Avances en la clasificación multinivel de comentarios: modelos y métricas en data science

Álvaro Ruiz Fernández
The Bridge

Tabla de contenidos

1. Contexto y desafíos
2. Preprocesamiento de datos
3. Modelos y evaluación
4. Mejoras y recomendaciones

CONTEXTO Y DESAFÍOS

¿Cuál es nuestro objetivo y qué retos enfrentamos?

- Clasificación de comentarios en función de distintos niveles de toxicidad.

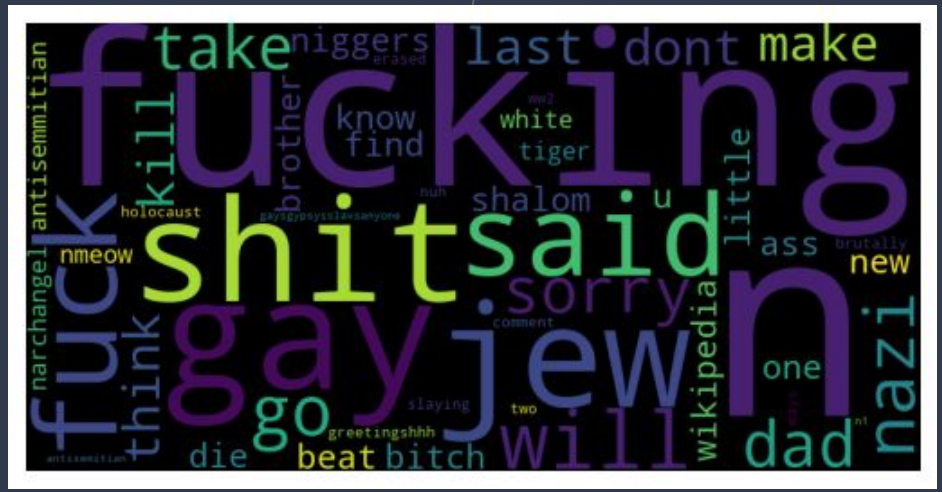
- Desbalance de clases, interpretación de texto, captura de contextos y matices.

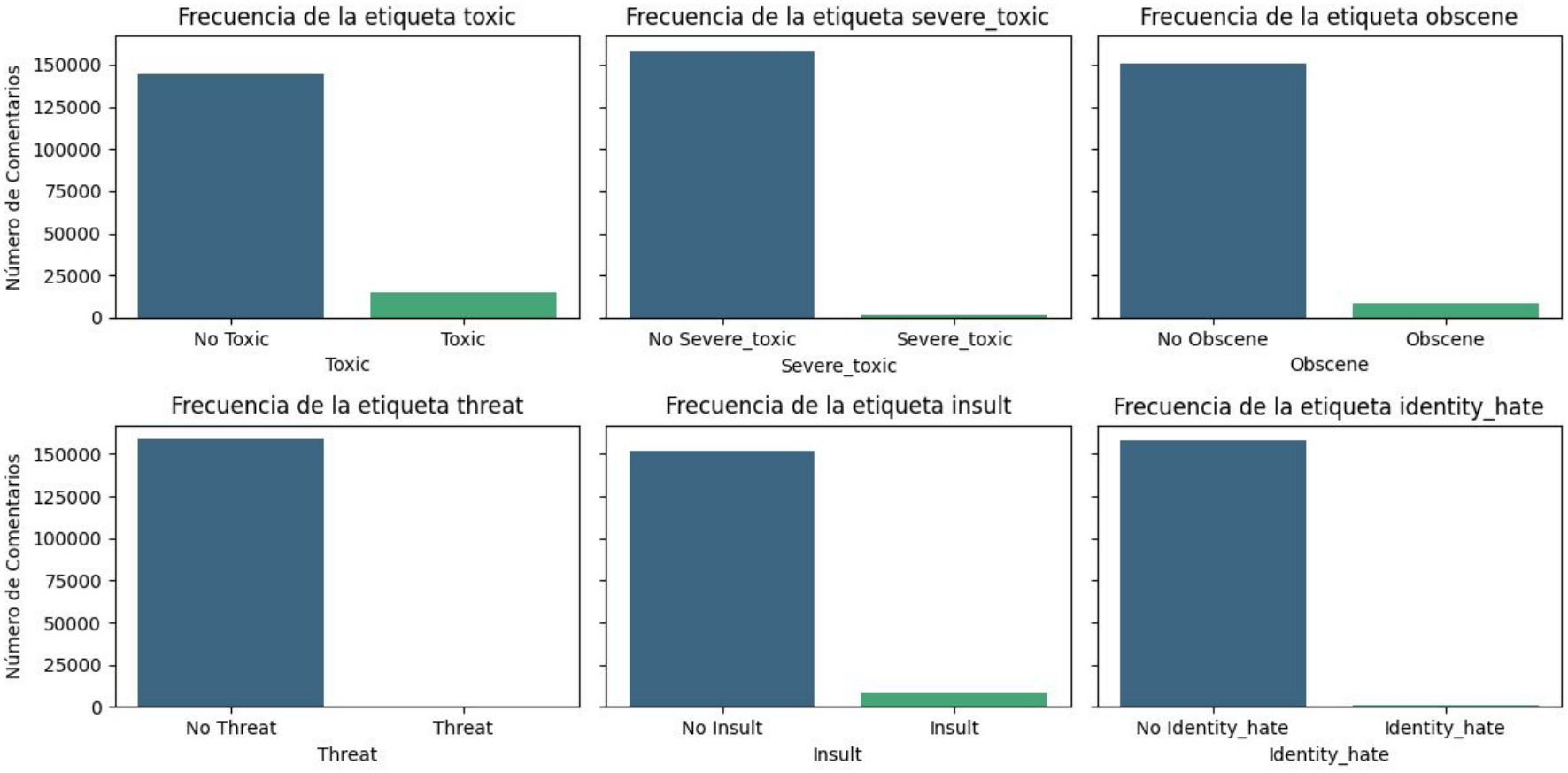
PREPROCESAMIENTO DE DATOS

- Limpieza de datos
- Tokenización: separación de palabras clave
- Vectorización: Conversión de texto a características numéricas

ODIO

TOXICIDAD





DESBALANCEO DE LOS DATOS

MODELOS ELABORADOS

1. Logistic Regression
2. Random Forest
3. Red Neuronal Recurrente
4. K-means clustering

Logistic Regression

- Vectorización con enegramas de 1 y 2 elementos
- Regresión logística multiclase

- Precision (micro): 85%
- Precision (macro): 80%

- Recall (micro): 66%
- Recall (macro): 42%

- Categorical Accuracy: 65%

Random Forest

- Vectorización con n-gramas de 1 y 2 elementos
- Random Forest con 200 árboles y una profundidad de 50

- Precision (micro): 89%
- Precision (macro): 81%

- Recall (micro): 33%
- Recall (macro): 13%

Categorical Accuracy: 56%

Red Neuronal Recurrente

- Red Neuronal Recurrente (RNN) con estructura bidimensional LSTM (long short-term memory)
  - Capa de embedding (32)
  - Capa LSTM bidirec. (32)
  - 3 Capas densas (128-256-128)
  - Capa de salida (6)

- Precision: 77%
- Recall: 70%
- Categorical Accuracy: 47%

COMPARACIÓN DE MODELOS

MEJORAS Y RECOMENDACIONES

- Optimización de hiperparámetros a través de un grid search
- Implementar otras técnicas para el rebalanceo de los datos y/o el reajuste de mismos
- Exploración de modelos más avanzados
    - Implementar la RNN actualizada
    - Implementar Transformer
    - Implementar BERT (Bidirectional Encoder Representations from Transformers)

¡GRACIAS POR VUESTRA ATENCIÓN!