Laboratorio 5 Mineria de texto

Allan Paniagua 18084 Gabriel Garcia 21352

September 2, 2024

Resumen

Este reporte presenta un analisis detallado de un conjunto de datos de tweets relacionados con desastres naturales. Utilizamos técnicas de procesamiento de lenguaje natural (NLP) y diversos modelos de machine learning para clasificar tweets como relacionados o no relacionados con desastres. Los modelos utilizados incluyen Regresión Logística, XGBoost y un Ensemble Voting Classifier, los cuales fueron optimizados utilizando RandomizedSearchCV.

1 Introducción

El objetivo de este estudio es analizar un conjunto de datos de tweets para identificar aquellos que se refieren a desastres naturales. Este tipo de análisis es crucial para agencias de respuesta ante emergencias y organizaciones humanitarias que dependen de información en tiempo real para tomar decisiones efectivas. El conjunto de datos utilizado en este estudio contiene 7613 tweets clasificados como relacionados (1) o no relacionados (0) con desastres.

2 Descripción del conjunto de datos

El conjunto de datos utilizado en este estudio contiene cinco columnas principales: *id*, *keyword*, *location*, *text* y *target*. La columna *text* contiene el contenido del tweet, mientras que *target* indica si el tweet está relacionado con un desastre (1) o no (0). A continuación se muestra una descripción inicial del conjunto de datos:

```
Descripción del conjunto de datos:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7613 entries, 0 to 7612
Data columns (total 5 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   id        7613 non-null   int64
 1   keyword   7552 non-null   object
```

```
 2   location  5080 non-null   object
 3   text      7613 non-null   object
 4   target    7613 non-null   int64
dtypes: int64(2), object(3)
memory usage: 297.5+ KB
None
```

3 Preprocesamiento de datos

El preprocesamiento de datos incluyo la eliminación de URLs, signos de puntuación, números y palabras no informativas específicas. También se realizó la lematización de las palabras y la eliminación de *stopwords*. A continuación se muestra una vista previa de los datos después del preprocesamiento:

```
Datos después del preprocesamiento:
\begin{tabular}{|l|l|}
\hline
\textbf{text} & \textbf{cleaned\_text} \\
\hline
our deeds are the reason of this earthquake may allah forgive & deed reason earthquake may a
forest fire near la ronge sask. canada & forest fire near la ronge sask canada \\
all residents asked to 'shelter in place' are notified officer & resident asked shelter plac
13,000 people receive wildfires evacuation order california & people receive wildfire evacua
just got sent this photo from ruby alaska as smoke wildfire pour & got sent photo ruby alask
\hline
\end{tabular}
```

4 Análisis exploratorio de datos

Se realizaron análisis de unigramas y bigramas para identificar las palabras y frases más comunes en los tweets de desastres y no desastres. A continuación se presentan las palabras más comunes en ambos tipos de tweets:

```
Palabras más comunes en tweets de desastres:
[('fire', 262), ('news', 140), ('via', 121), ('disaster', 118), ('california', 115), ('suici

Palabras más comunes en tweets de no desastres:
[('new', 168), ('ha', 151), ('dont', 141), ('one', 135), ('body', 116), ('time', 105), ('vic
```

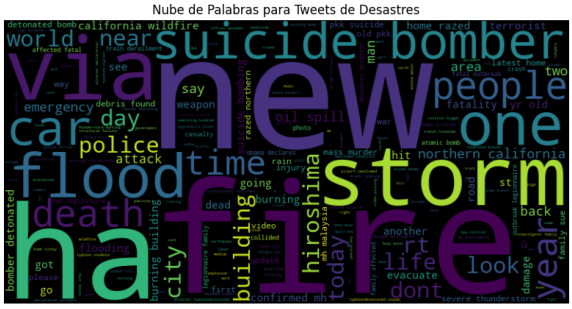

Figure 1: Nube de palabras para Tweets de desastres

Figure 2: Nube de palabras para Tweets de No desastres

5 Modelos de predicción y resultados

Se entrenaron varios modelos de machine learning incluyendo Regresión Logística, XGBoost y un Ensemble Voting Classifier. Los modelos fueron evaluados utilizando métricas como precisión, recall, y curva ROC.

5.1 Regresión logística

Informe de clasificación para Logistic Regression Optimizado:

	precision	recall	f1-score	support
0	0.80	0.86	0.83	874
1	0.79	0.71	0.75	649
accuracy			0.80	1523

macro avg	0.80	0.79	0.79	1523
weighted avg	0.80	0.80	0.80	1523

5.2 XGBoost

Informe de clasificación para XGBoost Optimizado:

	precision	recall	f1-score	support
0	0.79	0.87	0.83	874
1	0.79	0.68	0.73	649
accuracy			0.79	1523
macro avg	0.79	0.78	0.78	1523
weighted avg	0.79	0.79	0.79	1523

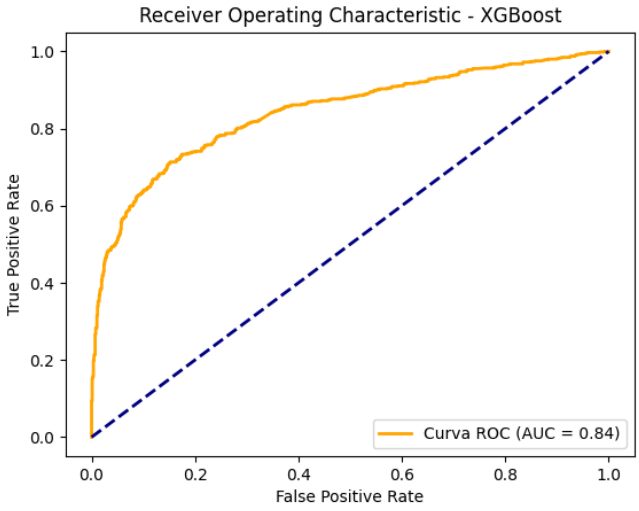

Figure 3: Curva Precision-Recall - XGBoost

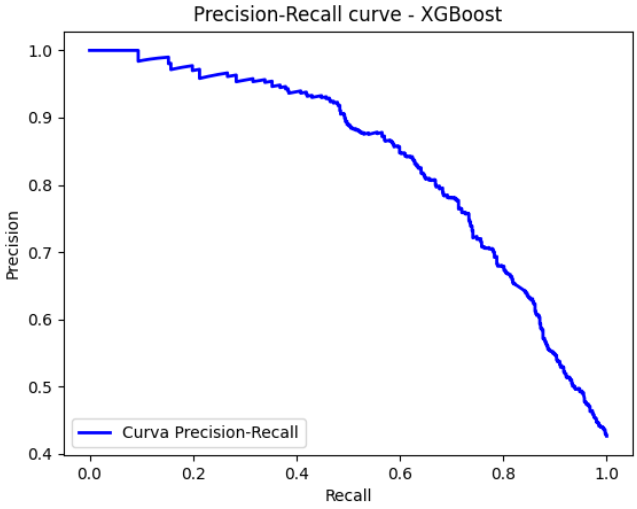

Figure 4: Curva ROC - XGBoost

6 Conclusiones

Los modelos desarrollados fueron capaces de clasificar tweets relacionados con desastres con una alta precisión y recall. XGBoost mostró un mejor rendimiento general en comparación con la regresión logística especialmente en términos de curva ROC y precision-recall. Estos resultados demuestran el potencial del machine learning en la clasificación de textos en tiempo real para aplicaciones críticas como la respuesta a emergencias.

7 Link de Github

https://github.com/Gegdgt/Lab5-Data-Science