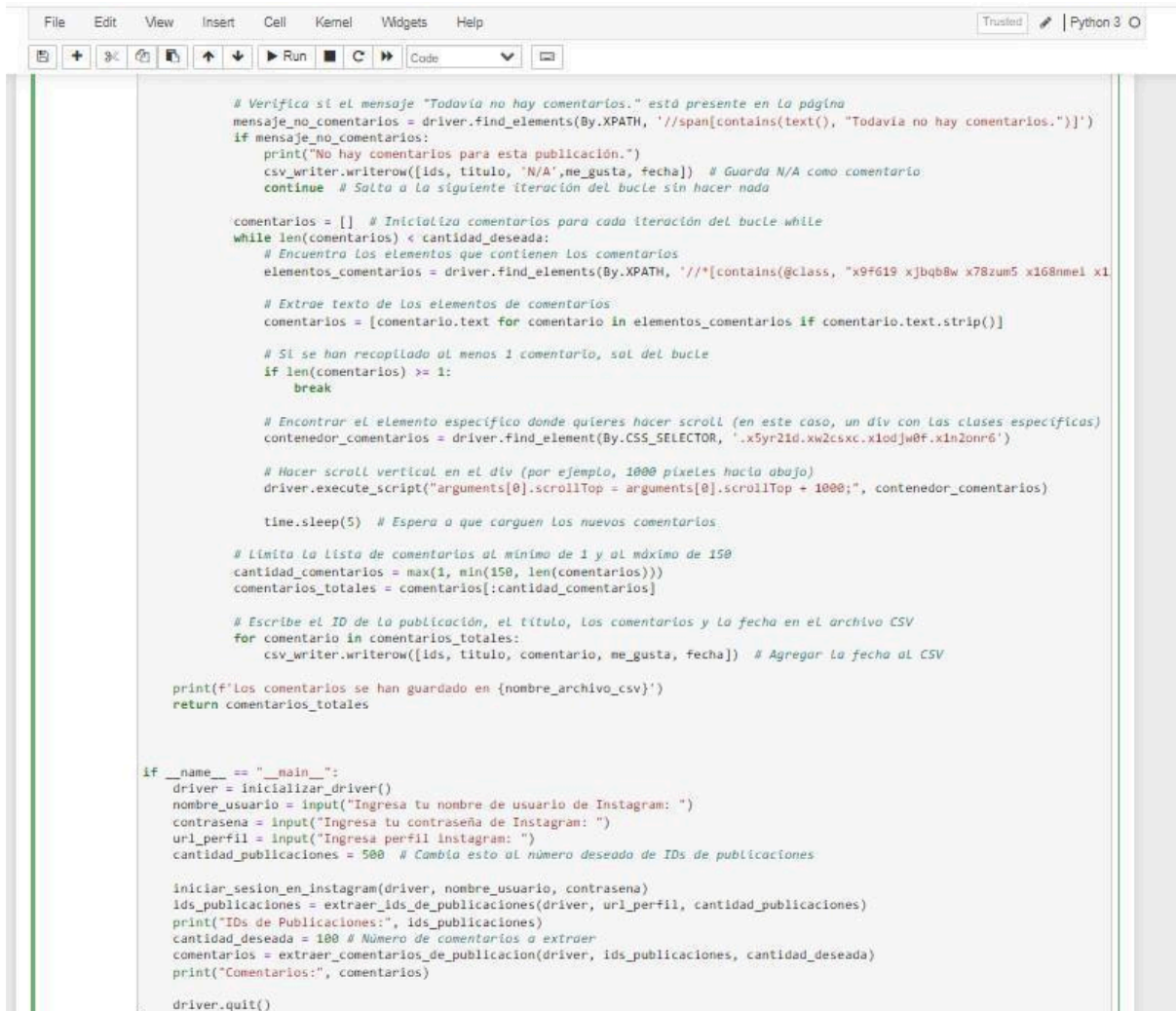
- **Preparación del Archivo CSV:** Abre un archivo CSV para escribir los comentarios y escribe la primera fila con los encabezados.
- **Iteración sobre las Publicaciones:**
 - Itera sobre la lista de IDs de publicaciones proporcionados.
 - Para cada publicación:
 - Espera 5 segundos para asegurar que la página se cargue completamente.
 - Inicializa una lista vacía `comentarios\_totales` para almacenar los comentarios de la publicación actual.
 - Imprime un mensaje indicando la extracción de comentarios para la publicación actual.
- **Obtención del Título de la Publicación:**
 - Accede a la página de la publicación mediante su URL.
 - Busca los elementos que contienen el título de la publicación utilizando XPath.
 - Si encuentra un solo elemento, utiliza el primer XPath; de lo contrario, utiliza el segundo.
 - Extrae el texto del elemento, que representa el título de la publicación.
- **Obtención de la Fecha de la Publicación:**
 - Busca el elemento `` que contiene la información de fecha y hora.
 - Extrae el atributo 'datetime', que contiene la fecha y hora de la publicación.
- **Obtención de la Cantidad de "Me Gusta":**
 - Intenta encontrar el elemento que contiene la cantidad de "Me gusta" utilizando dos XPath diferentes.
 - Extrae el número de "Me gusta" del texto del elemento encontrado.
- **Impresión de Resultados:** Imprime el título de la publicación y el número de "Me gusta" para verificar la extracción correcta.

El código sigue con la extracción de comentarios, pero esta parte específica se enfoca en la obtención del título, la fecha y la cantidad de "Me gusta" de una publicación en Instagram.

Descripción de las funciones para la extracción de comentarios de Instagram parte tres.

Figura 35

Código de extracción de datos en Instagram tres

The image shows a Jupyter Notebook interface with a menu bar (File, Edit, View, Insert, Cell, Kernel, Widgets, Help) and a toolbar with icons for file operations, running, and code execution. The code is written in Python and is designed to extract comments from Instagram. It starts by checking if the message "Todavía no hay comentarios." is present on the page. If it is, it prints a message and continues to the next iteration. If not, it prints the message and continues. It then initializes a list for comments and enters a while loop that continues until the number of comments is less than the desired quantity. Inside the loop, it finds the comment elements using XPath, extracts the text, and appends it to the list. It also checks if the list is not empty and breaks the loop. After the loop, it finds the container for comments, scrolls down by 1000 pixels, and waits for 5 seconds. It then limits the list of comments to a maximum of 150 and writes the data to a CSV file. Finally, it prints the total number of comments and returns the list. The code is wrapped in an if __name__ == "__main__": block, which initializes the driver, prompts for user credentials, and calls the functions to extract IDs and comments.

```
# Verifica si el mensaje "Todavía no hay comentarios." está presente en la página
mensaje_no_comentarios = driver.find_elements(By.XPATH, '//span[contains(text(), "Todavía no hay comentarios.")]')
if mensaje_no_comentarios:
    print("No hay comentarios para esta publicación.")
    csv_writer.writerow([ids, titulo, 'N/A', me_gusta, fecha]) # Guarda N/A como comentario
    continue # Salta a la siguiente iteración del bucle sin hacer nada

comentarios = [] # Inicializa comentarios para cada iteración del bucle while
while len(comentarios) < cantidad_deseada:
    # Encuentra los elementos que contienen los comentarios
    elementos_comentarios = driver.find_elements(By.XPATH, '//*[@contains(@class, "x9f619 xjbqb8w x78zum5 x168nmel x1')
    # Extrae texto de los elementos de comentarios
    comentarios = [comentario.text for comentario in elementos_comentarios if comentario.text.strip()]

    # Si se han recopilado al menos 1 comentario, sal del bucle
    if len(comentarios) >= 1:
        break

    # Encontrar el elemento específico donde quieres hacer scroll (en este caso, un div con las clases específicas)
    contenedor_comentarios = driver.find_element(By.CSS_SELECTOR, '.x5yr21d.xw2csxc.x1odjw0f.x1n2onr6')

    # Hacer scroll vertical en el div (por ejemplo, 1000 pixeles hacia abajo)
    driver.execute_script("arguments[0].scrollTop = arguments[0].scrollTop + 1000;", contenedor_comentarios)

    time.sleep(5) # Espera a que carguen los nuevos comentarios

# Límite la lista de comentarios al mínimo de 1 y al máximo de 150
cantidad_comentarios = max(1, min(150, len(comentarios)))
comentarios_totales = comentarios[:cantidad_comentarios]

# Escribe el ID de la publicación, el título, los comentarios y la fecha en el archivo CSV
for comentario in comentarios_totales:
    csv_writer.writerow([ids, titulo, comentario, me_gusta, fecha]) # Agregar la fecha al CSV

print(f'Los comentarios se han guardado en {nombre_archivo_csv}')
return comentarios_totales

if __name__ == "__main__":
    driver = inicializar_driver()
    nombre_usuario = input("Ingresa tu nombre de usuario de Instagram: ")
    contrasena = input("Ingresa tu contraseña de Instagram: ")
    url_perfil = input("Ingresa perfil Instagram: ")
    cantidad_publicaciones = 500 # Cambia esto al número deseado de IDs de publicaciones

    iniciar_sesion_en_instagram(driver, nombre_usuario, contrasena)
    ids_publicaciones = extraer_ids_de_publicaciones(driver, url_perfil, cantidad_publicaciones)
    print("IDs de Publicaciones:", ids_publicaciones)
    cantidad_deseada = 100 # Número de comentarios a extraer
    comentarios = extraer_comentarios_de_publicacion(driver, ids_publicaciones, cantidad_deseada)
    print("Comentarios:", comentarios)

    driver.quit()
```

Nota. Elaboración propia.

Este bloque de código está diseñado para extraer comentarios de publicaciones en Instagram, teniendo en cuenta la posibilidad de que no haya comentarios o que la página sea dinámica. Aquí está la descripción del código:

- **Verificación de la Existencia de Comentarios:** Utiliza XPath para encontrar el mensaje "Todavía no hay comentarios". Si el mensaje está presente, imprime un mensaje indicando que no hay comentarios para esa publicación y guarda información en el archivo CSV (marca 'N/A' como comentario).
- **Extracción de Comentarios:**
 - Utiliza un bucle `while` para esperar y recopilar comentarios.

- Busca los elementos que contienen los comentarios mediante XPath.
 - Extrae el texto de los elementos de comentarios.
 - Si se han recopilado al menos 1 comentario, sale del bucle.
 - Si no se han recopilado comentarios, hace scroll en un contenedor específico para cargar más comentarios.
- **Escritura en el Archivo CSV:** Limita la cantidad de comentarios recopilados al mínimo de 1 y al máximo de 150. Escribe el ID de la publicación, el título, los comentarios y la cantidad de "Me gusta" en el archivo CSV.
 - **Iteración sobre Múltiples Publicaciones:**
 - Utiliza un bucle para iterar sobre varias publicaciones.
 - Inicializa el driver y realiza el inicio de sesión en Instagram.
 - Extrae los IDs de las publicaciones y posteriormente extrae comentarios para cada publicación.
 - Imprime los comentarios recopilados y guarda la información en el archivo CSV.

Este código, cuando se ejecuta, solicitará al usuario su nombre de usuario y contraseña de Instagram, la URL del perfil de Instagram y el número deseado de comentarios e IDs de publicaciones. Luego, realiza la extracción de comentarios y muestra los resultados.

Descripción de las funciones para la extracción de comentarios de YouTube parte uno.

Figura 36

Código de extracción de datos en YouTube uno

```
YouTube

# yutu v2.5
from googleapiclient.discovery import build
import googleapiclient.errors
import csv

# Clave de la API de YouTube
clave_api = 'AIzaSyCvbicyzII27Kx9hNNsLU8h7dPwc1bZcpY' # Reemplaza con tu clave API

# Crea una instancia de la API de YouTube Data
youtube = build('youtube', 'v3', developerKey=clave_api)

# Nombre del canal de YouTube
nombre_canal_youtube = 'facultaddeingenieriaupla' # Reemplaza con el nombre del canal que deseas procesar

try:
    # Archivo CSV donde se guardarán los datos
    nombre_archivo_csv = 'comentarios_youtube_upla_version_final_v2.csv'

    with open(nombre_archivo_csv, 'w', newline='', encoding='utf-8') as csvfile:
        # Crea un objeto escritor CSV
        csv_writer = csv.writer(csvfile)

        # Escribe la cabecera del archivo CSV
        csv_writer.writerow(
            ['ID del video', 'Titulo del video', 'Autor', 'Publicado', 'Likes', 'Compartir', 'Visitas', 'Comentario'])

        # Realiza una solicitud para obtener los últimos videos del canal
        respuesta_videos = youtube.search().list(
            part='id,snippet',
            type='channel',
            q=nombre_canal_youtube,
            maxResults=1 # Limitamos a 1 resultado ya que estamos buscando un canal específico
        ).execute()

        # Extraer el ID del canal si se encuentra
        if 'items' in respuesta_videos:
            id_canal = respuesta_videos['items'][0]['id']['channelId']

            # Realiza una solicitud para obtener los últimos videos del canal por su ID
            videos_response = youtube.search().list(
                part='id,snippet',
                channelId=id_canal,
                maxResults=5000, # Puedes ajustar este número según tus necesidades
                order='date' # Obtener los videos más recientes
            ).execute()
```

Nota. Elaboración propia.

Este código está diseñado para interactuar con la API de YouTube y obtener información sobre los videos de un canal específico. Aquí está la descripción del código:

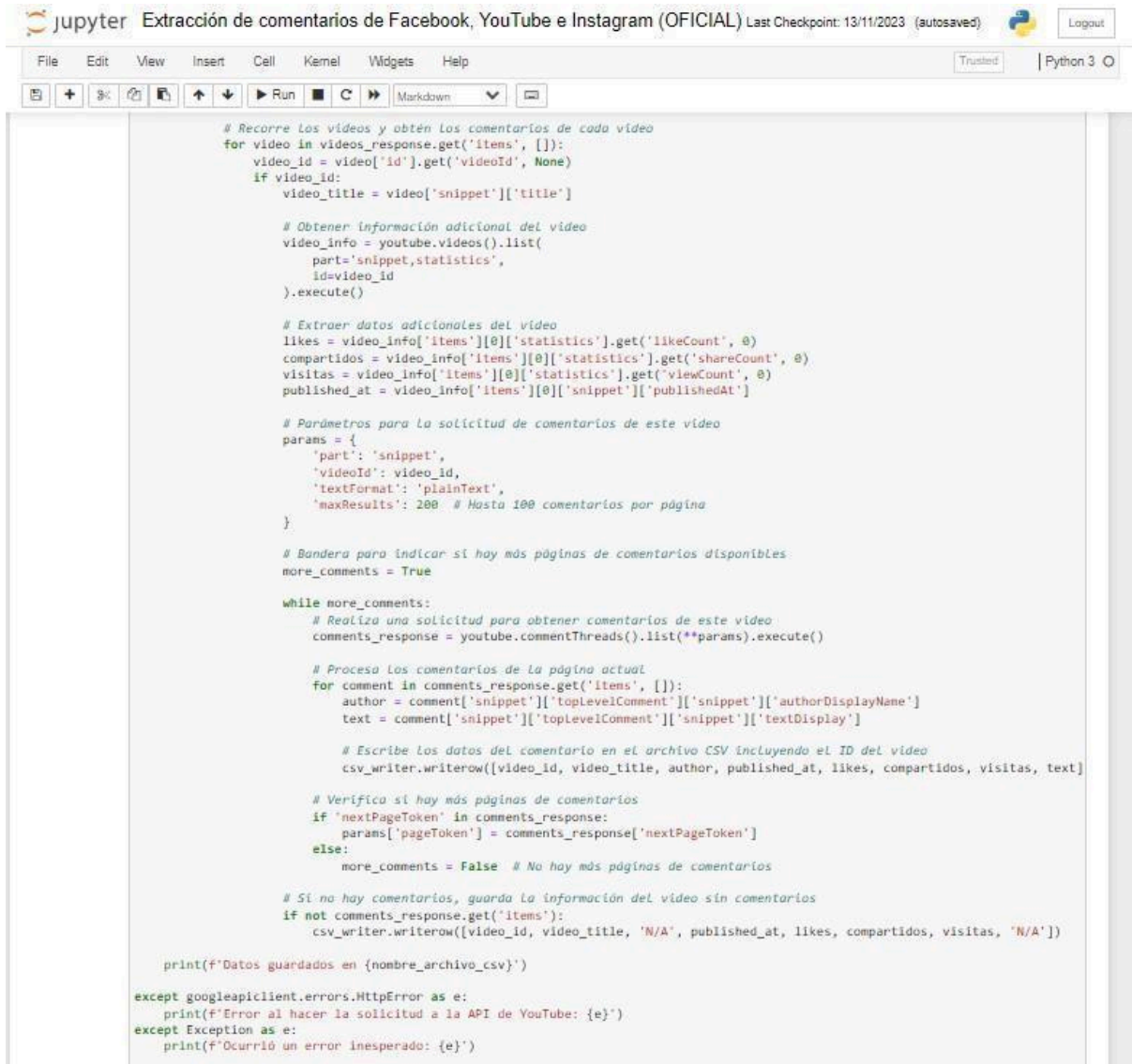
- **Importación de Módulos:** Importa los módulos necesarios de la biblioteca `googleapiclient` para trabajar con la API de YouTube y la manipulación de archivos CSV.
- **Clave de la API de YouTube:** Define la clave de la API de YouTube necesaria para realizar solicitudes a la API.
- **Creación de una Instancia de la API de YouTube Data:**
 - Utiliza la función `build` para crear una instancia de la API de YouTube Data.
 - Utiliza la clave de la API como `developerKey` en la instancia.

- **Nombre del Canal de YouTube:** Define el nombre del canal de YouTube que se desea procesar. En este caso, el nombre del canal es 'facultaddeingenieriaupla'.
- **Manejo de Errores:** Utiliza un bloque ``try`` para manejar posibles excepciones que puedan surgir durante la ejecución del código.
- **6. Archivo CSV:**
 - Define el nombre del archivo CSV donde se guardarán los datos, en este caso, 'comentarios_youtube_upla_version_final_v2.csv'.
 - Abre el archivo en modo de escritura ('w') y crea un objeto escritor CSV.
- **Escritura de la Cabecera del Archivo CSV:** Escribe la primera fila del archivo CSV que contendrá los encabezados de las columnas.
- **Obtención de los Últimos Videos del Canal:**
 - Realiza una solicitud a la API de búsqueda (``youtube.search().list``) para obtener los últimos videos de un canal específico (``type='channel'``).
 - Utiliza el nombre del canal (``nombre_canal_youtube``) para la búsqueda y limita los resultados a 1.
 - Extrae el ID del canal si se encuentra.
- **Obtención de los Detalles de los Videos:**
 - Utiliza el ID del canal para realizar una segunda solicitud para obtener los últimos videos del canal.
 - Limita los resultados a 5000 videos y ordena por fecha (``order='date'``).
 - Ejecuta la solicitud y obtiene la respuesta (``videos_response``).

Descripción de las funciones para la extracción de comentarios de YouTube parte dos.

Figura 37

Código de extracción de datos en YouTube dos



```
# Recorre los videos y obtén los comentarios de cada video
for video in videos_response.get('items', []):
    video_id = video['id'].get('videoId', None)
    if video_id:
        video_title = video['snippet']['title']

        # Obtener información adicional del video
        video_info = youtube.videos().list(
            part='snippet,statistics',
            id=video_id
        ).execute()

        # Extraer datos adicionales del video
        likes = video_info['items'][0]['statistics'].get('likeCount', 0)
        compartidos = video_info['items'][0]['statistics'].get('shareCount', 0)
        visitas = video_info['items'][0]['statistics'].get('viewCount', 0)
        published_at = video_info['items'][0]['snippet']['publishedAt']

        # Parámetros para la solicitud de comentarios de este video
        params = {
            'part': 'snippet',
            'videoId': video_id,
            'textFormat': 'plainText',
            'maxResults': 200 # Hasta 100 comentarios por página
        }

        # Bandera para indicar si hay más páginas de comentarios disponibles
        more_comments = True

        while more_comments:
            # Realiza una solicitud para obtener comentarios de este video
            comments_response = youtube.commentThreads().list(**params).execute()

            # Procesa los comentarios de la página actual
            for comment in comments_response.get('items', []):
                author = comment['snippet']['topLevelComment']['snippet']['authorDisplayName']
                text = comment['snippet']['topLevelComment']['snippet']['textDisplay']

                # Escribe los datos del comentario en el archivo CSV incluyendo el ID del video
                csv_writer.writerow([video_id, video_title, author, published_at, likes, compartidos, visitas, text])

            # Verifica si hay más páginas de comentarios
            if 'nextPageToken' in comments_response:
                params['pageToken'] = comments_response['nextPageToken']
            else:
                more_comments = False # No hay más páginas de comentarios

        # Si no hay comentarios, guarda la información del video sin comentarios
        if not comments_response.get('items'):
            csv_writer.writerow([video_id, video_title, 'N/A', published_at, likes, compartidos, visitas, 'N/A'])

        print(f'Datos guardados en {nombre_archivo_csv}')

except googleapiclient.errors.HttpError as e:
    print(f'Error al hacer la solicitud a la API de YouTube: {e}')
except Exception as e:
    print(f'Ocurrió un error inesperado: {e}')
```

Nota. Elaboración propia.

Este código interactúa con la API de YouTube para obtener información sobre los comentarios de los videos de un canal específico y guarda esos datos en un archivo CSV. Aquí está una descripción del código:

- **Iteración sobre los Videos:**

- Utiliza un bucle `for` para iterar sobre los videos obtenidos en la respuesta de la API (`videos\_response`).
- Extrae información clave de cada video, como el ID del video (`video\_id`), el título (`video\_title`), la fecha de publicación

(`published_at`), la cantidad de likes (`likes`), la cantidad de compartidos (`compartidos`), y la cantidad de visitas (`visitas`).

- **Solicitud de Comentarios:**

- Utiliza el ID del video para realizar una solicitud adicional (`youtube.commentThreads().list`) para obtener los comentarios asociados a ese video.
- Utiliza un bucle `while` para manejar la paginación de los comentarios, ya que cada solicitud puede devolver hasta 100 comentarios y hay múltiples páginas de comentarios.

- **Procesamiento de Comentarios:**

- Procesa los comentarios obtenidos en cada página y extrae información como el autor (`author`) y el texto del comentario (`text`).
- Escribe los datos del comentario, junto con la información del video, en el archivo CSV usando el objeto escritor CSV.

- **Manejo de Situaciones sin Comentarios:** Verifica si no hay comentarios (`not comments_response.get('items')`) y, en ese caso, guarda la información del video con comentarios marcados como 'N/A' en el archivo CSV.

- **Manejo de Errores:** Utiliza bloques `except` para manejar posibles errores, como errores de solicitud HTTP (`googleapiclient.errors.HttpError`) y otros errores generales. Imprime mensajes de error correspondientes en caso de cualquier problema.

- **Impresión del Mensaje de Éxito:** Imprime un mensaje indicando que los datos se han guardado exitosamente en el archivo CSV especificado (`nombre_archivo_csv`).

Unificando los comentarios en un nuevo archivo para poder facilitar el análisis.

Es crucial poder contar con un solo archivo en el cual se pueda trabajar con todos los datos recopilados y unificados, normalizando y estandarizando sus datos para poder obtener una mayor congruencia en los datos.

Figura 38

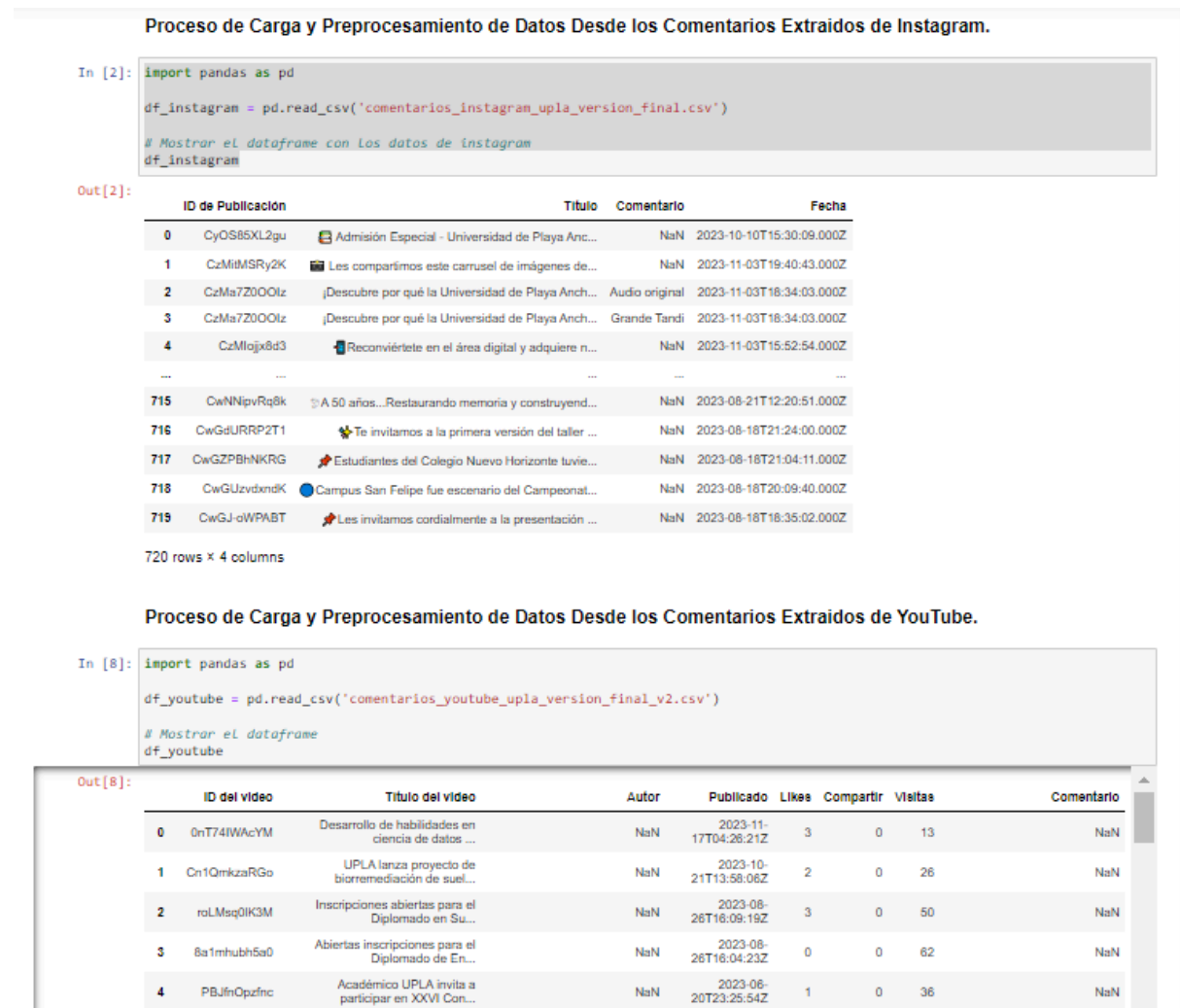
Unificación de los datos



Nota. Elaboración propia.

Figura 39

Unificación de los datos



Nota. Elaboración propia.

Llamado de Archivos y Consolidación en un DataFrame Único

En esta sección, se procede a la integración y manipulación de los datos provenientes de distintas plataformas de redes sociales: Facebook, Instagram y YouTube. Se emplea la librería Pandas en Python para cargar, transformar y combinar los datos de cada plataforma. A continuación, se describe el proceso paso a paso:

- **Carga de Datos Individuales:**

- Facebook: Se carga el archivo 'comentarios_facebook_upla_version_final.csv' en un DataFrame llamado 'df_facebook'.
- Instagram: Se carga el archivo 'comentarios_instagram_upla_version_final.csv' en un DataFrame llamado 'df_instagram'.

- YouTube Se carga el archivo 'comentarios_youtube_upla_version_final.csv' en un DataFrame llamado `df\_youtube`.
- **Eliminar Columna Irrelevante en Facebook:** Se elimina la columna 'numero_publicacion' del DataFrame `df\_facebook`, ya que se considera irrelevante para el análisis. La modificación se realiza in situ mediante `inplace=True`.
- **Renombrar Columnas para Consistencia:** Se renombran las columnas de cada DataFrame para lograr consistencia en la nomenclatura. Esto asegura uniformidad en los datos a lo largo de todas las plataformas.
- **Asignación de Columna 'Fuente':** Se crea una nueva columna llamada 'Fuente' en cada DataFrame (df_facebook , df_instagram , df_youtube) para identificar el origen de cada comentario (Facebook, Instagram o YouTube).
- **Unión de DataFrames:** Se combinan los tres DataFrames en uno solo llamado `df\_total` utilizando la función `pd.concat()`. La opción `ignore\_index=True` reindexa el DataFrame resultante.
- **Guardado del DataFrame Consolidado:** El DataFrame consolidado (`df\_total`) se guarda en un nuevo archivo CSV llamado 'comentarios_redes_sociales_upla.csv'. El parámetro `index=False` evita que se añadan índices adicionales al archivo CSV.
- **Verificación de los Primeros Registros:** Se muestran las primeras filas del DataFrame consolidado (`df\_total`) para verificar y asegurar la correcta combinación de datos.

Este proceso sienta las bases para análisis posteriores y proporciona un conjunto de datos unificado que puede ser explorado y utilizado para investigaciones más avanzadas. La columna 'Fuente' facilita la identificación del origen de cada comentario en el análisis subsiguiente.

4.2.2.- Estandarización y Normalización del Conjunto de Datos: Un Paso Esencial para el Análisis

En el ámbito del análisis de datos, la calidad y coherencia de los conjuntos de datos son elementos cruciales que afectan directamente la efectividad de cualquier análisis subsiguiente. El proceso de estandarización y normalización se convierte, por lo tanto, en un paso esencial para garantizar que los datos estén limpios, coherentes y listos para ser sometidos a evaluaciones precisas.

1. Eliminación de Comentarios Irrelevantes: Antes de proceder con cualquier análisis, es imperativo depurar el conjunto de datos de cualquier información que carezca de relevancia o que haya sido extraída de manera innecesaria. Esta limpieza no solo simplifica el conjunto de datos, sino que también mejora la eficacia de los análisis subsiguientes al centrarse en la información verdaderamente significativa.

2. Normalización de Texto en Minúsculas: Para garantizar la consistencia en el análisis de texto, la normalización de todas las cadenas de texto a minúsculas se vuelve esencial. Esto asegura que las comparaciones y búsquedas en el conjunto de datos se realicen de manera uniforme, evitando posibles incongruencias debidas a las variaciones en el uso de mayúsculas y minúsculas.

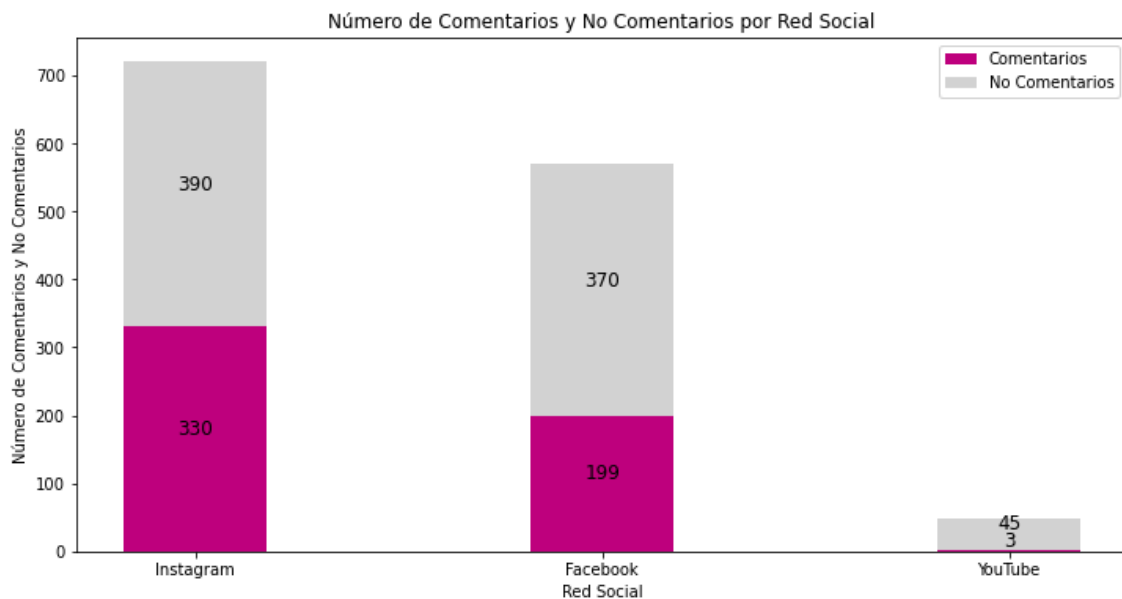
3. Normalización de Fecha y Hora: La normalización de las fechas y horas en el conjunto de datos es crucial para establecer una representación coherente del tiempo. Este proceso unifica formatos dispares, como "2023-11-03 12:15:03" y "2023-10-10T15:30:09.000Z", en un formato estandarizado. Esto facilita la manipulación y comparación temporal durante el análisis.

4. Manejo de Valores Nulos: La presencia de valores nulos en un conjunto de datos puede impactar significativamente en el análisis. La estrategia para manejar estos valores puede incluir la eliminación de filas con nulos, la imputación de valores basada en estadísticas descriptivas, o la aplicación de métodos más avanzados según las necesidades específicas del análisis.

La estandarización y normalización del conjunto de datos son prácticas esenciales para garantizar la calidad y coherencia de los datos antes de someterlos a cualquier análisis más profundo. Estos procesos proporcionan una base sólida para la obtención de resultados precisos y valiosos en el ámbito del análisis de datos.

Figura 40

Cantidad de comentarios y no comentarios por red social



Nota. Elaboración propia.

El gráfico de barras presenta la distribución del número de comentarios y no comentarios después de la normalización de datos en tres redes sociales principales: Facebook, Instagram y YouTube. Este gráfico de barras ofrece una representación clara y concisa de la distribución del número de comentarios después del proceso de normalización, permitiendo una rápida identificación de las diferencias en la participación de usuarios entre las diversas redes sociales analizadas.

4.2.3.- Desarrollo del análisis

Extracción general del conjunto de datos a utilizar

En esta sección, se procederá a realizar la extracción y procesamiento de un conjunto de datos relevante. Para el propósito de este trabajo, se ha utilizado un conjunto de datos provenientes de redes sociales, que incluye títulos y comentarios relacionados con una institución educativa específica, en este caso, la Universidad Playa Ancha (UPLA).

Proceso de Tokenización y Análisis de Tópicos

Para llevar a cabo el proceso de tokenización se mostrará una parte del dataframe que se usará para llevar a cabo el análisis.

Figura 41

Cantidad de comentarios y no comentarios por red social

	Título	Autor	Comentario	Publicado	Reacciones	Tipo de Reacciones	Cantidad de Reacciones	Respuestas	ID de Publicación	Fuente	Me gusta	Compartir	Visitas
0	El Observatorio de Participación Social y Ter...	NaN	NaN	2023-11-03 12:15:03	NaN	NaN	NaN	NaN	pfbid0eXv5U5dxNb5B13MdSTeogI55ZpFBA3qjd477VwHf...	Facebook	NaN	NaN	NaN
1	Admisión Especial - Universidad de Playa Anc...	NaN	NaN	2023-11-03 11:50:04	NaN	NaN	NaN	NaN	pfbid02Ec5CkbCw4aKoVzW/Xom1BmqEzK7i61F7kyRwFW...	Facebook	NaN	NaN	NaN
2	Convocatoria: Cuartas Jornadas de Estudios s...	NaN	NaN	2023-11-03 11:23:03	NaN	NaN	NaN	NaN	pfbid0A83ZmGSi3dd5yAGkUZ5qKbgHFwQ3BNXxmqJLKWB4...	Facebook	NaN	NaN	NaN
3	Reconviértete en el área digital y adquiere n...	Andres Andres Leiva	Luis Campos	NaN		NaN	NaN		pfbid0JPMJPbEGd4nYTLZ482Sj3MMchWgRQnhHWYUqSx...	Facebook	NaN	NaN	NaN
4	Doctorado Consorciado en Educación/inTe pre...	NaN	NaN	2023-11-03 10:10:43	NaN	NaN	NaN	NaN	pfbid0248eSA7JMdYvjuf71n5L2x3faZM7fuHRJ1v5jo5...	Facebook	NaN	NaN	NaN
1332	"Importancia de los servicios ecosistémic...	NaN	NaN	2021-06-12T02:17:00Z	NaN	NaN	NaN	NaN	Gx8sZuIPZPk	YouTube	5.0	0.0	53.0
1333	"Cambio climático: causas y principales e...	NaN	NaN	2021-06-12T01:53:54Z	NaN	NaN	NaN	NaN	d4O3iwrvtbKQA	YouTube	1.0	0.0	27.0
1334	Charla "Contaminación en zonas costeras&qu...	NaN	NaN	2021-06-12T01:27:44Z	NaN	NaN	NaN	NaN	wEaMpoA1q78	YouTube	0.0	0.0	21.0
1335	"Descubrimientos accidentales en la cienc...	NaN	NaN	2021-06-11T23:35:52Z	NaN	NaN	NaN	NaN	wMG1yII7E8U	YouTube	0.0	0.0	35.0
1336	"Métodos para el muestreo de diversidad a...	NaN	NaN	2021-06-11T22:19:54Z	NaN	NaN	NaN	NaN	9pWSWMNW_gE	YouTube	0.0	0.0	78.0

1337 rows × 13 columns

Para abordar este análisis, se ha aplicado una técnica de procesamiento de texto conocida como tokenización. El proceso de tokenización implica dividir el texto en unidades más pequeñas llamadas tokens, que pueden ser palabras o frases más pequeñas. En este caso, se utilizó la biblioteca Natural Language Toolkit (nltk) en Python para llevar a cabo la tokenización y el filtrado de palabras irrelevantes, también conocidas como stopwords.

Además, se emplearon modelos de Latent Dirichlet Allocation (LDA) para identificar tópicos dentro de los títulos y comentarios. Los tópicos representan temas recurrentes en el conjunto de datos y permiten clasificar y comprender la información de manera más efectiva.

Figura 42

Código del proceso de tokenización

```
In [18]: import pandas as pd
from gensim.corpora import Dictionary
from gensim.models import LdaModel
import string
from nltk.corpus import stopwords
import nltk
import re

# Función para eliminar caracteres no alfanuméricos (incluyendo emojis)
def eliminar_no_alfanumericos(texto):
    # Utilizar una expresión regular para eliminar caracteres no alfanuméricos
    return re.sub(r'[^a-zA-Z0-9ÑÁÉÍÚÁÊÏÓðáëíóú\']', '', texto)

# Tokenización y eliminación de stopwords
stopwords_espanol = set(stopwords.words('spanish'))

# Crear una copia para el análisis de tópicos
df_analisis_topicos = df_total.copy()

# Procesar títulos
titulos = df_analisis_topicos['Titulo'].fillna('').apply(eliminar_no_alfanumericos)
titulos = titulos.apply(lambda x: x.lower().split())
titulos = titulos.apply(lambda x: [palabra for palabra in x if palabra not in stopwords_espanol])

# Procesar comentarios
comentarios = df_analisis_topicos['Comentario'].fillna('').apply(eliminar_no_alfanumericos)
comentarios = comentarios.apply(lambda x: x.lower().split())
comentarios = comentarios.apply(lambda x: [palabra for palabra in x if palabra not in stopwords_espanol])

# Crear diccionarios y corpora para títulos y comentarios
diccionario_titulos = Dictionary(titulos)
diccionario_comentarios = Dictionary(comentarios)

corpus_titulos = [diccionario_titulos.doc2bow(titulo) for titulo in titulos]
corpus_comentarios = [diccionario_comentarios.doc2bow(comentario) for comentario in comentarios]

# Entrenar modelos LDA para títulos y comentarios
modelo_lda_titulos = LdaModel(corpus_titulos, num_topics=5, id2word=diccionario_titulos, passes=15)
modelo_lda_comentarios = LdaModel(corpus_comentarios, num_topics=5, id2word=diccionario_comentarios, passes=15)

# Imprimir tópicos dominantes para títulos
print("Tópicos para Títulos:")
for idx, topic in modelo_lda_titulos.print_topics(-1):
    print(f'Tópico {idx}: {topic}')

# Imprimir tópicos dominantes para comentarios
print("\nTópicos para Comentarios:")
for idx, topic in modelo_lda_comentarios.print_topics(-1):
    print(f'Tópico {idx}: {topic}')

# Asignar tópicos a cada título y comentario en el DataFrame de análisis
df_analisis_topicos['Topico_Titulos'] = [max(modelo_lda_titulos[corpus_titulos[i]], key=lambda x: x[1])[0] for i in range(len(corpus_titulos))]
df_analisis_topicos['Topico_Comentarios'] = [max(modelo_lda_comentarios[corpus_comentarios[i]], key=lambda x: x[1])[0] for i in range(len(corpus_comentarios))]

# Asignar tópicos a cada título y comentario en el DataFrame original
df_total['Topico_Titulos'] = df_analisis_topicos['Topico_Titulos']
df_total['Topico_Comentarios'] = df_analisis_topicos['Topico_Comentarios']
```

A continuación, se presentan los tópicos dominantes identificados en los títulos y comentarios:

Tópicos para los títulos de las publicaciones.

- **Artes y Reconocimientos:** Este tópico parece relacionarse con logros y reconocimientos en el ámbito de las artes, con menciones específicas a la universidad y premios. Puede indicar un enfoque en logros artísticos y reconocimientos relacionados con la institución educativa, promoviendo eventos y éxitos en el campo de las artes.
- **Historia y Memoria:** Se aborda la historia y la memoria, con referencias a períodos específicos de tiempo y elementos relacionados con la institución.

Sugiere discusiones sobre la historia de la institución y eventos pasados, mostrando la importancia de la memoria histórica en las conversaciones.

- **Educación y Estudiantes:** Enfocado en temas de educación y el impacto en los estudiantes, con menciones a la universidad y la enseñanza. Indica discusiones centradas en la educación y su impacto en la comunidad estudiantil, abordando temas relevantes para los alumnos.
- **Actividades y Disciplinas Físicas:** Relacionado con actividades y disciplinas específicas, como actividad física, kinesiólogía y docencia. Muestra una diversidad de temas, desde actividades académicas hasta disciplinas específicas, reflejando la amplitud de la oferta educativa.
- **Género y Ciencias:** Aborda temas de género y su relación con la educación, además de menciones a ciencias y eventos en septiembre. Sugiere discusiones sobre la relación entre género y educación, así como eventos específicos que pueden estar relacionados con estas temáticas.

Tópicos para los comentarios de las publicaciones.

- **Logros y Reconocimientos:** En este tópico, se evidencia un enfoque claro en la celebración de logros y reconocimientos. Los comentarios sugieren que la comunidad valora y destaca los éxitos individuales y colectivos. La expresión de términos como "gran" y "merecido" indica una apreciación por el esfuerzo y la excelencia.
- **Agradecimiento a Personas que Hacen la Diferencia:** Este tópico destaca comentarios centrados en expresiones de agradecimiento hacia personas que marcan la diferencia. La comunidad reconoce y valora las contribuciones individuales que tienen un impacto positivo. Este enfoque en la diferencia personal subraya la importancia de las acciones individuales en el tejido social de la comunidad.
- **Reconocimiento al Esfuerzo y Trabajo Destacado:** Aquí, se resalta el reconocimiento al esfuerzo y trabajo destacado. Los comentarios reflejan una apreciación por el arduo trabajo y el rendimiento excepcional. Términos como "trabajo" y "merecido" indican un énfasis en valorar el esfuerzo individual y los logros notables.
- **Agradecimientos y Reconocimiento a Personas y Trabajadores:** Este tópico evidencia un sentido de gratitud y reconocimiento extendido no solo a logros, sino también a personas y trabajadores en general. La comunidad reconoce la labor de aquellos que contribuyen al bienestar y éxito colectivo.

Existe una conexión entre el reconocimiento individual y el agradecimiento a la labor colectiva.

- **Felicitaciones a la Universidad y a Personas Queridas:** En este último tópico, se destaca la expresión de felicitaciones tanto a la universidad como a personas queridas. La comunidad no sólo celebra los logros institucionales, sino que también comparte felicitaciones personales. Este enfoque indica un ambiente emocional y conectado, donde el orgullo se extiende tanto a la institución como a las relaciones personales.

Análisis de Tópicos.

La identificación y clasificación de tópicos proporcionan una valiosa herramienta para explorar y entender las principales temáticas discutidas en los títulos y comentarios. Estos tópicos actúan como categorías temáticas que organizan de manera significativa el contenido, permitiendo una visión estructurada de las conversaciones en línea vinculadas a la institución educativa. Al asignar documentos a estos tópicos, se facilita la interpretación y análisis de las discusiones, brindando una visión detallada de los temas más relevantes y destacados en la comunidad. Este enfoque facilita la comprensión global y la extracción de información clave en el contexto de la institución educativa analizada.

La división de tópicos en cinco categorías proporciona una estructura organizativa que permite una representación más detallada y específica de las discusiones en línea. Al tener cinco tópicos, se logra una mayor granularidad en la clasificación de los documentos, lo que facilita la identificación de las variadas temáticas abordadas en los títulos y comentarios.

Esta subdivisión ofrece una perspectiva más matizada de los temas emergentes, permitiendo un análisis más preciso y detallado de las conversaciones en comparación con una clasificación en un número menor de categorías. Así, la elección de cinco tópicos se traduce en una herramienta analítica más refinada, capaz de capturar la diversidad y complejidad de las interacciones en línea relacionadas con la institución educativa.

Visualización de Tópicos con WordCloud

Se utilizarán visualizaciones de nubes de palabras (WordClouds) para resaltar las palabras clave en cada tópico. Las WordClouds proporcionarán una representación gráfica de las palabras más frecuentes, lo que permitirá una comprensión más intuitiva de los temas dominantes.

Continuaremos con la visualización y análisis de estas WordClouds para cada tópico de los títulos.

Figura 43

Tópicos encontrados en las publicaciones



Análisis de WordCloud para Títulos.

Tópico: Artes y Reconocimientos

Palabras clave identificadas: ** artes, universidad, país, reconocimiento, nacional

Análisis:

- La presencia de "artes" y "reconocimiento" indica un enfoque en logros artísticos y reconocimientos relacionados con la institución educativa.
- "Universidad" sugiere que las discusiones pueden estar centradas en logros específicos de la institución.
- "País" y "nacional" podrían indicar un reconocimiento a nivel nacional.

Tópico: Historia y Memoria

Palabras clave identificadas:** universidad, años, memoria, horas, comunidad, casa
Análisis:

- La presencia de "universidad" sugiere que las discusiones están relacionadas con la institución educativa.
- "Años" y "memoria" indican una posible reflexión sobre el pasado y la historia de la universidad.
- "Comunidad" y "casa" podrían apuntar a la conexión entre la institución y la comunidad.

Tópico: Educación y Estudiantes

Palabras clave identificadas:** universidad, estudiantes, educación, trabajo, academia.

Análisis:

- La presencia de "universidad" y "estudiantes" señala discusiones centradas en la institución y su comunidad estudiantil.
- "Educación" sugiere conversaciones relacionadas con los procesos educativos.
- "Trabajo" y "academia" podrían indicar discusiones sobre la labor académica y el entorno educativo.

Tópico: Actividades y Ciencias Físicas

Palabras clave identificadas:** universidad, estudiantes, ciencias, actividad, física, docente.

Análisis:

- La presencia de "universidad" y "estudiantes" indica discusiones relacionadas con la institución y su población estudiantil.
- "Ciencias" sugiere la inclusión de temáticas científicas.
- "Actividad física" y "docente" podrían señalar discusiones sobre actividades académicas relacionadas con las ciencias físicas.

Tópico: Género y Ciencias

Palabras clave identificadas:** universidad, educación, estudiantes, ciencia, género, diversidad de género.

Análisis:

- La presencia de "universidad" y "estudiantes" indica que las discusiones están centradas en la institución y su comunidad.
- "Ciencia" sugiere que se abordan temas científicos.
- "Género" y "diversidad de género" indican discusiones sobre cuestiones de género en el contexto educativo.

Análisis General.

Este análisis integral permite identificar la diversidad de temas abordados en la comunidad en línea relacionada con la institución educativa. Se observan una

amplia gama de discusiones, desde logros artísticos y reconocimientos hasta reflexiones sobre la historia y la memoria de la universidad.

La presencia constante de términos como "universidad" y "estudiantes" indica un enfoque recurrente en la institución y su comunidad estudiantil. Además, las discusiones abarcan temas educativos, académicos, científicos y sociales, proporcionando una visión completa de la vida universitaria y sus múltiples facetas.

Este análisis facilita la comprensión de los temas más destacados y permite a los interesados obtener información valiosa sobre las prioridades y dinámicas de la comunidad en línea asociada a la institución educativa analizada.

Continuaremos con la visualización y análisis de estas WordClouds para cada tópico de los comentarios.

Figura 44

Tópicos encontrados en los comentarios

Tópico en Comentarios: Logros y Reconocimiento



Tópico en Comentarios: Reconocimiento al esfuerzo y trabajo destacado



Tópico en Comentarios: Agradecimientos a personas que hacen la diferencia



Tópico en Comentarios: Agradecimientos y reconocimiento a personas y trabajadores



Tópico en Comentarios: Felicitaciones a la universidad y a personas queridas



Análisis de WordCloud para Comentarios

Tópico: Logros y Reconocimientos

Palabras Clave Identificadas: felicitaciones, profe, grande, merecido, reconocimiento

Análisis:

- El WordCloud refleja un énfasis en términos relacionados con felicitaciones, reconocimientos y elogios hacia profesores, destacando el orgullo y la admiración.

Tópico: Agradecimiento a Personas que Hacen la Diferencia

Palabras Clave Identificadas: felicitaciones, excelente, docente, gran, merecido

Análisis:

- El WordCloud muestra términos relacionados con agradecimientos, destacando la excelencia y el reconocimiento a docentes por sus contribuciones notables.

Tópico: Reconocimiento al Esfuerzo y Trabajo Destacado

Palabras Clave Identificadas: trabajo, felicitaciones, orgullo, maestro, memoria

Análisis del WordCloud:

- El WordCloud resalta términos vinculados con reconocimientos y aprecio hacia el trabajo y el esfuerzo, expresando orgullo y recuerdo.

Tópico: Agradecimientos y Reconocimiento a Personas y Trabajadores

Palabras Clave Identificadas: trabajo, muchas, gracias, mejor, upla, profesor

Análisis del WordCloud:

- Este WordCloud refleja agradecimientos y reconocimientos, haciendo énfasis en la labor de trabajadores y profesores, destacando la gratitud y la mejora continua.

Tópico: Felicitaciones a la Universidad y a Personas Queridas

Palabras Clave Identificadas: felicidades, excelente, universidad, querida, personas

Análisis del WordCloud:

- El WordCloud muestra términos relacionados con felicitaciones, destacando la excelencia de la universidad y expresiones hacia personas queridas.

Análisis General

Los WordClouds reflejan la naturaleza positiva y de reconocimiento de los comentarios en línea asociados a la institución educativa. Se observa una variedad de términos que indican agradecimiento, felicitaciones y reconocimientos hacia profesores, trabajadores, la institución y personas queridas. Este análisis visual destaca la naturaleza apreciativa y positiva de la comunidad en línea, proporcionando una visión clara de los temas predominantes en cada tópico de comentarios.

Explorando las Emociones: Análisis de Sentimientos en Comentarios de Redes Sociales de la Universidad de Playa Ancha

Este estudio se sumerge en el análisis de sentimientos a través de comentarios recopilados de las redes sociales de la Universidad de Playa Ancha, abordando un conjunto de 1314 filas. Para llevar a cabo este análisis, se emplea la biblioteca pysentimiento, una herramienta versátil que respalda diversos idiomas, destacando especialmente su diseño original para el español.

La librería también abarca distintas tareas que enriquecen el análisis textual, tales como la evaluación de sentimientos, la detección de discursos de odio, la identificación de ironía, el análisis de emociones, el etiquetado de puntos de venta, la detección contextualizada del discurso de odio y el análisis de sentimiento dirigido.

pysentimiento se distingue por su fundamento en modelos preentrenados de transformers, destacando el uso de BETO y RoBERTuito en Español como modelos base. A continuación, se presenta un resumen de los resultados obtenidos mediante pysentimiento para diferentes tareas y modelos:

Tabla 6

Tareas y Modelos de entrenamiento usados por pysentimiento

Tarea/Modelo	Sentimiento	Emoción	Ironía
Beto	67.2 +- 0.6	52.2 +- 1.4	52.2 +- 1.4
Robertuito	70.2 +- 0.2	55.3 +- 0.8	74.1 +- 0.7

Nota. (Pysentimiento, 2023).

Este análisis se apoya en un código eficiente que implementa un sistema de análisis de sentimientos. La función “predecir_sentimiento” toma un comentario como entrada y devuelve una etiqueta de sentimiento, ya sea positivo, negativo o neutro. Además, el código incluye una gestión robusta de errores mediante un bloque try-except para abordar posibles problemas durante el procesamiento, como comentarios no procesables.

A continuación en la figura se muestra el código utilizado para llevar a cabo el análisis de sentimiento.

Figura 45

Código que predice el sentimiento en los comentarios

```
from pysentimiento import create_analyzer

# Crea un analizador de sentimientos para español
analyzer = create_analyzer(task="sentiment", lang="es")

# Función para predecir sentimiento de un comentario y manejar casos especiales
def predecir_sentimiento(comentario):
    try:
        resultado = analyzer.predict(comentario)
        if resultado:
            # Obtiene la clase con la probabilidad más alta
            clase_predominante = max(resultado.probas, key=resultado.probas.get)
            # Obtiene el porcentaje de la clase predominante
            porcentaje_predominante = resultado.probas[clase_predominante] * 100
            return clase_predominante, porcentaje_predominante
        else:
            return "No Comentario", None
    except Exception as e:
        print(f"Error al procesar el comentario: {comentario}")
        return "No Comentario", None

# Filtra los comentarios que no son de tipo float y realiza el análisis de sentimiento solo para ellos
df_total[['Sentimiento', 'Porcentaje']] = df_total['Comentario'].apply(lambda x: pd.Series(predecir_sentimiento(x))
                                                                    if not isinstance(x, float)
                                                                    else pd.Series(["No Comentario", None]))

# Muestra el DataFrame con las nuevas columnas 'Sentimiento' y 'Porcentaje'
df_total
```

El código comienza importando la función `create\_analyzer` de la biblioteca `pysentimiento`. Luego, se utiliza para crear un analizador de sentimientos específico para la tarea de análisis de sentimientos en español, y este analizador se almacena en la variable 'analyzer'.

La funcionalidad principal se encuentra en la función `predecir\_sentimiento`. Esta función toma un comentario como entrada y utiliza el analizador previamente creado para predecir su sentimiento. La función incorpora un manejo de excepciones para gestionar posibles errores durante el proceso de análisis. En caso de un error, se imprime un mensaje indicando el problema y se devuelve la etiqueta "No Comentario". Si la predicción es exitosa, se extrae la clase con la probabilidad más alta y se calcula el porcentaje de esa clase predominante.

Posteriormente, la función `predecir\_sentimiento` se aplica a la columna 'Comentario' en un DataFrame llamado `df\_total`. Se filtran los comentarios que no son de tipo float, asegurando que la función de análisis de sentimientos se aplique únicamente a comentarios válidos. Los resultados de esta operación se almacenan en dos nuevas columnas llamadas 'Sentimiento', que indica si el comentario es positivo, negativo o neutro, y 'Porcentaje', que almacena el nivel de confianza expresado como un porcentaje en la clasificación emocional del comentario.

Finalmente, el DataFrame modificado se muestra como salida del programa, ahora incluyendo las columnas adicionales que reflejan el sentimiento asociado a cada comentario y el porcentaje de confianza de esa predicción. Este enfoque

proporciona una visión más detallada de la composición emocional de los comentarios en la columna 'Comentario' del DataFrame.

4.2.1.- Visualización de los resultados del análisis

En búsqueda de datos significativos en el conjunto de datos formado.

Una vez aplicada las técnicas de tokenización y de evaluar el tono emocional de los comentarios que ayudarán a tener un conjunto de datos más formado, nos dirigimos especialmente a analizar dicho conjunto de datos para poder obtener así una mayor especificación de los datos.

Visión General

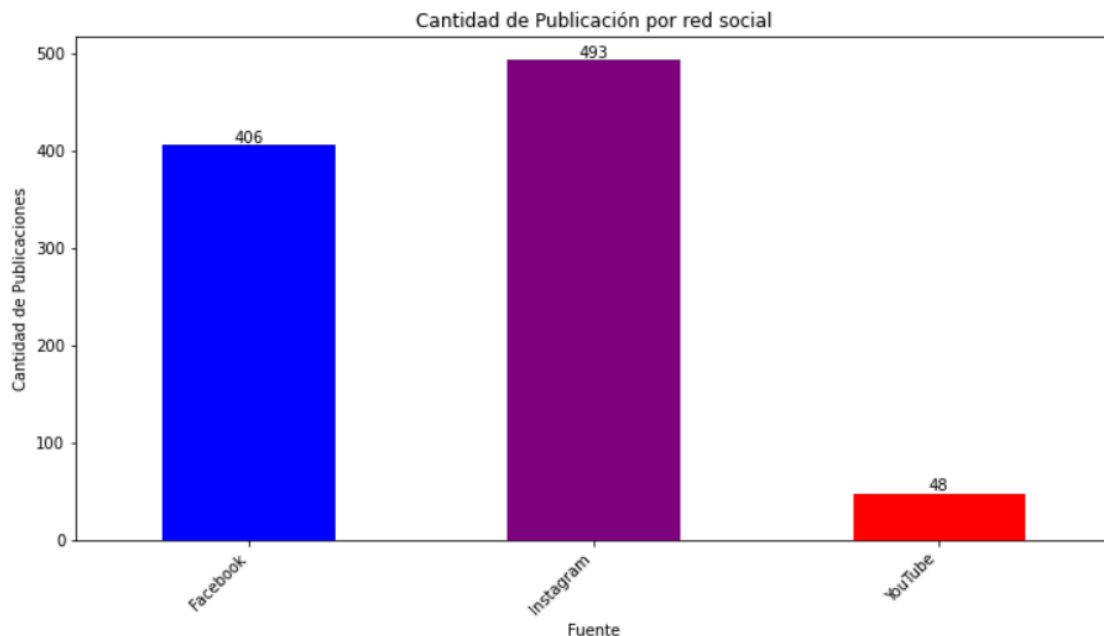
Cantidad de publicaciones que se extrajeron es de: 947 de 1500 esperados.

La discrepancia se atribuye principalmente a que la cuenta oficial de la Universidad de Playa Ancha (@upla_comunica) en YouTube tiene los comentarios desactivados para sus 545 videos. Como solución, hemos extraído comentarios de la cuenta de Facultad de Ingeniería UPLA (@facultaddeingenieriaupla), que cuenta con 81 videos. Además, se observan algunos errores ocasionales al obtener datos de las publicaciones a través de librerías como facebook_scraper, aunque su impacto es mínimo.

Vista general de la cantidad de publicaciones por redes sociales.

Figura 46

Cantidad de publicaciones por red social



Con base en los datos proporcionados, se presenta la cantidad de publicaciones por red social:

- Facebook: 406 publicaciones
- Instagram: 493 publicaciones
- YouTube: 48 publicaciones

Este gráfico ilustra la distribución de las publicaciones extraídas de la Universidad de Playa Ancha en estas plataformas. No obstante, es esencial considerar algunas observaciones clave:

Debido a limitaciones inherentes de las librerías y al contenido más restringido en YouTube por parte de la universidad, se observa una marcada disminución en la cantidad de publicaciones en esta plataforma en comparación con Facebook e Instagram.

Consideraciones Adicionales:

Resulta particularmente valioso fomentar la interacción en la plataforma de YouTube. Actualmente, los comentarios están desactivados en la página oficial de la Universidad de Playa Ancha en YouTube, lo que restringe la oportunidad de recibir feedback directo y construir una comunidad más activa.

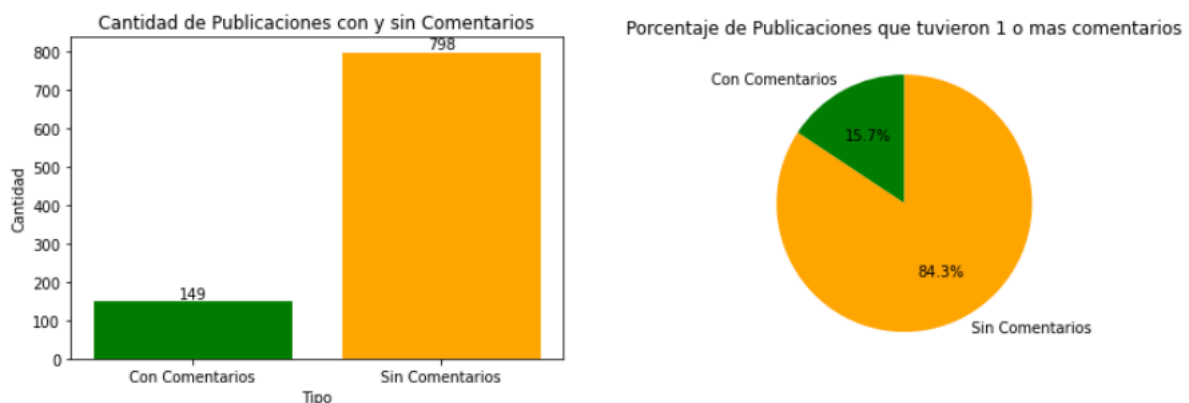
Habilitar los comentarios permitiría a la audiencia expresar opiniones, hacer preguntas y compartir experiencias, enriqueciendo la relación entre la universidad y su comunidad en línea. Esta interacción no solo ofrece información valiosa sobre expectativas y preferencias del público, sino que también cultiva un sentido de comunidad y pertenencia. Considerar el desbloqueo de comentarios puede ser un paso significativo hacia una comunicación más abierta y colaborativa. ¡Tu participación es fundamental para construir un espacio de diálogo enriquecedor.

Participación Activa vs. Pasiva: Desglose de comentarios en publicaciones.

Dado los siguientes gráficos

Figura 47

Participación activa vs pasiva en general



1. Participación Activa vs. Pasiva

- La mayor parte de las publicaciones (84.3%) no recibió comentarios, indicando una participación pasiva por parte de la comunidad. Esto podría sugerir que la mayoría de las publicaciones no generan suficiente interacción o que el contenido no motiva a los usuarios a comentar.
- El 15.7% de las publicaciones con comentarios muestran una participación activa, lo cual es valioso ya que indica que cierto contenido genera respuestas o interacciones directas. Explorar qué tipo de publicaciones reciben comentarios podría ofrecer perspectivas valiosas sobre el contenido que resuena mejor con la audiencia.

2. Interacción de la Comunidad

- El bajo porcentaje de comentarios podría indicar que la comunidad no se siente tan comprometida para interactuar directamente con las publicaciones. Es importante explorar las razones detrás de esto: ¿es el contenido poco atractivo, la falta de incentivos para comentar, o algún otro factor?
- Puede ser útil analizar las publicaciones específicas que generan comentarios para entender qué características o temas despiertan el interés de la comunidad.

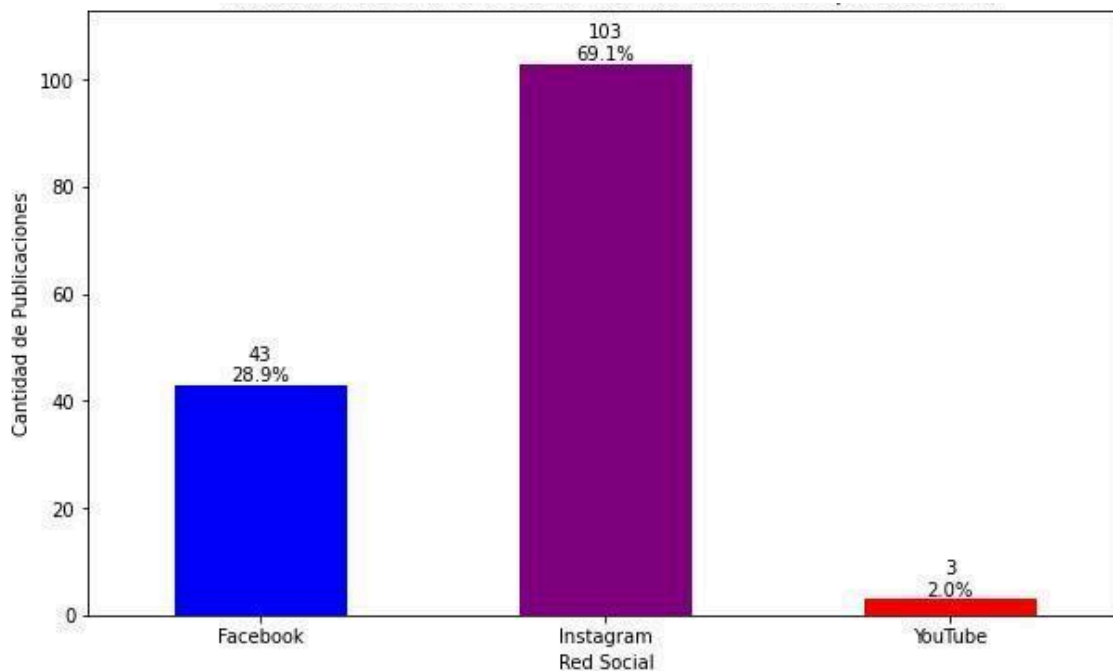
3. Estrategias para Fomentar la Participación

- Con base en estos datos, podrías considerar desarrollar estrategias específicas para fomentar la participación de la comunidad. Esto podría incluir la creación de contenido más interactivo, encuestas, o cualquier otra táctica que aliente a los usuarios a dejar comentarios.

Exploración de la Participación: Publicaciones con Comentarios por Red Social

Figura 48

Cantidad de publicaciones con uno o más comentarios por red social



Descripción: En el gráfico de barras se presenta la distribución de publicaciones con comentarios según la red social. Se observa que de las 149 publicaciones con comentarios extraídas:

- **Instagram** lidera en cantidad de publicaciones con comentarios, contribuyendo con más de la mitad de las interacciones registradas.
- **Facebook** muestra una participación significativa, aunque su proporción es menor en comparación con Instagram.
- **YouTube**, a pesar de tener menos publicaciones con comentarios, representa aproximadamente un 2% del total recopilado.

Estas cifras son una fracción de las 947 publicaciones recopiladas en total, donde la mayoría no presenta interacciones. Este análisis destaca la relevancia de considerar la participación del público en redes sociales al evaluar el alcance y la interacción de las publicaciones de la Universidad de Playa Ancha.

Cantidad Total de Comentarios por Red Social

Figura 49

Cantidad total de comentarios por red social

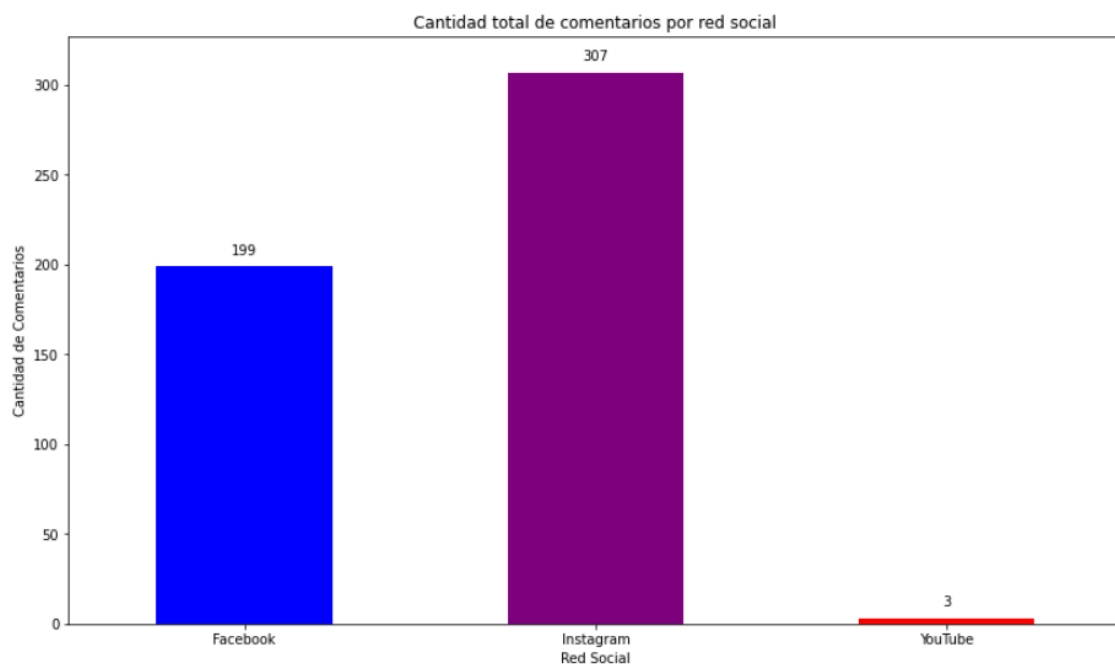
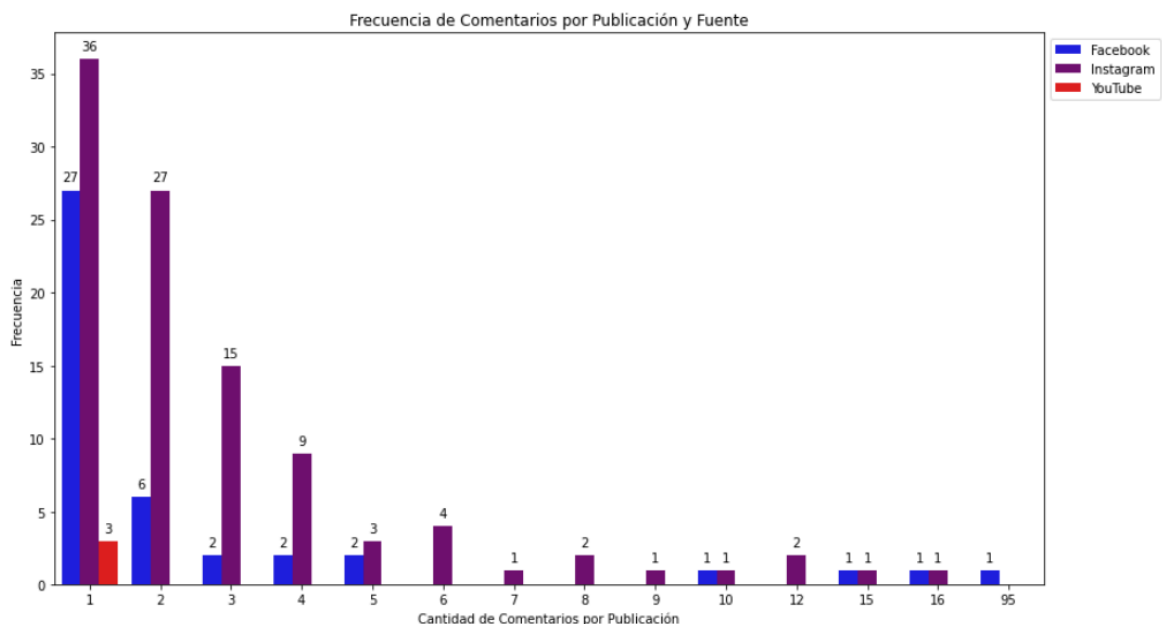


Figura 50

Frecuencia de comentarios por publicación y fuente



En el primer gráfico de barras, se presenta la cantidad total de comentarios recopilados por cada red social. Los resultados indican que Instagram lidera con 307 comentarios, seguido por Facebook con 199 comentarios, mientras que YouTube cuenta con solo 3 comentarios. Esta información destaca la interacción significativa de la audiencia en Instagram en comparación con las otras plataformas.

En el segundo gráfico de barras, se analiza la frecuencia de comentarios por publicación y fuente. Se observa que:

- Para una frecuencia de 1, Facebook tiene 27 publicaciones, Instagram cuenta con 36, y YouTube presenta 3.
- A medida que aumenta la frecuencia, Facebook e Instagram muestran una disminución en el número de publicaciones, mientras que YouTube tiene una frecuencia de 0 para frecuencias mayores a 1.
- La mayoría de las publicaciones tienen una frecuencia baja, con pocos casos de frecuencias mayores.

Este análisis sugiere que, aunque Facebook e Instagram tienen una cantidad considerable de publicaciones con bajos recuentos de comentarios, Instagram destaca en términos de participación continua y mayor frecuencia de comentarios por publicación. Además, YouTube muestra una participación más limitada, con la mayoría de sus publicaciones teniendo frecuencia cero.

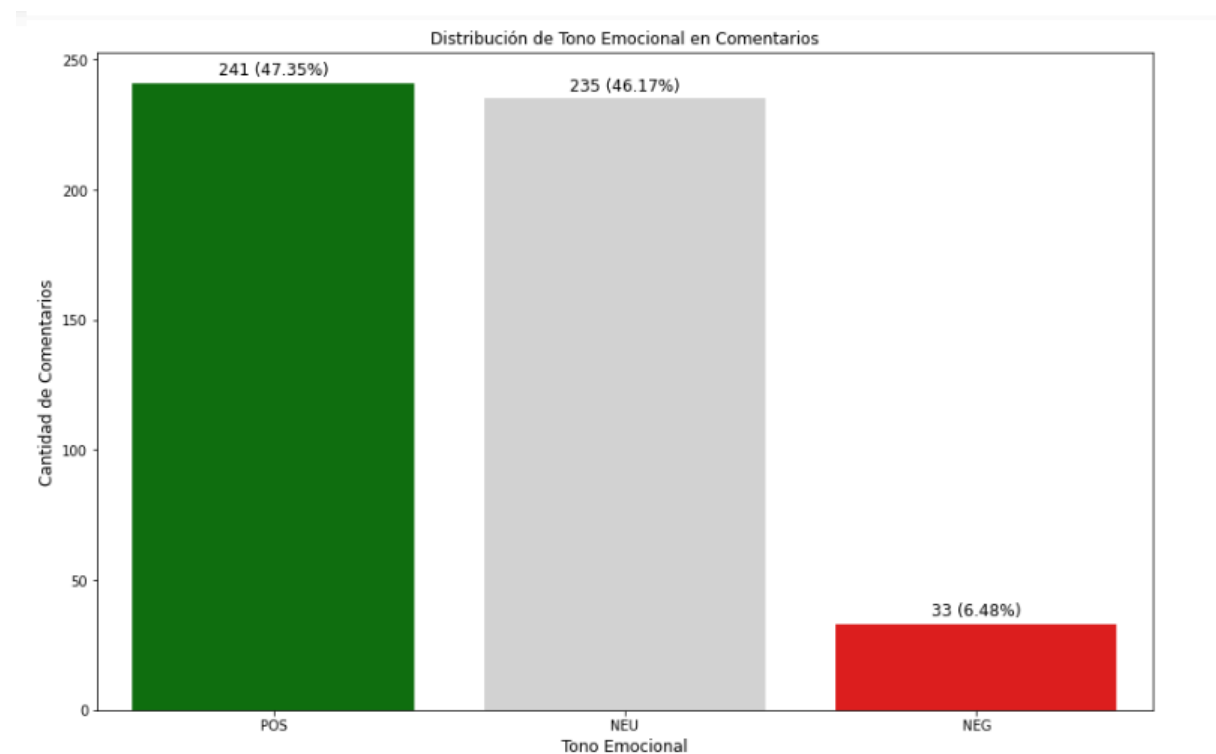
Estos hallazgos pueden ser útiles para comprender mejor cómo la audiencia interactúa con las publicaciones en cada plataforma y guiar estrategias para mejorar la participación en aquellas con menor interacción.

Análisis de la Distribución del Tono Emocional en los Comentarios

En este análisis, se mostrará la cantidad de comentarios en función de su tono emocional, proporcionando una visión general de la distribución de los sentimientos expresados en los comentarios recopilados.

Figura 51

Distribución de tono emocional en comentarios



En el gráfico de barras generado, se observa la prevalencia de tres tonos emocionales principales: positivo, neutro y negativo.

Tono Emocional Positivo (POS): 241 comentarios (47.35%)

- Representa la mayor proporción de comentarios.
- Indica una respuesta generalmente positiva por parte de los usuarios.

Tono Emocional Neutro (NEU): 235 comentarios (46.17%)

- Se encuentra cerca del tono positivo en términos de cantidad.
- Sugiere una respuesta neutral o falta de carga emocional significativa en los comentarios.

Tono Emocional Negativo (NEG): 33 comentarios (6.48%)

- Representa la menor proporción de comentarios.
- Indica una respuesta generalmente negativa o crítica por parte de los usuarios.

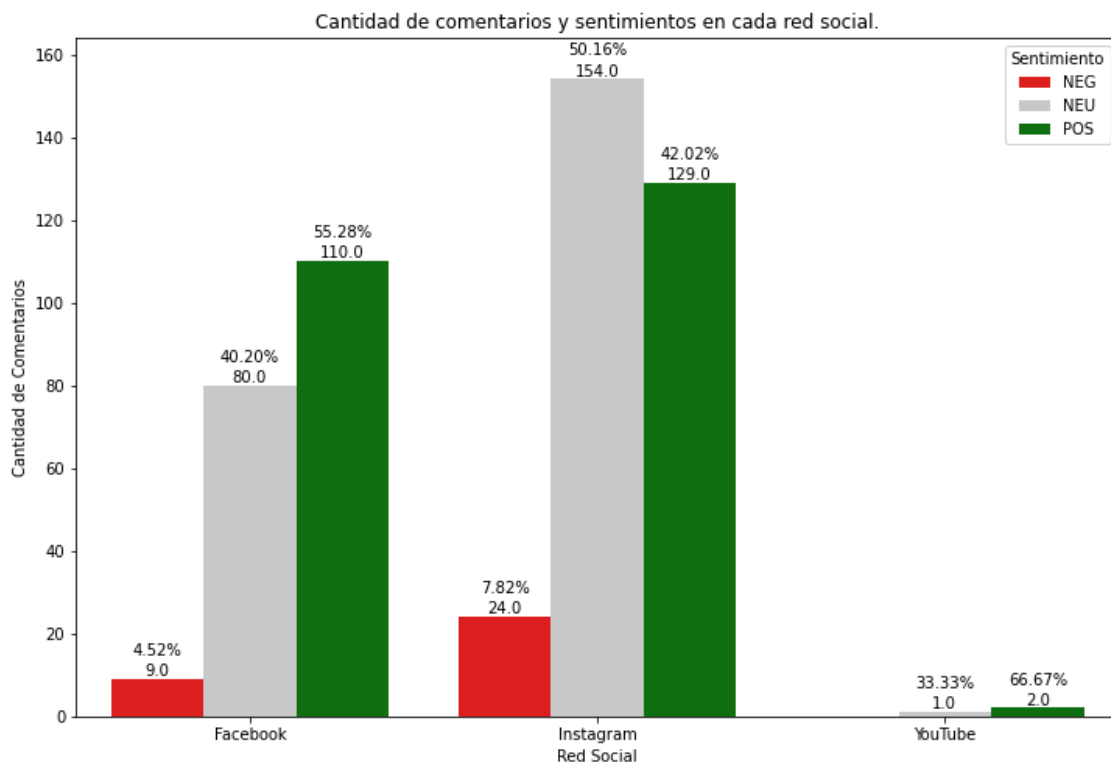
Este análisis resalta la predominancia de tonos emocionales positivos y neutros en los comentarios, indicando en general una respuesta favorable o neutral por parte de los usuarios. El bajo porcentaje de comentarios con tono emocional negativo sugiere que la mayoría de las interacciones en este contexto específico son positivas o, al menos, neutrales. Este conocimiento puede ser valioso para comprender la percepción y la recepción de los contenidos analizados.

Análisis de la Cantidad de Comentarios y Sentimientos en Redes Sociales

Este análisis proporciona una exploración exhaustiva de la cantidad de comentarios catalogados por sentimiento en tres destacadas plataformas de redes sociales: Facebook, Instagram y YouTube. El objetivo es ofrecer una visión clara de cómo se distribuyen las emociones en cada red social, facilitando la comprensión de la interacción del público en estas plataformas digitales.

Figura 52

Distribución de tono emocional en comentarios



El gráfico de barras apiladas presenta la cantidad total de comentarios para cada red social, desglosados por sentimiento. Cada barra representa la suma de comentarios para una red social específica, y la variación de colores en las barras refleja la proporción de sentimientos: negativos (NEG), neutros (NEU) y positivos (POS).

Análisis por Red Social:

1. Facebook:

- Total de Comentarios: 199
- Comentarios Negativos: 9 (4.52%)
- Comentarios Neutros: 80 (40.20%)
- Comentarios Positivos: 110 (55.28%)

2. Instagram:

- Total de Comentarios: 307
- Comentarios Negativos: 24 (7.82%)
- Comentarios Neutros: 154 (50.16%)
- Comentarios Positivos: 129 (42.02%)

3. YouTube:

- Total de Comentarios: 3
- Comentarios Neutros: 1 (33.33%)
- Comentarios Positivos: 2 (66.67%)

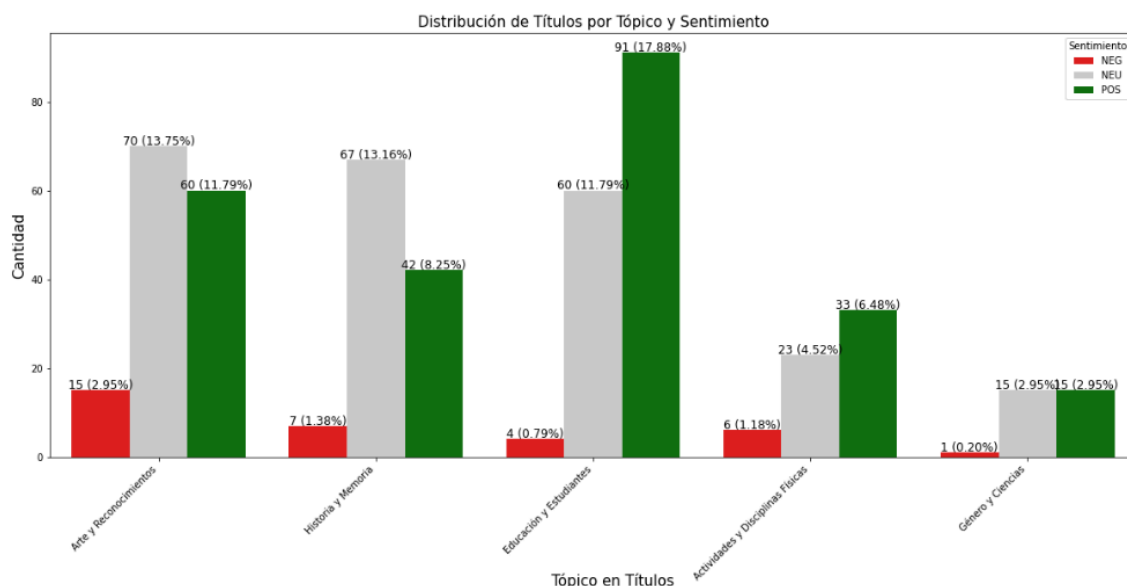
Este análisis destaca la prevalencia del tono emocional neutro y positivo en las tres redes sociales analizadas. Sin embargo, se observa que Instagram registra la mayor cantidad de comentarios negativos, seguida por Facebook. En contraste, YouTube presenta una proporción significativa de comentarios positivos. Este panorama refleja la complejidad de las interacciones emocionales en distintos entornos digitales y subraya la importancia de comprender las dinámicas específicas de cada plataforma.

Distribución de Títulos por Tópico y Sentimiento: Una Inmersión en la Expresión Emocional en los Diversos Contextos Temáticos.

Este análisis profundiza en la manera en que se dispersan los sentimientos en los títulos según distintos tópicos. Cada tópico refleja una temática única, y cada título se clasifica en términos de su tono emocional: negativo, neutro o positivo. Esta exploración brinda una perspectiva detallada sobre cómo la audiencia percibe y expresa sus emociones en diversos ámbitos temáticos.

Figura 53

Distribución de tópicos de los títulos y su tono emocional



Análisis Detallado:

1. Arte y Reconocimiento:

- Destaca por tener la mayor cantidad de comentarios negativos entre los tópicos abordados, representando el 2.95% del total de comentarios.
- A su vez, presenta la mayor proporción de comentarios neutrales en comparación con otros tópicos, alcanzando un 13.75%.
- Se posiciona como el segundo tópico con mayor cantidad de comentarios positivos, registrando un 11.79%.

2. Historia y Memoria:

- Se destaca como el segundo tópico con más comentarios neutros, abarcando el 13.16%.
- Se encuentra en la cuarta posición en comentarios negativos, con un 1.38%.

3. Educación y Estudiantes:

- Sobresale al contar con la mayor cantidad de comentarios positivos, representando el 17.88% del total de comentarios.
- Se ubica como el segundo tópico con menor cantidad de comentarios negativos, registrando solo un 0.78%.
- Exhibe una notable cantidad de comentarios neutros, alcanzando el 11.79%.

4. Actividades y Disciplinas Físicas:

- Se posiciona como el segundo tópico con menos participación, recopilando un total de 62 comentarios.

5. Género y Ciencias:

- Representa el tópico con menor participación, abarcando solo el 6.1% del total de comentarios.
- A pesar de su baja participación, muestra una proporción equitativa entre comentarios neutros y positivos, con solo un comentario negativo.

Observaciones Generales

- El tópico "Educación y Estudiantes" destaca al tener la mayor cantidad de comentarios positivos (17.88%), señalando una percepción generalmente favorable.
- A pesar de la diversidad de tópicos, la mayoría de los comentarios se clasifican como neutros, indicando una inclinación hacia expresiones moderadas en diversas áreas temáticas.
- "Género y Ciencias" presenta una proporción equilibrada entre comentarios neutros y positivos, con una baja presencia de comentarios negativos.
- Temas que abarcan "educación y estudiantes" generan mayor interés en la comunidad abarcando el 30.5% de los comentarios totales, solo teniendo un 0,79% de comentarios negativos el segundo mejor típico.

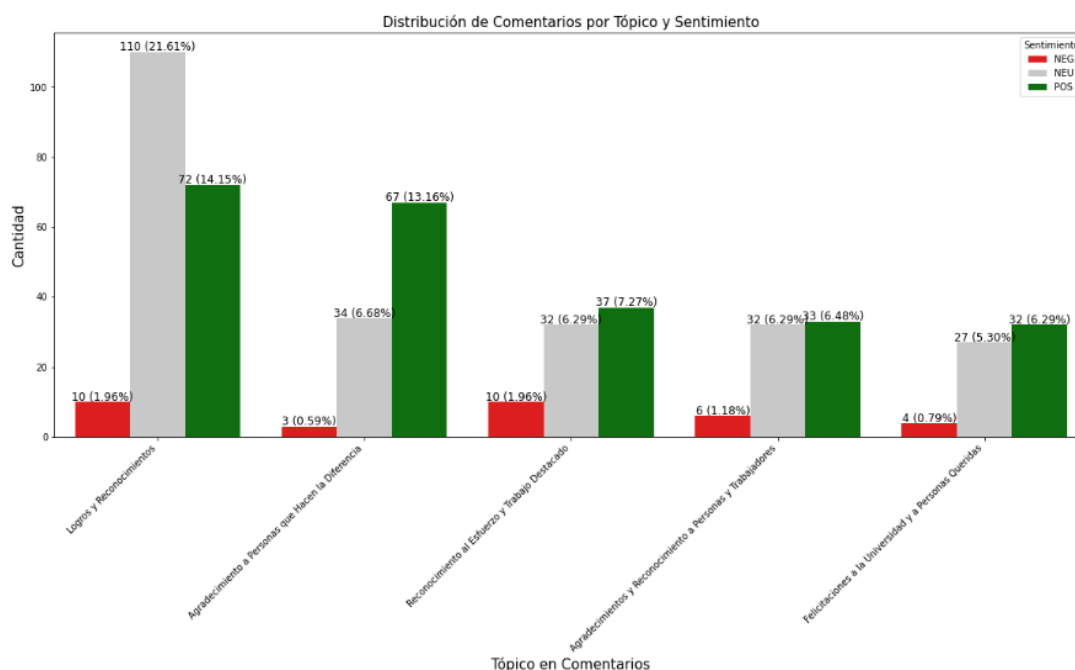
Este análisis ofrece una visión enriquecedora de cómo las emociones se entrelazan con temas específicos, proporcionando información valiosa sobre la resonancia emocional en distintas áreas temáticas.

Análisis Profundo de la Distribución de Comentarios por Tópico: Explorando las Emociones en Expresiones Específicas

Este análisis se sumerge en la distribución de comentarios en diversos tópicos específicos, ofreciendo una perspectiva detallada de las emociones expresadas en cada categoría. Cada tópico refleja un tema particular, y cada comentario se clasifica como negativo, neutro o positivo. A continuación, se presenta un análisis pormenorizado para cada tópico:

Figura 54

Distribución de comentarios por tópicos y sentimientos



1. Logros y Reconocimientos:

- Posee la mayor cantidad de comentarios neutros, abarcando el 21.61% de los comentarios totales, indicando una participación activa de la audiencia.
- También exhibe la mayor cantidad de comentarios positivos, representando el 14.15%, sugiriendo una interacción positiva y significativa.
- Abarca la mayor cantidad de comentarios negativos, con un 1.96%. En conjunto, este tópico engloba el 37.72% de los comentarios totales, superando un tercio del total.

2. Agradecimiento a Personas que Hacen la Diferencia:

- Predomina con comentarios positivos, sugiriendo una generación más positiva de interacciones.
- Tiene la menor proporción de comentarios negativos en comparación con otros tópicos.
- Representa el 20.43% de los comentarios totales, con solo un 2.8% clasificados como negativos, el menor entre todos.

3. Reconocimiento al Esfuerzo y Trabajo Destacado:

- A pesar de tener una proporción modesta de comentarios positivos y neutros, destaca por la mayor cantidad de comentarios negativos, con un 1.96%.
- Este tópico abarca el 12.6% de los comentarios negativos en total, siendo el más alto entre los demás.
- Presenta el índice de negatividad más elevado.

4. Agradecimientos y Reconocimiento a Personas y Trabajadores:

- Exhibe una distribución equitativa entre comentarios neutros y positivos.
- Similar al tópico anterior, representa el 1.18% de los comentarios totales clasificados como negativos.

5. Felicitaciones a la Universidad y a Personas Queridas:

- Aunque tiene la menor interacción, destaca por tener una proporción notable de comentarios positivos y neutros.
- Es el segundo tópico con la menor cantidad de comentarios negativos.

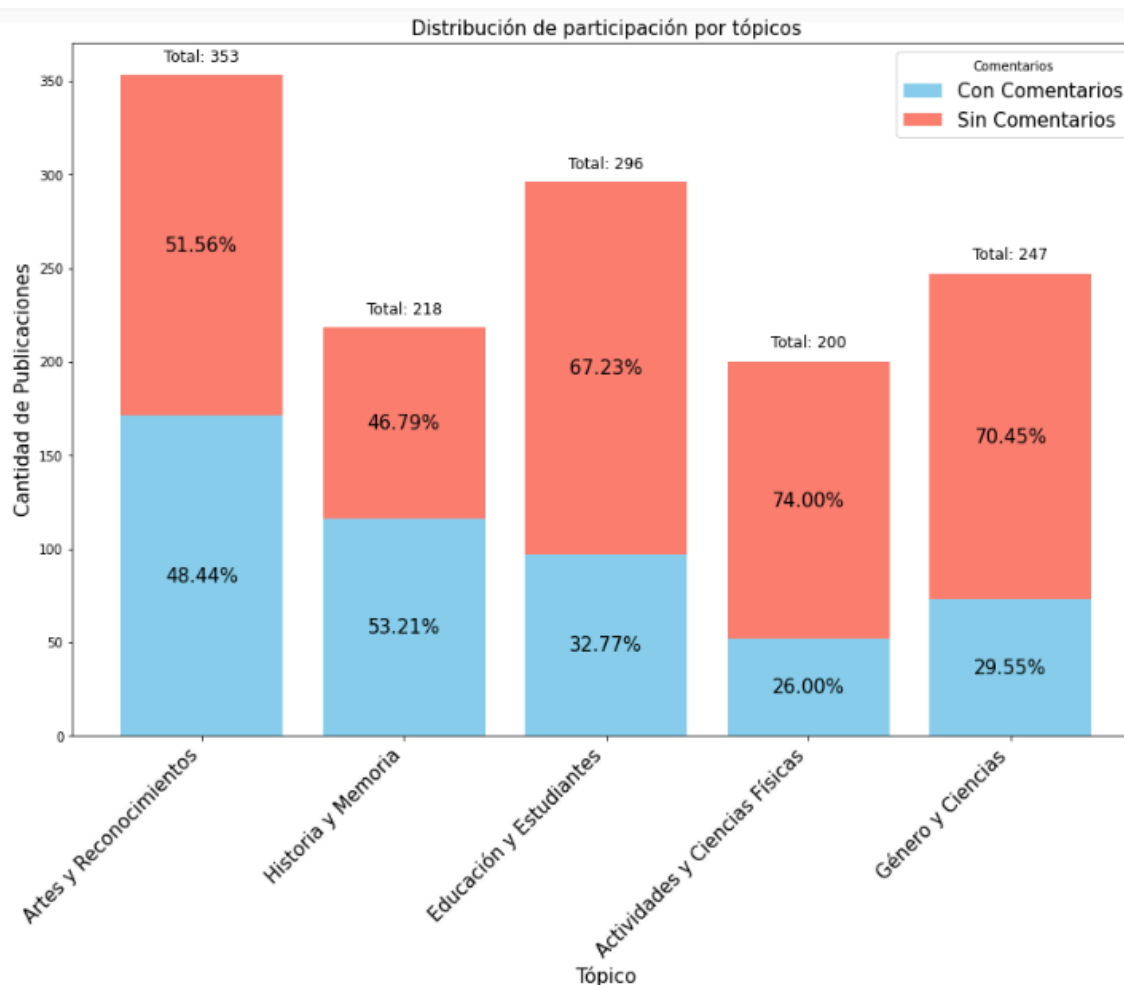
Observaciones Generales

- "Agradecimiento a Personas que Hacen la Diferencia" se distingue por su alta proporción de comentarios positivos, indicando una apreciación notable en este contexto específico.
- "Logros y Reconocimientos" sobresale por su cantidad significativa de comentarios neutros, sugiriendo una respuesta equilibrada de la audiencia.
- Aunque los tópicos presentan variaciones en las emociones expresadas, la mayoría de los comentarios se clasifican como neutros, reflejando una tendencia hacia la expresión moderada.
- En general, todos los tópicos muestran una mayor cantidad de comentarios positivos y neutros, indicando una dinámica positiva en las interacciones.

Análisis de Participación por Tópicos

Figura 55

Distribución de participación por tópicos



1. Artes y Reconocimientos:

- Es el tópico con la mayor cantidad de publicaciones.
- La participación con comentarios está por debajo del 50%, lo que sugiere una distribución equilibrada de comentarios y sin comentarios.

2. Historia y Memoria:

- Presenta un alto porcentaje de participación con comentarios, indicando un interés activo en este tópico.
- La cantidad total de publicaciones es significativa, sugiriendo un interés general en temas históricos y de memoria.

3. Educación y Estudiantes:

- Aunque es el tópico con más publicaciones, tiene el porcentaje más bajo de participación con comentarios.

- La alta proporción de publicaciones sin comentarios podría indicar una falta de participación activa en este tópico.

4. Actividades y Ciencias Físicas:

- Tiene el porcentaje más bajo de participación con comentarios, indicando un menor interés o involucramiento en este tópico.
- La mayoría de las publicaciones no generan comentarios, lo que sugiere un nivel más bajo de interacción.

5. Género y Ciencias:

- Aunque tiene una cantidad considerable de publicaciones, el porcentaje de participación con comentarios es relativamente bajo.
- La mayoría de las publicaciones en este tópico no generan comentarios, lo que podría indicar un interés más pasivo.

Conclusiones Generales:

- El tópico de Historia y Memoria parece ser el más interesante, con un alto porcentaje de participación activa.
- Artes y Reconocimientos tiene la mayor cantidad total de publicaciones, pero con una participación más equilibrada.
- Educación y Estudiantes podría ser el menos interesante, dado su bajo porcentaje de participación con comentarios y la alta proporción de publicaciones sin comentarios.
- Actividades Ciencias y Físicas, Género y Ciencias también tienen participaciones más bajas y pueden requerir estrategias para fomentar la interacción.

Análisis de la Distribución Mensual de Comentarios

Cantidad de publicaciones recopiladas en el segundo semestre del 2023:

- Agosto: 364 Publicaciones.
- Septiembre: 342 Publicaciones.
- Octubre: 504 Publicaciones.
- Noviembre: 55 Publicaciones.

El siguiente análisis se basa en la recopilación de datos de publicaciones realizadas durante los meses de agosto, septiembre, octubre y noviembre de 2023.

1. Distribución Mensual:

- Agosto: Se recopilaron un total de 364 publicaciones durante este mes.
- Septiembre: La cantidad de publicaciones se mantuvo elevada, alcanzando un total de 342.
- Octubre: Este mes registró la mayor cantidad de publicaciones, con un total de 504.

- Noviembre: Aunque disminuyó significativamente en comparación con octubre, se recopilaron 55 publicaciones en este mes.

2. Tendencias Mensuales:

- La tendencia general muestra un aumento significativo de publicaciones en octubre, posiblemente debido a eventos relevantes o temas populares durante ese período.
- Agosto y septiembre mantuvieron cifras relativamente estables, indicando una actividad constante en esos meses.
- Noviembre muestra una disminución en la actividad, lo cual podría deberse a factores estacionales o eventos específicos de ese mes.

3. Comparación con el Análisis de Comentarios:

- Sería interesante comparar estos datos con la cantidad de comentarios en cada mes para entender la participación y la interacción de la audiencia en relación con la cantidad de publicaciones.

4. Implicaciones para Estrategias de Contenido:

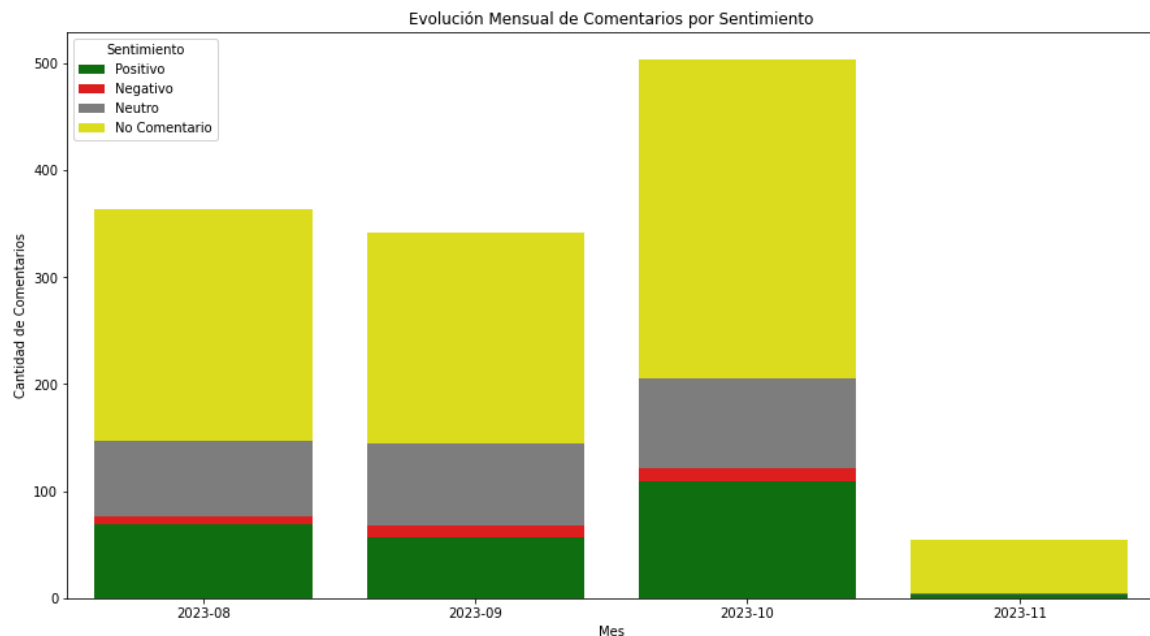
- Con base en la alta actividad registrada en octubre, podría ser beneficioso analizar las temáticas o eventos específicos que generaron este aumento y considerar estrategias similares en el futuro.
- Noviembre, al tener una disminución en la cantidad de publicaciones, podría ser un mes para evaluar y ajustar estrategias de contenido.

El análisis de la cantidad de publicaciones recopiladas proporciona información clave sobre las tendencias mensuales, lo que puede ser útil para adaptar estrategias de contenido y mejorar el compromiso con la audiencia.

Análisis de Sentimientos: Evaluación de la Reacción del Público ante las Publicaciones Mensuales.

Figura 56

Evolución mensual de comentarios por sentimiento



1. Tendencia General:

- A lo largo de los meses registrados (agosto a noviembre de 2023), se observa un aumento en la cantidad total de comentarios, alcanzando el punto máximo.

2. Proporción de Sentimientos:

- En general, el mes de octubre tiene la mayor proporción de comentarios en todas las categorías, con números significativos de comentarios positivos y neutros. Esto sugiere una participación más activa durante ese período.

3. Cambios Significativos:

- Se destaca un cambio notable en septiembre, donde la cantidad total de comentarios disminuye en comparación con agosto y luego vuelve a aumentar en octubre. Podría ser interesante investigar eventos o cambios en la dinámica que coincidan con estos cambios.

4. Proporción de Sentimientos en Cada Mes:

- Aunque hay variabilidad en la proporción de comentarios positivos, negativos y neutros entre los meses, la categoría "No Comentario" domina en todos los meses. Esto indica que la mayoría de las interacciones son de naturaleza informativa o de aprobación.

5. Identificación de Eventos:

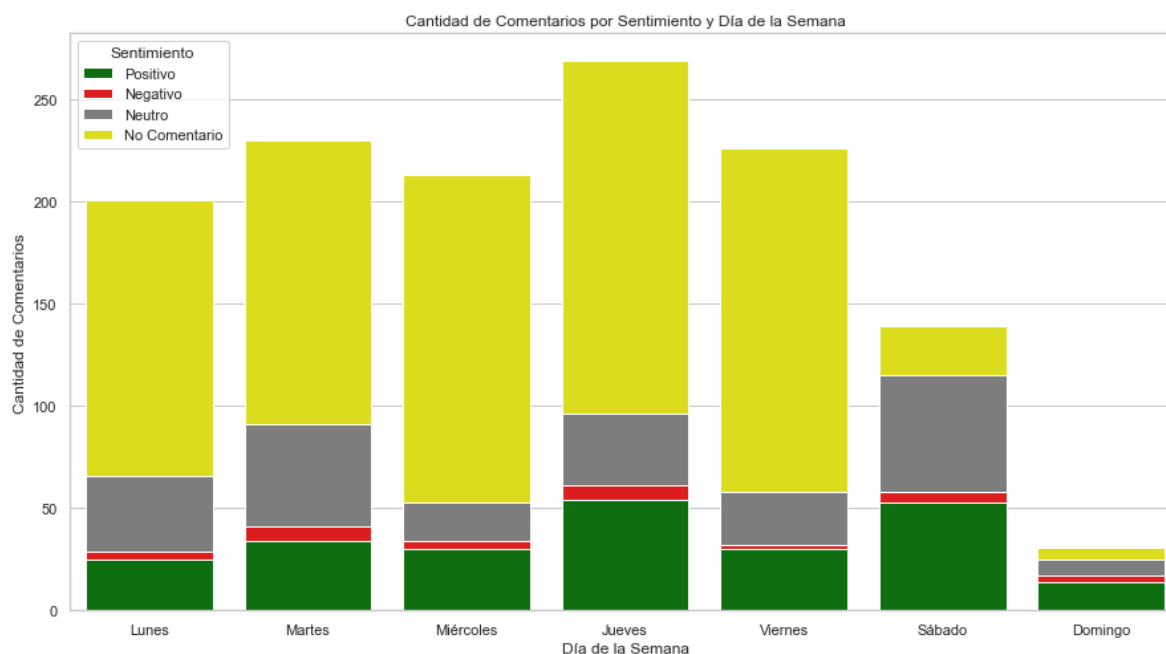
- Para comprender mejor los cambios en la dinámica de comentarios, sería beneficioso investigar eventos específicos o cambios en la actividad durante los meses de septiembre y octubre.

La distribución mensual de comentarios sugiere patrones de participación y cambios a lo largo del tiempo. Sería valioso profundizar en eventos específicos o factores externos que puedan haber influido en estas tendencias. Además, la preponderancia de comentarios sin sentimiento podría indicar una dinámica particular en la interacción de la audiencia.

Análisis de Sentimientos por Día de la Semana

Figura 57

Cantidad de comentarios y su sentimiento en relación a los días de la semana



Los datos presentan la distribución de sentimientos (positivo, negativo, neutro y sin comentario) en función del día de la semana. Aquí se presenta un análisis de los patrones y tendencias observadas:

1. Frecuencia de Publicaciones:

- Jueves y Sábado: Los días jueves y sábado registran la mayor cantidad de publicaciones. Estos días parecen ser los más activos en términos de comentarios.
- Domingo y Miércoles: Por el contrario, el domingo y miércoles tienen la menor cantidad de comentarios.

2. Sentimientos Predominantes:

- Positivo: Los días jueves, sábado y viernes son los días donde predominan los comentarios positivos.
- Negativo: Los días jueves también presentan la mayor cantidad de comentarios negativos, seguido de domingo y martes.
- Neutro: El miércoles tiene una cantidad significativa de comentarios neutros, seguido de jueves y martes.

3. Implicaciones para Estrategias de Publicación:

- La identificación de patrones en la respuesta del público puede ser útil para planificar estrategias de publicación y moderación.
- Días con mayor participación (jueves y sábado) podrían ser ideales para lanzar contenido importante o promociones.
- La baja actividad en días como domingo y miércoles podría indicar la necesidad de ajustar las estrategias de publicación en esos días.

4. Monitoreo Continuo:

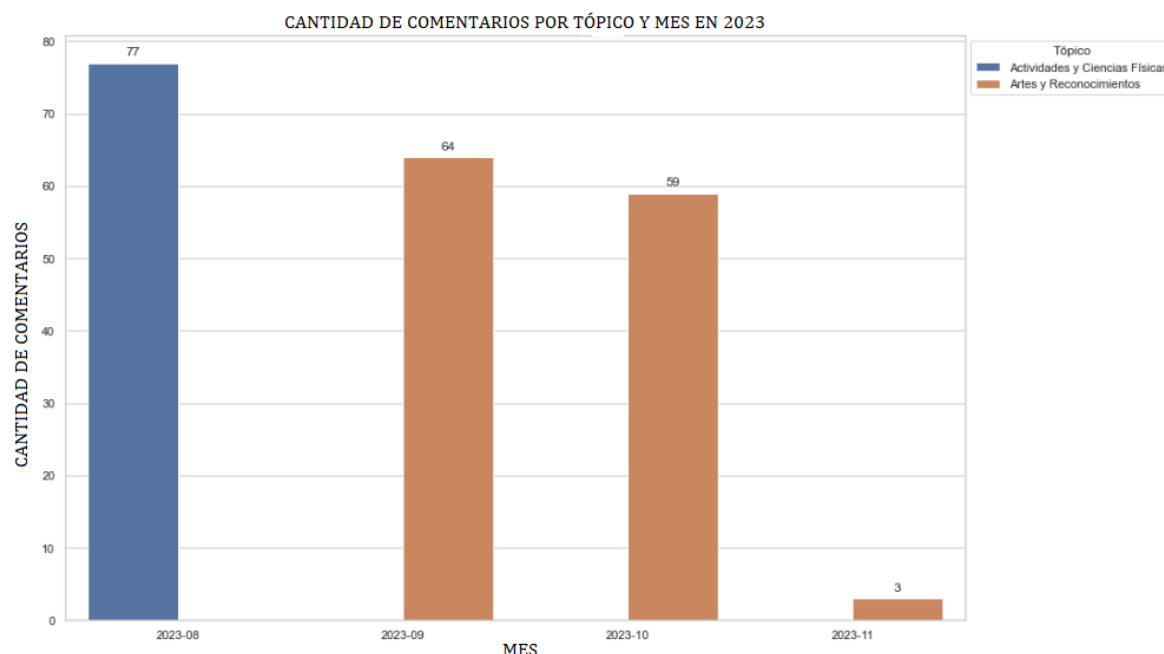
- Es importante tener en cuenta que los patrones de participación pueden cambiar con el tiempo. Se recomienda un monitoreo continuo y ajustes en las estrategias según sea necesario.

Este análisis proporciona una visión general de la interacción del público en diferentes días de la semana, lo que puede ser valioso para optimizar la planificación de contenido y la gestión de la comunidad en línea.

Análisis de Comentarios por Tópico y Mes en 2023

Figura 58

Cantidad de comentarios por tópicos en diferentes meses



1. Actividades y Ciencias Físicas (Tópico 3):

- Agosto Este tópico fue el más comentado en agosto, con un total de 77 comentarios. Podría sugerir un interés destacado en actividades físicas y ciencias relacionadas durante ese mes.

2. Artes y Reconocimientos (Tópico 0):

- Septiembre y Octubre: El tópico de Artes y Reconocimientos (Tópico 0) fue el más comentado en septiembre y octubre, con 64 y 59 comentarios respectivamente. Esto indica un fuerte interés o discusión en torno a temas relacionados con artes y reconocimientos durante esos meses.
- - ****Noviembre:**** Aunque este tópico sigue siendo el más comentado en noviembre, la cantidad de comentarios disminuyó significativamente a solo 3. Sería necesario investigar más para comprender la razón detrás de esta disminución.

3. Recomendaciones:

- Para el tópico de Actividades y Ciencias Físicas (Tópico 3), considerar la posibilidad de continuar generando contenido relacionado con actividades físicas y ciencias, ya que ha demostrado ser popular.
- Para el tópico de Artes y Reconocimientos (Tópico 0), analizar la razón detrás de la disminución de comentarios en noviembre. Puede ser útil explorar nuevas formas de mantener el interés en este tema.

4. Comparación Mensual:

- La variación en los tópicos más comentados mes a mes resalta la importancia de adaptar estrategias de contenido según las tendencias y preferencias del público en evolución.

5. Consideración Adicional:

- Sería beneficioso complementar este análisis con información cualitativa, como el contenido específico discutido en cada tópico, para obtener una comprensión más profunda de las preferencias del público.

Análisis final sobre la minería de datos y la información encontrada.

Este análisis proporciona información valiosa sobre los tópicos que generan más comentarios, lo que puede guiar la estrategia de contenido y el compromiso de la audiencia en el futuro.

Análisis y Recomendaciones para Estrategias de Contenido y Participación en Redes Sociales de la Universidad de Playa Ancha

En el estudio de la participación y la dinámica de temas en las redes sociales de la Universidad de Playa Ancha, se observan tendencias significativas que proporcionan oportunidades para optimizar las estrategias de contenido y mejorar la interacción con la comunidad en línea.

Participación Activa vs. Pasiva:

La predominancia de publicaciones sin comentarios (84.3%) sugiere una participación pasiva. Para fomentar una interacción más activa, se recomienda desarrollar contenido interactivo, encuestas y tácticas que incentiven a los usuarios a expresar sus opiniones.

Análisis de Tópicos:

El tópico "Educación y Estudiantes" destaca al generar la mayor cantidad de comentarios positivos (17.88%), indicando una percepción generalmente favorable. Por otro lado, "Género y Ciencias" muestra una proporción equilibrada entre comentarios neutros y positivos, a pesar de tener una participación más baja.

Distribución de Comentarios por Tópico:

La popularidad fluctuante de los tópicos a lo largo de los meses resalta la importancia de adaptar las estrategias de contenido según las tendencias y preferencias cambiantes del público.

Estrategias de Contenido:

Octubre se destaca como el mes con la mayor cantidad de comentarios, sugiriendo la relevancia de eventos o temas específicos durante ese período. Noviembre, al

mostrar una disminución, ofrece la oportunidad de evaluar y ajustar las estrategias de contenido.

Participación por Día de la Semana:

Identificar patrones en la respuesta del público según el día de la semana puede informar sobre la planificación estratégica de publicaciones. Se sugiere lanzar contenido importante en los días más activos, como jueves y sábado.

Análisis de la Distribución Mensual de Comentarios:

La disminución en la actividad de noviembre puede atribuirse a factores estacionales o eventos específicos de ese mes. Evaluar y ajustar las estrategias de contenido durante períodos similares puede ser beneficioso.

Análisis de Sentimientos por Día de la Semana:

El análisis de sentimientos revela que los días jueves y sábado son los más activos en términos de comentarios positivos y negativos. Ajustar las estrategias según estos patrones puede mejorar la participación.

Análisis de Comentarios por Tópico y Mes:

La investigación de la disminución en los comentarios de "Artes y Reconocimientos" en noviembre proporciona una oportunidad para comprender las razones detrás de este cambio y ajustar estrategias en consecuencia.

Recomendaciones Generales:

- Se sugiere un monitoreo continuo de las tendencias y ajustes periódicos en las estrategias.
- La inclusión de contenido interactivo y variado puede enriquecer la participación.
- Habilitar los comentarios en YouTube puede promover una relación más activa con la comunidad.
- Complementar análisis cuantitativos con información cualitativa sobre el contenido específico discutido en cada tópico.

En su conjunto, estas conclusiones y recomendaciones no solo apuntan a mejorar la presencia en redes sociales de la Universidad de Playa Ancha, sino también a establecer un canal efectivo de retroalimentación con su entorno. Reconociendo la importancia del feedback generado en plataformas digitales, se vislumbra la oportunidad de obtener mejoras sustanciales tanto a nivel digital como territorial.

Este proceso no solo impulsa la participación activa de la audiencia, sino que también abre la puerta a posibles eventos, investigaciones y mejoras contextuales que aborden de manera integral el entorno en el que la universidad está inmersa, generando así un impacto significativo en su comunidad.

Capítulo 5: Conclusiones

La culminación de la investigación, centrada en la evaluación de la efectividad de los canales de comunicación en las redes sociales de la Universidad de Playa Ancha (UPLA), ha proporcionado conocimientos valiosos que no solo alcanzan los objetivos establecidos, sino que también abren puertas hacia futuras aplicaciones y mejoras significativas en la gestión universitaria.

La elección de la metodología, específicamente la implementación de KDD (Knowledge Discovery in Databases), se fundamentó en su enfoque integral y en la experiencia previa con esta técnica en análisis de menor escala. La selección de KDD se justificó por su generalidad, evitando términos excesivamente específicos que podrían desviar del objetivo. Además, la familiaridad previa con esta metodología garantizó una ejecución eficiente del proceso.

El resultado del proceso fue altamente satisfactorio, destacando la generación de información crucial para la toma de decisiones. La planificación emergió como un elemento crítico, subrayando la necesidad de una estrategia detallada para la ejecución exitosa de la metodología. Esta fase inicial no solo facilitó la recolección de datos significativos, sino que también resaltó la importancia de un enfoque estructurado para garantizar la coherencia y validez de los resultados.

En el examen de los elementos descubiertos, se destaca la trascendencia de las métricas en la evaluación de los canales de comunicación de la UPLA en entornos digitales. Estas métricas proporcionaron una visión detallada del alcance, visibilidad y percepción general de la institución. La implementación efectiva de métricas específicas para evaluar la percepción de las comunicaciones, analizar el alcance y la visibilidad de las publicaciones, así como realizar un análisis de sentimientos, permitió una evaluación integral de los canales de comunicación de la UPLA.

En relación con los objetivos establecidos, se confirma su logro exitoso:

1. Evaluar la efectividad de los canales de comunicación: Participación Activa vs. Pasiva, la mayor parte de las publicaciones (84.3%) no recibió comentarios, indicando una participación pasiva por parte de la comunidad. Esto podría sugerir que la mayoría de las publicaciones no generan suficiente interacción o que el contenido no motiva a los usuarios a comentar.

El 15.7% de las publicaciones con comentarios muestran una participación activa, lo cual es valioso ya que indica que cierto contenido genera respuestas o interacciones directas. Explorar qué tipo de publicaciones reciben comentarios podría ofrecer perspectivas valiosas sobre el contenido que resuena mejor con la audiencia.

La evaluación de las "Estrategias para Fomentar la Participación" revela información crucial sobre la interacción de la comunidad en los canales de comunicación de la Universidad de Playa Ancha (UPLA). Al analizar la participación activa y pasiva en las publicaciones, se observa que la mayoría de ellas (84.3%) no genera comentarios, indicando una participación pasiva. Este hallazgo sugiere que el contenido puede no estar generando la interacción deseada o no motivar a los usuarios a expresar sus opiniones.

El 15.7% de las publicaciones con comentarios demuestra una participación activa, lo que señala que cierto contenido específico genera respuestas o interacciones directas. Para comprender mejor este fenómeno, se podría explorar en detalle qué tipo de publicaciones reciben comentarios. Este análisis proporciona perspectivas valiosas sobre el contenido que resuena mejor con la audiencia y las posibles áreas de mejora para incrementar la participación activa en los canales de comunicación de la UPLA.

2. Revisar y determinar métricas para medir la percepción de las comunicaciones: El Análisis de la Distribución del Tono Emocional en los Comentarios revela patrones significativos en la percepción de las comunicaciones de la Universidad de Playa Ancha (UPLA). La visualización mediante el gráfico de barras destaca tres tonos emocionales principales: positivo, neutro y negativo.

El Tono Emocional Positivo (POS), con 241 comentarios (47.35%), emerge como la categoría dominante, sugiriendo una respuesta generalmente positiva de la comunidad universitaria. Este hallazgo subraya la efectividad de ciertos elementos comunicativos que generan una reacción favorable por parte de los usuarios.

El Tono Emocional Neutro (NEU), que abarca 235 comentarios (46.17%), se sitúa en proximidad al tono positivo en términos de cantidad. Esto indica una respuesta neutral o una falta de carga emocional significativa en los comentarios, sugiriendo una percepción moderada o imparcial hacia las comunicaciones de la UPLA.

Por último, el Tono Emocional Negativo (NEG), con 33 comentarios (6.48%), representa la menor proporción. Este hallazgo indica una respuesta generalmente negativa o crítica por parte de algunos usuarios, destacando áreas donde las comunicaciones pueden mejorar para evitar percepciones desfavorables.

En conjunto, este análisis proporciona una evaluación detallada de la percepción emocional en los comentarios, permitiendo identificar áreas de fortaleza y oportunidades de mejora en las estrategias comunicativas de la UPLA.

3. Analizar el alcance y visibilidad de las publicaciones: El Análisis del Alcance y Visibilidad de las Publicaciones destaca patrones distintivos en la percepción de diversos tópicos abordados por la Universidad de Playa Ancha (UPLA). A continuación, se resumen los hallazgos más relevantes:

Arte y Reconocimiento:

- Destaca por la mayor cantidad de comentarios negativos (2.95% del total), señalando áreas de mejora en la percepción de esta categoría.
- Presenta la mayor proporción de comentarios neutrales (13.75%), indicando una respuesta moderada hacia las publicaciones en este tópico.
- Se posiciona como el segundo tópico con la mayor cantidad de comentarios positivos (11.79%), revelando áreas de contenido que generan una respuesta positiva en la audiencia.

Historia y Memoria:

- Sobresale como el segundo tópico con más comentarios neutros (13.16%), sugiriendo una respuesta moderada o imparcial de la audiencia.
- Se encuentra en la cuarta posición en comentarios negativos (1.38%), mostrando una recepción generalmente favorable, aunque con menor participación.

Educación y Estudiantes:

- Destaca al contar con la mayor cantidad de comentarios positivos (17.88% del total), indicando una percepción generalmente favorable.
- Se ubica como el segundo tópico con la menor cantidad de comentarios negativos (0.78%), evidenciando una recepción positiva y una baja crítica en esta categoría.
- Exhibe una notable cantidad de comentarios neutros (11.79%), sugiriendo una respuesta moderada hacia las publicaciones relacionadas con la educación y los estudiantes.

Actividades y Disciplinas Físicas:

- Se posiciona como el segundo tópico con menos participación, recopilando un total de 62 comentarios. Aunque la participación es menor, se destaca la necesidad de explorar estrategias para aumentar la interacción en este ámbito.

Género y Ciencias:

- Representa el tópico con menor participación (6.1% del total de comentarios). A pesar de la baja participación, muestra una proporción equitativa entre comentarios neutros y positivos, con solo un comentario negativo. Se sugiere explorar estrategias para fomentar la participación en este tema específico.

A partir de los análisis detallados sobre la percepción y participación en diversos tópicos, se identifican oportunidades clave para orientar futuras estrategias de la Universidad de Playa Ancha (UPLA) en redes sociales. Estas observaciones brindan valiosa información para mejorar la relevancia y el impacto de las publicaciones, así como para fortalecer la conexión con la comunidad.

4. Realizar análisis de sentimientos:

El análisis de sentimientos proporciona una visión detallada de la respuesta emocional de la comunidad en las distintas plataformas sociales de la Universidad de Playa Ancha (UPLA). Los resultados, presentados de manera clara en el gráfico de barras apiladas, permiten extraer conclusiones significativas sobre la percepción general en cada red social.

Facebook: Con un total de 199 comentarios, Facebook muestra una distribución equilibrada. Aunque la proporción de comentarios negativos es mínima (4.52%), es crucial observar la participación activa en comentarios positivos (55.28%). Este hallazgo sugiere un compromiso positivo por parte de la audiencia en esta plataforma, destacando la importancia de fomentar interacciones constructivas.

Instagram: El análisis en Instagram, con un total de 307 comentarios, revela una dinámica interesante. A pesar de un leve aumento en comentarios negativos (7.82%), la plataforma destaca por la significativa participación neutra (50.16%). Este equilibrio entre comentarios neutros y positivos (42.02%) subraya la importancia de abordar contenidos que generen tanto interacción como respuestas emocionales.

YouTube: Si bien YouTube presenta un conjunto más reducido de comentarios (3 en total), la plataforma destaca por la ausencia de comentarios negativos y la predominancia de comentarios positivos (66.67%). Aunque la participación es menor en comparación con otras plataformas, este resultado sugiere un ambiente generalmente favorable en YouTube.

En conjunto, estos resultados resaltan la diversidad de respuestas emocionales en las distintas redes sociales de la UPLA. Es fundamental considerar estas variaciones al diseñar estrategias de comunicación, adaptándolas a las dinámicas específicas de cada plataforma para fomentar una conexión más efectiva con la audiencia.

Este análisis destaca la prevalencia del tono emocional neutro y positivo en las tres redes sociales analizadas. Sin embargo, se observa que Instagram registra la mayor cantidad de comentarios negativos, seguida por Facebook. En contraste, YouTube presenta una proporción significativa de comentarios positivos. Este panorama refleja la complejidad de las interacciones emocionales en distintos entornos

digitales y subraya la importancia de comprender las dinámicas específicas de cada plataforma.

Dirigiendo la mirada hacia el futuro, se vislumbra un papel trascendental para la minería de datos en el ámbito universitario. Más allá de la expansión de la investigación actual, la aplicación constante de técnicas de minería de datos podría trascender a la identificación de posibles carreras de interés en la zona territorial.

La capacidad predictiva inherente a la minería de datos no solo se limitaría a la evaluación del presente, sino que se proyecta hacia el futuro, anticipando posibles escenarios como el abandono de carreras por parte de los alumnos. Este enfoque visionario podría brindar una herramienta invaluable para diseñar estrategias de retención y acciones preventivas, abriendo un abanico de posibilidades para moldear de manera proactiva el rumbo de la educación universitaria.

La presente investigación, al sumergirse en la exploración de la presencia digital de la UPLA, ha sentado cimientos sólidos para futuras mejoras y aplicaciones de la minería de datos en la gestión universitaria. Al alcanzar y superar los objetivos establecidos, este proyecto de título no solo contribuye al entendimiento actual de la efectividad de las estrategias de comunicación, sino que también despliega sus alas hacia el futuro, destacándose por su capacidad de anticipación y prevención de desafíos potenciales. En este horizonte emergente, la minería de datos se erige como una brújula orientadora, señalando direcciones estratégicas para potenciar la conexión con la comunidad universitaria, adaptarse a las dinámicas cambiantes del entorno digital y modelar un futuro educativo más resiliente y eficiente.

Así, esta investigación no solo se erige como una instantánea del presente digital de la UPLA, sino como un precursor visionario que proyecta el camino que la institución puede recorrer en la era de la información. Los hallazgos y metodologías aquí desarrollados no sólo capturan el pulso actual de la comunicación universitaria, sino que también sientan las bases para un futuro en el cual la minería de datos se convierte en un aliado estratégico, impulsando la toma de decisiones informadas y la construcción proactiva del mañana educativo. En este trayecto, la investigación no concluye; más bien, se transforma en el primer paso de una nueva fase, una fase donde la universidad, equipada con las herramientas avanzadas de la minería de datos, se aventura audazmente hacia un horizonte de innovación continua y mejora constante.

Este horizonte amplio y prometedor también abre espacio para explorar nuevas aplicaciones de la minería de datos en la gestión universitaria. Podría profundizar en la personalización de la experiencia estudiantil, utilizando análisis detallados para adaptar los programas académicos a las necesidades individuales de los alumnos. Asimismo, la predicción de tendencias académicas y la identificación temprana de posibles casos de ausentismo se presentan como áreas estratégicas para mejorar la eficiencia y el bienestar estudiantil. Estas perspectivas delinean un panorama en el

cual la minería de datos no solo responde a las demandas presentes, sino que también se erige como un faro que guía hacia un futuro universitario más inclusivo, resiliente y acorde con las expectativas cambiantes de la comunidad educativa.

La minería de datos se posiciona como una herramienta esencial para el futuro, desplegando oportunidades significativas para el crecimiento continuo y la toma de decisiones informadas en el ámbito universitario. Este estudio no solo establece cimientos sólidos, sino que también allana el camino hacia una gestión universitaria más proactiva y guiada por datos. En un entorno educativo en constante evolución, esta perspectiva innovadora no solo promete adaptación, sino también prosperidad, marcando el inicio de una nueva era donde la universidad se nutre de la riqueza de la información para moldear un futuro educativo más eficiente y orientado al éxito.

Referencias Bibliográficas

- Amazon. (n.d.). *¿En qué consiste la limpieza de datos? - Explicación de la limpieza de datos*
- AWS. AWS. Retrieved 2023, from
<https://aws.amazon.com/es/what-is/data-cleansing/>
- Barney, N. (2023, Enero). *What Is Sentiment Analysis (Opinion Mining)? | Definition from TechTarget*. TechTarget. Retrieved 2023, from
<https://www.techtarget.com/searchbusinessanalytics/definition/opinion-mining-sentiment-mining>
- Beto. (n.d.). *dccuchile/beto: BETO - Spanish version of the BERT model*. GitHub. Retrieved 2023, from <https://github.com/dccuchile/beto>
- Biblioteca del congreso Nacional de Chile. (2022, DIC 28). *Ley Chile - Ley 21094 - Biblioteca del Congreso Nacional de Chile*. BCN. Retrieved Abril 20, 2023, from
<https://www.bcn.cl/leychile/navegar?idNorma=1119253&idVersion=2022-12-28&idParte=9920281>
- Cambria et al. (2012). *Sentic Computing: Techniques, Tools, and Applications*. Retrieved 2023, from
https://books.google.cl/books?id=8DPLZ8klJrkC&printsec=frontcover&hl=es&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false
- Cantor, L. (2022, Enero). *Growing app complexity: Paving the way for digital lifestyles and immersive experiences*. Sandvine. Retrieved 2023, from
https://www.sandvine.com/hubfs/Sandvine_Redesign_2019/Downloads/2022/Phenomena%20Reports/GIPR%202022/Sandvine%20GIPR%20January%202022.pdf
- Castañeda Luis, Benítez Joshua, Zuniga Carlos y Vásquez Kevin. (2021, September 22). *Big Data ¿Tecnología o una prioridad empresarial para la gestión de datos digitales?*
Retrieved 2023, from
https://www.academia.edu/download/38742054/Paper_Big_Data.pdf

- Chapman, A. (2017, November 9). *Principles and methods of data cleaning*. YouTube: Home. Retrieved 2023, from <https://books.google.es/books?hl=es&lr=&id=44gDJTEJoVIC&oi=fnd&pg=PA1&dq=What+is+data+cleaning&ots=vUfQ-8i1uo&sig=EqohyXizP1tgm6rcL-3N-eKuB5k#v=onepage&q&f=false>
- Datahack. (2020, January 27). *LAS 10 V'S DEL BIG DATA*. Datahack. Retrieved 2023, from <https://www.datahack.es/10-vs-del-big-data/>
- Duarte, F. (2023, December 13). *Amount of Data Created Daily (2024)*. Exploding Topics. Retrieved 2023, from <https://explodingtopics.com/blog/data-generated-per-day>
- Gandhi et al. (2022, Junio 13). *Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and*. SenticNet. Retrieved 2023, from <https://sentic.net/multimodal-sentiment-analysis-review.pdf>
- Hutto, G. (2014). *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*. View of VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. Retrieved 2023, from <https://ojs.aaai.org/index.php/ICWSM/article/view/14550/14399>
- Joyanes, L. (2013). *Big Data Análisis de grandes volúmenes de datos en organizaciones E-Books & Papers for Statisticians*. Academia.edu. Retrieved 2023, from https://www.academia.edu/39908268/Big_Data_An%C3%A1lisis_de_grandes_vol%C3%BAmenes_de_datos_en_organizaciones_E_Books_and_Papers_for_Statisticians
- Kan, Kamber y Pei. (2012). *Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)*. Sabancı University myWeb Service. Retrieved 2023, from <https://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf>

- Lane et al. (2018). *Natural Language Processing in Action: Understanding, analyzing, and generating text with Python MEAP V10 - Chapter 1*. AWS. Retrieved 2023, from https://manning-content.s3.amazonaws.com/download/a/c9fc557-b088-4f1f-87a8-7ea2e488d262/Lane_NLPiA_MEAP_V10_ch1.pdf
- Lui, B. (2011, September 15). *Sentiment Analysis and Opinion Mining*. Computer Science. Retrieved 2023, from <https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf>
- Marr, B. (2021). *How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read*. Bernard Marr. Retrieved 2023, from <https://bernardmarr.com/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/>
- Mckinney, W. (2011, Enero). *Pandas: a Foundational Python Library for Data Analysis and Statistics*. ResearchGate. Retrieved 2023, from https://www.researchgate.net/publication/265194455_pandas_a_Foundational_Python_Library_for_Data_Analysis_and_Statistics
- Medhat, Hassan y Korashy. (2014, Diciembre). *Sentiment analysis algorithms and applications: A survey*. Retrieved 2023, from <https://www.sciencedirect.com/science/article/pii/S2090447914000550>
- Numpy. (n.d.). *NumPy documentation — NumPy v1.26 Manual*. NumPy. Retrieved 2023, from <https://numpy.org/doc/stable/>
- Oracle. (n.d.). *¿Qué es Big Data?* Oracle. Retrieved Abril, 2023, from <https://www.oracle.com/cl/big-data/what-is-big-data/>
- Oracle. (2022). *The Evolution of Big Data and the Future of the Data Platform*. Retrieved Abril, 2023, from <https://www.oracle.com/cl/a/ocom/docs/big-data/big-data-evolution.pdf>
- Pang Lee, B. L. (2008, October 16). *Opinion mining and sentiment analysis*. Cornell Computer Science. Retrieved 2023, from <https://www.cs.cornell.edu/home/llee/omsa/omsa.pdf>

- Pedregosa, Varoquaux, Gramfort, A., et al., (2011, Diciembre 10). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research. Retrieved 2023, from <https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf>
- Pereira, T. (2016). *El proceso de descubrimiento de conocimiento en bases de datos - The Process of Knowledge Discovery on Databases*. Ediciones UCC. Retrieved 2023, from <https://ediciones.ucc.edu.co/index.php/ucc/catalog/download/36/40/230?inline=1>
- pysentimiento/pysentimiento: A Python multilingual toolkit for Sentiment Analysis and Social NLP tasks*. (n.d.). GitHub. Retrieved 2023, from <https://github.com/pysentimiento/pysentimiento>
- Robertuito. (n.d.). *pysentimiento/robertuito: A pre-trained language model for social media text in Spanish*. GitHub. Retrieved 2023, from <https://github.com/pysentimiento/robertuito>
- Sayed, H. (2021, Junio 7). *What is Sentiment Analysis? Types and Use Cases*. TAUS. Retrieved 2023, from <https://www.taus.net/resources/blog/what-is-sentiment-analysis-types-and-use-cases>
- Schmidt, I. B. (2016). *A glimpse inside the mind of a data scientist*. IBM. Retrieved 2023, from <https://www.ibm.com/downloads/cas/W6GEX9LL>
- Tableau. (n.d.). *Big Data Analytics: What It Is, How It Works, Benefits, And Challenges*. Tableau. Retrieved 2023, from <https://www.tableau.com/learn/articles/big-data-analytics>
- Taylor, P. (2023, November 16). *Data growth worldwide 2010-2025*. Statista. Retrieved 2023, from <https://www.statista.com/statistics/871513/worldwide-data-created/>
- transformers*. (n.d.). *huggingface/transformers: 🤗 Transformers: State-of-the-art Machine Learning for Pytorch, TensorFlow, and JAX*. GitHub. Retrieved 2023, from <https://github.com/huggingface/transformers>
- Universidad de Playa Ancha. (n.d.). *Upla. Inicio - Universidad de Playa Ancha - Portada*. Retrieved 2023, from <https://www.upla.cl/portada/>

Universidad de Playa Ancha Noticias. (n.d.). *Noticias de la Universidad de Playa Ancha – Dirección General de Comunicaciones*. Universidad de Playa Ancha. Retrieved 2023, from <https://www.upla.cl/noticias/>

Vasilis, T. (2022, October 20). *What is ethical scraping and how do you do it?* Apify Blog. Retrieved 2023, from <https://blog.apify.com/what-is-ethical-web-scraping-and-how-do-you-do-it/>

Virtanen, Gommers, Oliphant, et al. (2020, Febrero 3). *SciPy 1.0: fundamental algorithms for scientific computing in Python*. Retrieved 2023, from <https://www.nature.com/articles/s41592-019-0686-2#Sec12>

Wiley. (2015). *Storytelling with data: a data visualization guide for business professionals*. Storytelling with Data. Retrieved 2023, from <https://www.asrepayesh.com/assets/asrepayesh.com/repo/file/storytelling-with-data-c-ole-nussbaumer-knaflig.pdf>