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

Requerimientos

Técnicas de procesamiento de habla en el proyecto:

Se utilizaron varias técnicas de procesamiento de lenguaje natural (NLP) para analizar y clasificar las emociones en tweets

- Tokenización: Dividimos el texto de los tweets en palabras individuales o tokens.
- Normalización: Convertimos el texto a minúsculas y eliminamos caracteres no alfabéticos para estandarizar el contenido.
- Eliminación de stopwords: Filtramos palabras comunes que no aportan mucho valor semántico (como "el", "la", "y", etc.).
- Vectorización TF-IDF: Convertimos el texto preprocesado en características numéricas utilizando la técnica de TF-IDF (Term Frequency-Inverse Document Frequency), que mide la relevancia de una palabra en un documento en relación con una colección de documentos.
- *Sobremuestreo*: Utilizamos técnicas de sobremuestreo para abordar el problema del desequilibrio de clases en nuestro conjunto de datos, asegurando que cada clase esté representada de manera más equilibrada.
- *Modelos de Clasificación:* Empleamos modelos de aprendizaje automático como Naive Bayes y Random Forest para entrenar y predecir las emociones en los tweets.

Impacto en la solución propuesta:

- 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.

Metodología, herramientas y tipos de datos implicados:

- **Metodología**

- Recolección de Datos: DataSet del sitio web, Kaggle
- Preprocesamiento de Datos: Aplicamos técnicas de limpieza y normalización del texto.
- Vectorización: Convertimos el texto en características numéricas utilizando TF-IDF.
- Balanceo de Clases: Aplicamos sobremuestreo para equilibrar las clases de emociones.
- Entrenamiento de Modelos: Entrenamos modelos de aprendizaje automático en el conjunto de datos preprocesado.
- Evaluación: Evaluamos el rendimiento de los modelos utilizando métricas como precisión, recall y F1-score.

- **Herramientas**

- Pandas: Para la manipulación y análisis de datos.
- NLTK: Para el preprocesamiento de texto, incluyendo tokenización y eliminación de stopwords.
- Scikit-learn: Para la vectorización TF-IDF, el balanceo de clases y el entrenamiento de modelos de aprendizaje automático.
- Matplotlib: Para la visualización de resultados.
- Google Colab: Para el entorno de desarrollo y ejecución de código.

- **Tipos de Datos**

- Datos de texto: Tweets en bruto que contienen texto libre.
- Etiquetas de Emoción: Categorías de emociones que corresponden a cada tweet.

Avances y división de trabajo de procesamiento del habla entre integrantes.

Integrante 1

- ★ Montar Google Drive y cargar el archivo CSV.
- ★ Realizar un análisis inicial de los datos (mostrar primeras filas, información del dataset y distribución de sentimientos).
- ★ Preprocesamiento básico del texto.
- ★ Crear y mostrar un gráfico de barras que visualice la distribución de los sentimientos en el conjunto de datos.

Integrante 2

- ★ Implementar el preprocesamiento avanzado del texto (limpieza adicional, tokenización y eliminación de stopwords).
- ★ Vectorizar el texto usando TF-IDF.

Integrante 3

- ★ Balancear las clases usando técnicas de sobremuestreo.
- ★ Dividir los datos en conjuntos de entrenamiento y prueba.

Integrante 4

- ★ Entrenar el modelo Naive Bayes.
- ★ Evaluar el modelo y generar un informe de clasificación.

Link: 🗀 Procesamiento del Habla