Tecnicatura Superior en Ciencia de Datos e Inteligencia Artificial

Materia: Técnicas de Procesamiento de Habla

Profesor: Moises Tinte

Integrantes:

- Nicolas González Da Silva
- Lilen Guzmán
- Silvia Carina Monzón
- Gabriela Cáceres

1. Comprensión del negocio

Objetivo del proyecto:

Desarrollar un sistema de análisis de sentimiento de tweets que clasifique correctamente las emociones expresadas en los mismos.

Beneficios:

- Permitir a las empresas comprender mejor las opiniones y sentimientos de sus clientes.
- Ayudar en la monitorización de las redes sociales para detectar cambios en el estado de ánimo del público.
- Proporcionar información útil para campañas de marketing y atención al cliente.

Criterios de éxito:

- Alta precisión en la clasificación de emociones.
- Capacidad de manejar grandes volúmenes de datos en tiempo real.
- Facilidad de uso e integración con otras herramientas de análisis de datos

2. Comprensión de los datos

Recolectar el conjunto de datos:

El dataset proporcionado ya contiene 40,000 tweets con las siguientes columnas: tweet_id, sentiment, y content.

Explorar los datos:

- Inspeccionar el dataset para comprender su estructura y contenido.
- Revisar el balance de clases en la columna sentiment para identificar posibles desbalances.

Identificar problemas de calidad:

- Revisar si hay datos faltantes o duplicados.
- Evaluar la distribución de las clases de sentimiento para ver si hay desbalance.

3. Preparación de los datos

Limpieza y preprocesamiento:

- Eliminar tweets duplicados si existen.
- Limpiar el texto de los tweets eliminando URLs, menciones, hashtags y caracteres especiales.
- Convertir el texto a minúsculas y realizar la tokenización.

Transformaciones necesarias:

- Tokenizar y vectorizar los tweets usando técnicas como TF-IDF o embeddings.
- Dividir los datos en conjuntos de entrenamiento, validación y prueba.

4. Modelado

Selección de algoritmos:

Considerar algoritmos como Naive Bayes, Support Vector Machines (SVM), o redes neuronales recurrentes (RNN) con LSTM.

Diseño del modelo:

- Definir la arquitectura del modelo (si se utiliza una red neuronal).
- Entrenar y ajustar los hiperparámetros del modelo.

5. Evaluación

Evaluar el rendimiento del modelo:

Utilizar métricas como precisión, recall, F1-score, y matriz de confusión.
Validación cruzada:

Realizar validación cruzada para asegurarse de que el modelo generaliza bien.
Ajustes:

Analizar los resultados y realizar ajustes necesarios en el modelo.