**Proyecto de procesamiento de lenguaje natural**

**Objetivo:** Realizar un análisis de sentimientos utilizando técnicas de procesamiento de lenguaje natural (PLN) para analizar comentarios sobre productos y predecir el sentimiento (positivo o negativo) en nuevos datos.

**Base datos:** son comentarios sobre los productos que se han comprado en amazon los últimos 2 años es un archivo parquet y consta de 380,000 filas y 3 columnas, las cuales son

* Title: es el titulo del comentario
* Text: es el comentario
* Polarity: contiene la clasificación {1= negativo, 2= positivo}

**Metodología del Proyecto**: CRISP-DM

**Fase: Entendimiento del negocio**

1. Determinar los objetivos del negocio
   1. Antecedentes

Amazon, como líder en comercio electrónico, recopila una gran cantidad de datos de comentarios generados por usuarios sobre productos. Estos comentarios contienen información valiosa que puede ser utilizada para mejorar la experiencia del cliente, identificar problemas con productos, y personalizar estrategias de marketing. Sin embargo, el volumen y diversidad de estos datos requieren técnicas avanzadas como el procesamiento de lenguaje natural (PLN) para extraer valor.

* 1. Objetivos del negocio

Mejorar la capacidad de respuesta a las reseñas mediante la identificación rápida de comentarios negativos. Generar insights para optimizar productos y estrategias de marketing.

Identificar automáticamente el sentimiento de los comentarios de clientes (positivo o negativo) para mejorar la atención al cliente y priorizar problemas críticos.

Proveer un análisis detallado de tendencias en los comentarios, identificando patrones frecuentes en productos con sentimientos negativos.

Reducir el tiempo y esfuerzo humano dedicado al análisis manual de comentarios.

Mejorar la experiencia del cliente al proporcionar a los equipos de producto y marketing información valiosa sobre los sentimientos de los consumidores.

Reducir el tiempo y los costos asociados con el análisis manual de comentarios.

Implementar un modelo que pueda predecir con precisión los sentimientos de nuevos comentarios en tiempo real.

* 1. Criterios de éxito

 Modelo con una precisión de al menos 85%.

 Reporte claro que identifique palabras clave asociadas con comentarios positivos y negativos.

Implementar un modelo que pueda clasificar comentarios con un nivel de precisión superior al 85%.

Generar reportes comprensibles para los tomadores de decisiones, destacando productos y categorías problemáticas.

Proveer un sistema escalable que pueda analizar comentarios en tiempo real.

Reducción en el tiempo de análisis de comentarios en un 70%.

Aumento en la satisfacción del cliente medido por encuestas post-compra.

1. Situación Actual
   1. Inventario de recursos

Base de datos: Archivo Parquet con 380,000 filas y 3 columnas (Title, Text, Polarity).

Herramientas: Python, bibliotecas (NLTK, SpaCy, scikit-learn, pandas, TensorFlow/PyTorch.

Datos: Archivo Parquet con 380,000 filas y 3 columnas (Title, Text, Polarity).

Tecnología disponible: Python con bibliotecas de PLN (NLTK, SpaCy), scikit-learn, TensorFlow/PyTorch, pandas y herramientas de visualización como Matplotlib y Seaborn.

Personal: Equipo de ciencia de datos con experiencia básica en análisis de sentimientos y procesamiento de lenguaje natural.

Infraestructura: Recursos computacionales disponibles para manejar grandes volúmenes de datos (e.g., Google Colab, AWS).

* 1. Requerimientos

Un pipeline eficiente para preprocesar y transformar datos textuales.

Un modelo de aprendizaje automático entrenado para análisis de sentimientos.

Herramientas para interpretar resultados y proporcionar visualizaciones intuitivas.

Acceso a la base de datos de comentarios.

Infraestructura de computación adecuada para el procesamiento de grandes volúmenes de datos.

Herramientas de visualización para interpretar y comunicar los resultados.

* 1. Supuestos, restricciones, riesgos y contingencias

Los datos son representativos y están equilibrados para ambas clases (positivo y negativo).

Los comentarios están en inglés y no requieren traducción.

No hay datos mal etiquetados en la columna Polarity.

Tiempo limitado para el desarrollo del proyecto (6 semanas).

Recursos computacionales moderados para el entrenamiento de modelos avanzados.

Riesgo: Distribución de clases desbalanceada en los datos.

Mitigación: Aplicar técnicas de remuestreo (e.g., submuestreo, sobremuestreo).

Riesgo: Comentarios con ruido textual (e.g., errores ortográficos).

Mitigación: Aplicar técnicas de limpieza avanzada (e.g., correctores ortográficos automáticos).

Supuestos: Se asume que los datos proporcionados son representativos y completos.

Restricciones: Limitaciones en el presupuesto y en el tiempo disponible.

Riesgos: Posibles problemas con la calidad de los datos, dificultades en la implementación del modelo, y cambios en las necesidades del negocio.

Contingencias: Planes de respaldo en caso de que el modelo inicial no cumpla con los criterios de éxito, incluyendo ajustes en la estrategia de modelado o en la pre-procesamiento de datos.

* 1. Terminología

Sentimiento: Percepción emocional expresada en los comentarios.

Polarity: Clasificación del sentimiento (1 = negativo, 2 = positivo).

NLP (Natural Language Processing): Procesamiento de Lenguaje Natural.

* 1. Costos y beneficios

Costos: Desarrollo del modelo, infraestructura de computación, herramientas de software, salarios del equipo.

Beneficios: Mejor comprensión de las opiniones de los clientes, mejora en la toma de decisiones basada en datos, mayor satisfacción del cliente.

1. Determinar las metas de la minería de datos
   1. Metas de la minería de datos

Desarrollar un modelo que clasifique con alta precisión y sensibilidad los comentarios como positivos o negativos.

Identificar palabras y frases frecuentes asociadas a cada sentimiento.

Diseñar un pipeline de análisis que pueda escalarse para manejar volúmenes crecientes de datos.

Desarrollar un modelo preciso para clasificar los sentimientos de los comentarios.

Identificar patrones y tendencias en los comentarios de los clientes.

Desarrollar un modelo preciso para clasificar los sentimientos de los comentarios.

Identificar patrones y tendencias en los comentarios de los clientes.

* 1. Criterio de éxito de la minería de datos

Métricas del modelo que cumplan o superen:

Precisión: >85%.

Sensibilidad: >80%.

Especificidad: >80%.

Generar visualizaciones que permitan a los stakeholders identificar rápidamente productos problemáticos.

Precisión, recall y F1-score del modelo superior a 0.85.

Tiempo de procesamiento de nuevos comentarios reducido significativamente.

1. Generar el plan de trabajo
   1. Plan del proyecto inicial

Etapa 1: Entendimiento del negocio (1 semana).

Etapa 2: Entendimiento de los datos y preprocesamiento (2 semanas).

Etapa 3: Entrenamiento y evaluación del modelo (2 semanas).

Etapa 4: Implementación y generación de reportes (1 semana).

* 1. Inventario de técnicas y herramientas

Técnicas:

Limpieza de texto: Tokenización, stopwords, lematización.

Representación vectorial: TF-IDF, embeddings preentrenados.

Modelos: Naive Bayes, BERT.

Tokenización, lematización, eliminación de stopwords, análisis de sentimiento con modelos supervisados.

Herramientas:

Python: NLTK, SpaCy, scikit-learn, TensorFlow.

Jupyter Notebook para análisis exploratorio y desarrollo.

Google Colab o AWS para ejecución en la nube.

**Fase: Entendimiento de datos**

1. Recolección inicial de datos

Descripción: Recolectar la base de datos inicial que contiene los comentarios sobre productos comprados en Amazon en los últimos 2 años.

Objetivo: Asegurar que todos los datos necesarios están disponibles para el análisis.

Verificar que los datos están completos y sin errores en la transferencia.

1. Descripción de datos

Descripción: Realizar un análisis exploratorio inicial para comprender las características básicas del conjunto de datos.

Objetivo: Obtener una visión general de la estructura y las características de los datos.

Inspeccionar las columnas Title, Text y Polarity.

Realizar un resumen estadístico de las columnas para entender la distribución de los datos.

Verificar la presencia de valores nulos o datos faltantes

1. Exploración de datos

Identificar distribuciones de Polarity (balance de clases).

Verificar estadísticas descriptivas de las longitudes de Title y Text.

Descripción: Examinar en detalle los datos para descubrir patrones, anomalías y relaciones.

Objetivo: Identificar cualquier problema potencial que pueda influir en el análisis de sentimientos.

Analizar la longitud de los comentarios (Text) y la frecuencia de las palabras.

Visualizar la distribución de la Polarity para entender el balance entre comentarios positivos y negativos.

Realizar análisis de frecuencia de términos en los títulos (Title) y textos (Text).

1. Validación de la calidad de datos

Identificar y manejar valores nulos, duplicados o ruido en los datos.

Descripción: Verificar que los datos son de alta calidad y adecuados para el análisis.

Objetivo: Asegurar que los datos cumplen con los estándares necesarios para el modelado.

Pasos:

Comprobar la coherencia y precisión de los datos.

Detectar y manejar valores atípicos.

Evaluar la calidad del etiquetado en la columna Polarity

**Fase: Preparación de datos**

1. Selección de datos

Elegir columnas Text y Polarity para el análisis principal.

1. Limpieza de datos

Eliminar duplicados y comentarios vacíos.

Normalizar texto: convertir a minúsculas, eliminar puntuación y palabras vacías (stopwords).

1. Construcción de datos
2. Integración de datos
3. Formato de datos

Codificación de la variable objetivo (Polarity) para el modelo (e.g., 1 = negativo, 2 = positivo).

**Fase: Transformación de los datos**

1. Análisis de componentes principales
   1. PCA - Evaluar su necesidad para la representación vectorial de palabras.
2. Discretización de datos
   1. Representación de texto
      1. Métodos clásicos: TF-IDF, Bag-of-Words.
      2. Métodos modernos: Word2Vec, GloVe, o embeddings preentrenados como BERT.
3. Numeración de datos

**Fase: Modelado**

1. Seleccionar la técnica del modelo

Modelos clásicos: Naive Bayes, Logistic Regression.

Modelos avanzados: Redes Neuronales Recurrentes (RNN), Transformers (BERT, RoBERTa).

1. Generar pruebas del diseño
2. Construir el modelo

Separar datos en conjuntos de entrenamiento, validación y prueba.

1. Evaluar el modelo

Evaluar modelos con métricas como Accuracy, F1-Score, y AUC-ROC.

**Fase: Evaluación**

1. Evaluar los resultados

Comparar rendimiento entre modelos.

Realizar análisis de errores para mejorar.

1. Revisar el proceso

Documentar los pasos realizados y resultados obtenidos.

1. Determinar siguientes pasos

**Fase: Implementación**

1. Implementación

Exportar el modelo entrenado en un formato como .pkl.

1. Plan de monitoreo y mantenimiento

Diseñar mecanismos para evaluar el rendimiento del modelo con nuevos datos.

1. Reporte final

Generar visualizaciones sobre las palabras más frecuentes y su relación con los sentimientos.

Documentar las recomendaciones finales.

1. Revisar el proyecto