INFORME DE TRABAJO FINAL

Aplicaciones de Data Science

Carrera de Ciencias de la Computación

Sección: 279

Docente:
Carlos Fernando Montoya Cubas

Alumnos:
Aragón Ovalle, Alfredo Mauricio - u202210494
Morales Oliveros, Tarik Gustavo - u202210472
Ramírez Cesti, Rodrigo Alonso - u202210690
Rivas Siesquén, Eduardo José - u202216407

Julio 2025

Índice

Descripción del caso de uso

El discurso de odio en las redes sociales es un fenómeno creciente que se ha relacionado con un aumento de síntomas de ansiedad y estrés, afectando la salud mental de los usuarios que lo reciben, así como con el deterioro de la calidad del debate público y de la libre expresión en línea (Saha et al., 2019)

Twitter, en particular, ha enfrentado diversas críticas por su incapacidad para moderar eficazmente contenido tóxico y violento, lo que ha contribuido a la proliferación del discurso de odio y la desinformación (ADL., 2023).

El análisis de sentimientos se presenta como una herramienta clave para abordar este problema, ya que nos permite identificar y clasificar automáticamente el sentimiento expresado por los usuarios en sus tweets. Esto facilita tanto la mejora de políticas de moderación en las plataformas como la investigación académica sobre dinámicas de opinión pública (Pujari & Malik, 2024).

Al identificar patrones de polaridad en los tweets, es posible comprender mejor la evolución temporal y el impacto social de conversaciones que giran en torno a temas sensibles (Dreißigacker et al., 2024).

En este proyecto, se utilizará el conjunto de datos Sentiment140, compuesto por 1600000 tweets originalmente etiquetados como negativos (0), neutrales (2) o positivos (4) (TensorFlow., s. f.).

La versión de los datos empleada en este trabajo ha sido transformada para contener una columna binaria “positive” que indica si la polaridad es positiva (True) o no (False).

Cada registro incluye la marca de tiempo (date), el identificador del usuario (user), el texto original del tweet (text) y la previamente mencionada etiqueta de polaridad.

El objetivo de este trabajo es desarrollar modelos de aprendizaje automático que nos permitan clasificar estos mensajes según su polaridad y estimar probabilidades de positividad, lo que nos ayudaría a:

1. Proveer dashboards de vigilancia que alerten sobre picos de sentimiento negativo.
2. Informar a equipos de moderación sobre la urgencia de intervención en determinados hashtags o temas.
3. Apoyar a estudios de salud pública en la medición de la relación entre toxicidad en línea y bienestar mental.

Se plantean tres preguntas de clasificación/predicción:

1. ¿El texto de un tweet dado expresa un sentimiento positivo o negativo?
2. ¿Un usuario específico publica mayoritariamente tweets positivos o negativos?
3. ¿Cuál es la probabilidad de que un tweet dado sea positivo (en una escala de 0 a 1)?

Responder estas preguntas nos permitirá no solo etiquetar tweets, sino también calibrar umbrales de decisión basados en probabilidades, lo cual es útil para el análisis de tendencias y tomar acciones automatizadas.

Descripción del conjunto de datos

El conjunto de datos Sentiment140 fue creado en 2009 por investigadores de Stanford. El mismo contiene 1 600 000 tweets en inglés obtenidos mediante la API de Twitter, balanceados inicialmente en 800 000 positivos y 800 000 negativos, además de una pequeña porción de neutrales.

Originalmente, cada tweet se clasifica en tres categorías de polaridad: negativo (0), neutral (2) o positivo (4). Sin embargo, debido a la poca cantidad de datos con la etiqueta de neutral, en este proyecto se decidió binarizar la etiqueta en la columna “positive” (True para positivos, False para negativos) (TensorFlow., s. f.).

El archivo CSV original cuenta con 6 columnas de datos:

- target: polaridad original del tweet (0 = negativo, 2 = neutral, 4 = positivo)
- id: identificador del tweet
- date: fecha de publicación
- query: término de búsqueda (o NO_QUERY)
- user: nombre de usuario del autor del tweet
- text: contenido completo del tweet

Para adaptarse a un análisis binario, se reemplazó la columna target por la columna “positive”, derivada de polarity, que indica con un valor booleano si el tweet se considera positivo (True) o no (False).

Todos los campos son obligatorios y no presentan valores faltantes, lo que facilita las tareas de carga y limpieza inicial.

A nivel de texto, los datos son semiestructurados, pues combinan información estructurada (fechas, usuarios, etiquetas) con contenido no estructurado (contenido del tweet).

Se presentan algunos datos duplicados; sin embargo, debido a que representan poco más del 0.01% del total, estos podrán ser eliminados sin inconvenientes en la fase de preprocesamiento.

Análisis exploratorio de los datos (EDA)

Como previamente se ha mencionado, el dataset consiste en 1600000 tweets en ingles.

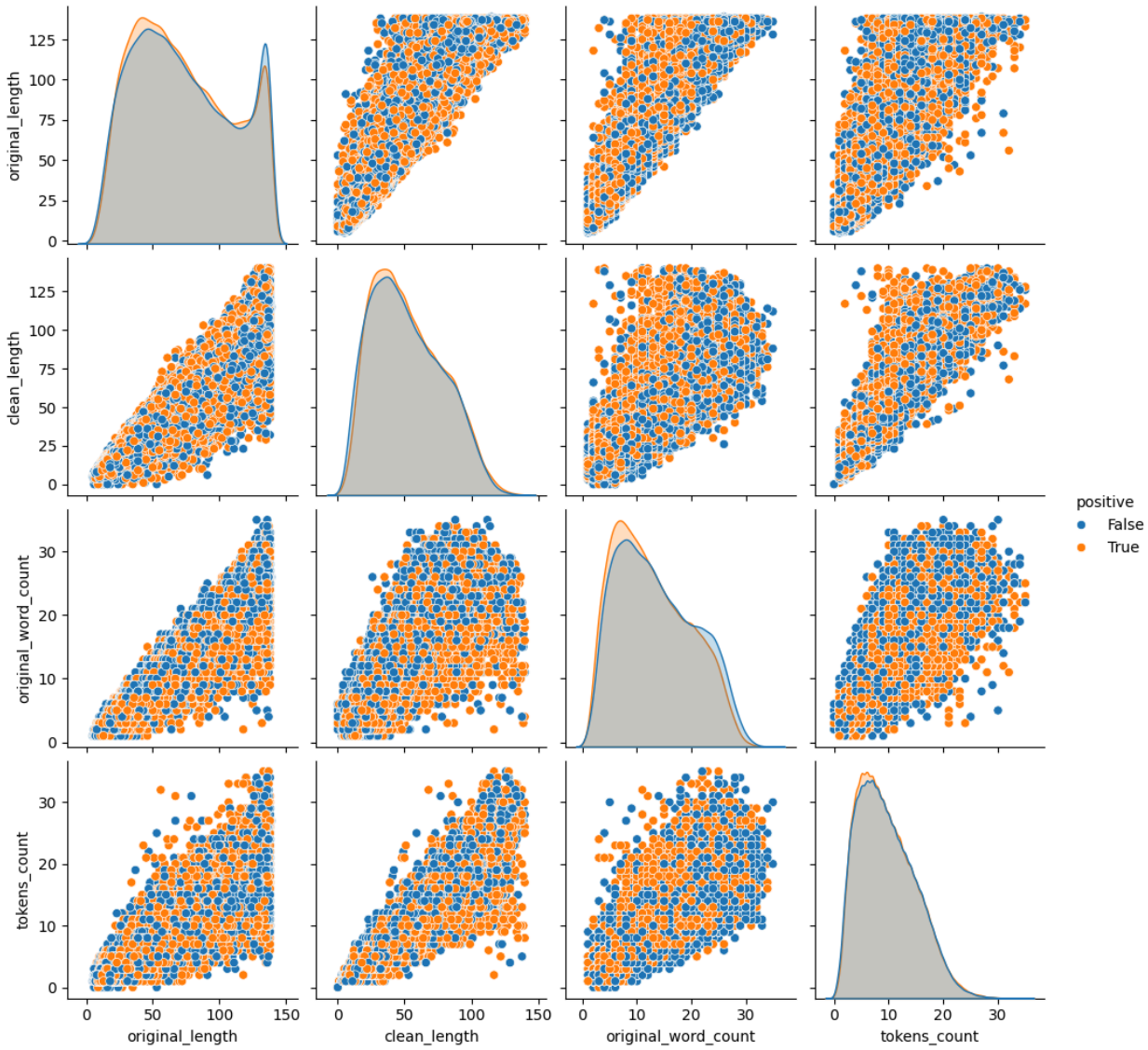
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1600000 entries, 0 to 1599999
Data columns (total 4 columns):
 #   Column  Non-Null Count    Dtype 
---  ------  --------------    ----- 
 0   date    1600000 non-null  object
 1   user    1600000 non-null  object
 2   text    1600000 non-null  object
 3   target  1600000 non-null  int64 
dtypes: int64(1), object(3)
memory usage: 48.8+ MB
```

Como parte del preprocesamiento de datos, se crearon nuevas columnas para:

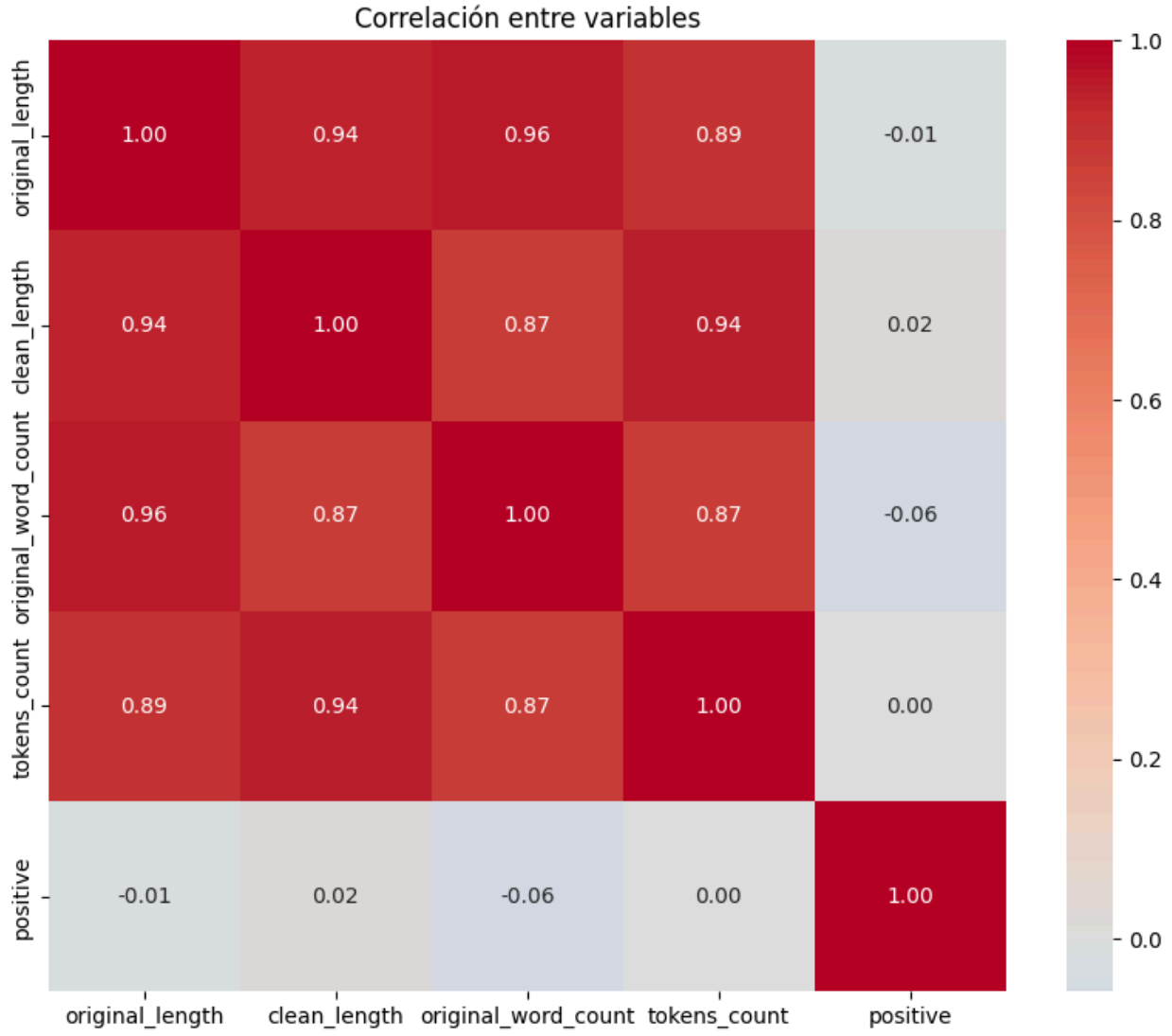
- El texto limpio y lematizado en tokens
- La longitud del texto original en número de caracteres
- La longitud del texto limpio en número de caracteres
- La cantidad de palabras en el texto original
- La cantidad de tokens en el texto limpio

	date	user	text	positive	tokens	original_length	clean_length	original_word_count	tokens_count
0	2009-04-06 22:19:45	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, t...	False	[@switchfoot, -, awww, ,, that's, bummer, ., s...	115	74	19	15
1	2009-04-06 22:19:49	scotthamilton	is upset that he can't update his Facebook by ...	False	[upset, can't, update, facebook, texting, ...,...	111	83	21	15
2	2009-04-06 22:19:53	mattycus	@Kenichan I dived many times for the ball. Man...	False	[@kenichan, dived, many, time, ball, ., manage...	89	64	18	13
3	2009-04-06 22:19:57	ElleCTF	my whole body feels itchy and like its on fire	False	[whole, body, feel, itchy, like, fire]	46	31	10	6
4	2009-04-06 22:19:57	Karoli	@nationwideclass no, it's not behaving at all....	False	[@nationwideclass, ,, behaving, ., mad, ., ?, ...	110	49	21	10

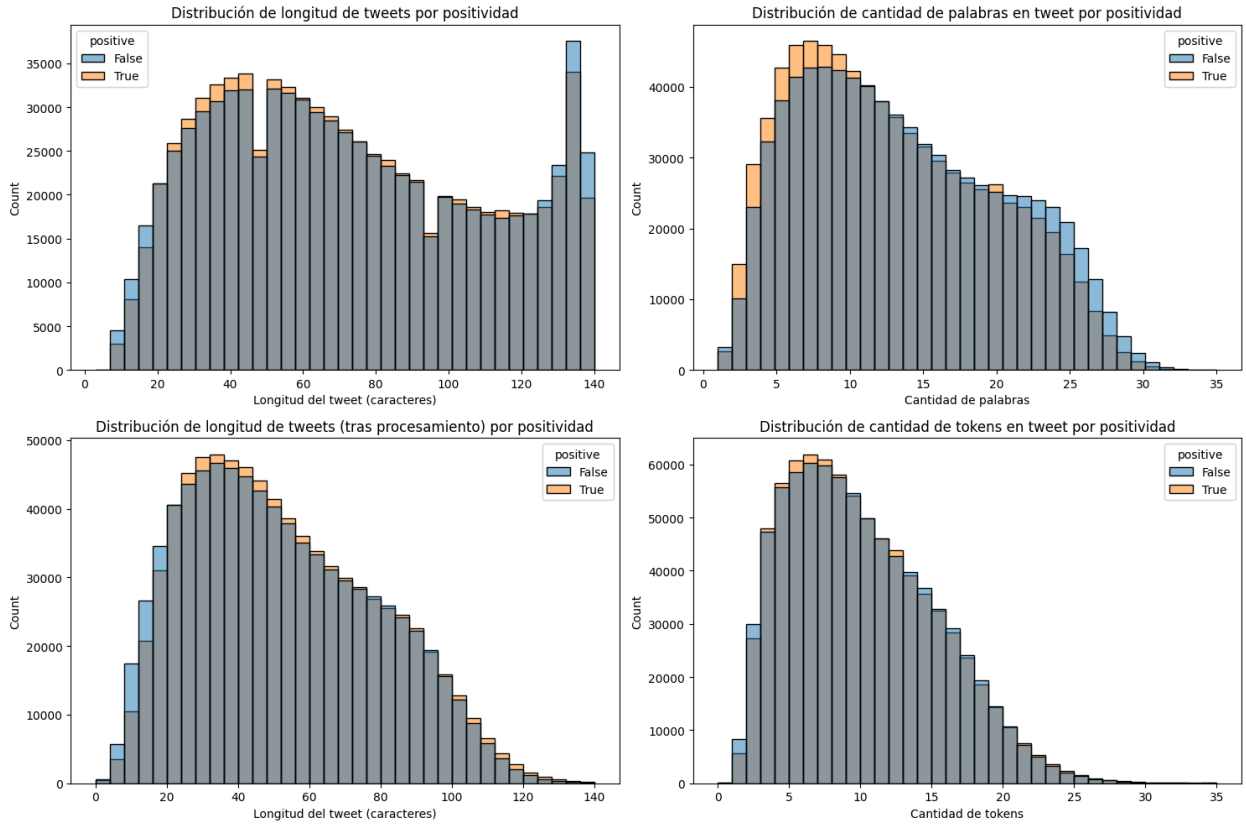
Estos datos numéricos fueron graficados en un pairplot, lo que nos mostró que no existe a simple vista una relación directa entre la longitud del texto y su positividad, lo cuál era algo que ya estábamos considerando de antemano.


original_length
clean_length
original_word_count
tokens_count
positive
False
True

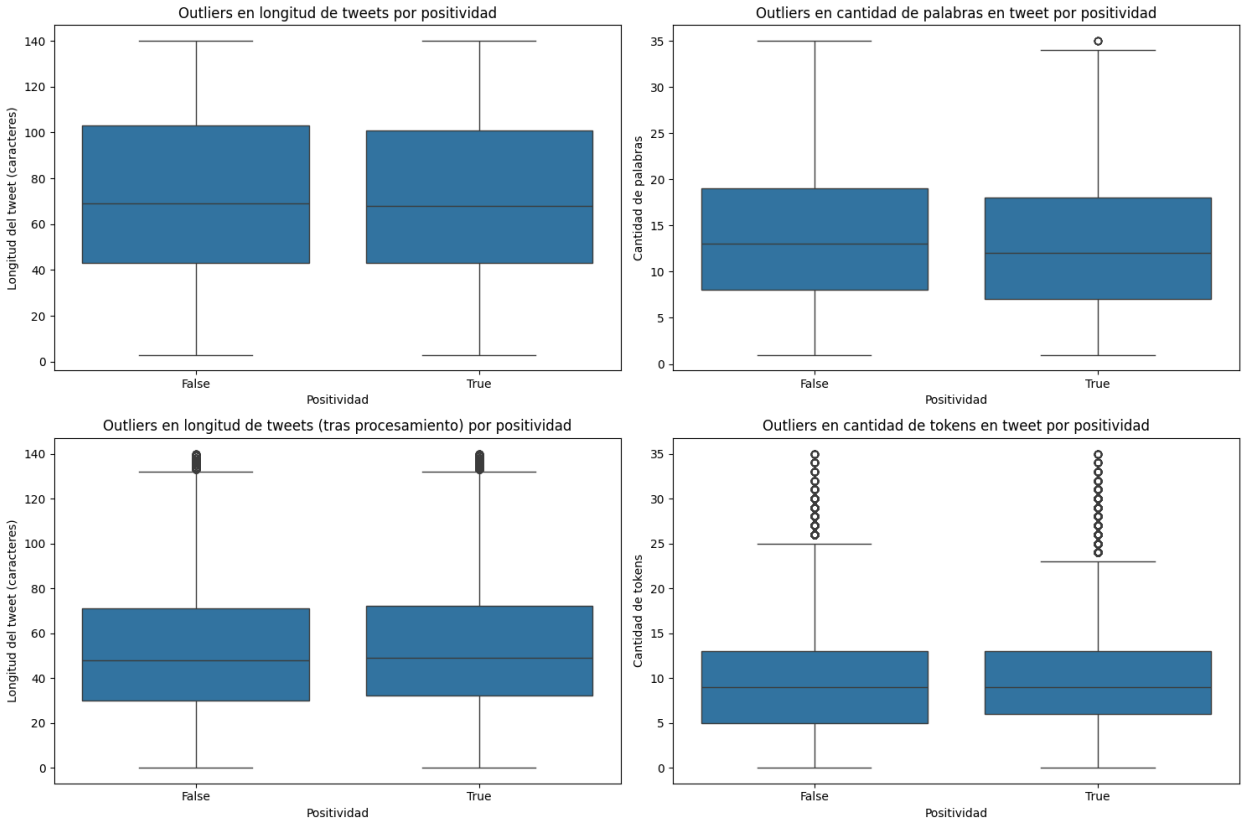
El heatmap realizado con la matriz de correlación nos hace llegar a la misma conclusión.



Sin embargo, estos nuevos datos nos sirvieron para graficarlos en un barplot, lo cuál nos dio información acerca de la distribución de longitud de los tweets.



Debido a que los tweets están limitados a 140 caracteres, no se encontraron outliers significativos.



Sin embargo, igualmente fue necesaria una limpieza de aquellos textos que no pudieron ser tokenizados exitosamente debido a que se encontraban en otro idioma. Así como aquellos textos duplicados, los cuales fueron removidos.

Quedándonos finalmente con 1598053 tweets, lo cual representa un 99.88% del dataset original, por lo que, pese a la limpieza, no hubo una pérdida significativa de datos.

Modelización

Propuesta de modelización

Debido a la naturaleza del tema a tratar, los textos de los tweets, que pueden llegar a ser bastante diferentes en cuanto a su extensión y redacción, deberán ser analizados para conocer exactamente lo que quieren comunicar. Sin embargo, debido a la gran cantidad de estos que se generan día a día en la red social, resulta complicado entender exactamente qué es lo que los usuarios quieren comunicar, y conocer la proporción de cuántos opinan así. Para muchas empresas y personas estos resultados y hallazgos representan puntos claves en sus estrategias, ya que a partir de ello pueden conocer estadísticas de desempeño de sus productos, reputación de nombre o marca, satisfacción del cliente, etc. De ahí la importancia de poder transformar toda su redacción en algo más sencillo como una simple opinión de si su percepción es positiva o no, y contabilizar ello. Para ello, el equipo de trabajo ha considerado pertinente adoptar un enfoque basado en el uso de modelos de lenguaje preentrenados y que han generado buenos resultados en otros casos previamente.

Inicialmente, se planteó el uso de modelos de lenguaje preentrenados como BERT (Bidirectional Encoder Representations from Transformers), uno de los modelos más completos entre las diferentes opciones que se pueden manejar y que se encuentran disponibles como código abierto en la web, debido a su capacidad para comprender el contexto bidireccional de las frases y su probado desempeño en tareas de análisis de sentimientos, como marco de aprendizaje automático desarrollado por Google (Lutkevich y Hashemi-Pour, s.f.). Sin embargo, debido a limitaciones de tiempo y recursos computacionales, se adoptó un enfoque clásico y eficiente basado en la vectorización de texto mediante TF-IDF (Term Frequency-Inverse Document Frequency), seguido por un clasificador Regresión Logística, técnica ampliamente probada en problemas de análisis de sentimientos. Esta combinación permite capturar la relevancia de las palabras en el contexto del corpus y generar predicciones robustas con bajos requerimientos computacionales.

El modelo podrá ser ajustado al conjunto de datos Sentiment140, luego de realizar un adecuado proceso de limpieza y binarización de la columna “target”. Como es habitual, se dividirán los datos en conjuntos de entrenamiento y evaluación. Este enfoque no solo permitiría clasificar los tweets del dataset, sino también estimar el comportamiento del modelo ante nuevos mensajes, facilitando la identificación de probabilidades de positividad o negatividad a partir del contenido textual. Y por último, la efectividad del modelo también se evaluaría mediante métricas como Accuracy, Precision y F1-Score, confirmando su capacidad para resolver la tarea planteada.

Publicación de los resultados

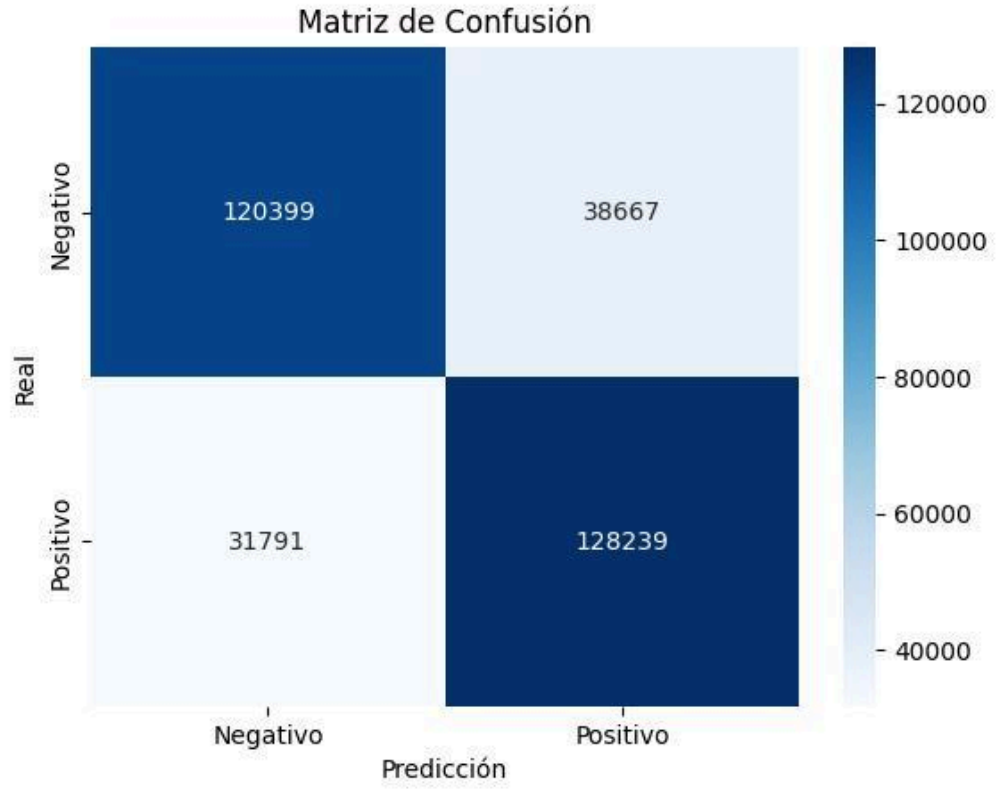
Una vez entrenado el modelo de regresión logística utilizando vectores TF-IDF, se evaluó su rendimiento sobre el conjunto de prueba compuesto por 319,096 tweets. Los resultados obtenidos son los siguientes:

```
Accuracy: 0.77919497580665538

Reporte de clasificación:
              precision    recall  f1-score   support

       False       0.79      0.76      0.77    159066
        True       0.77      0.80      0.78    160030

    accuracy                           0.78    319096
   macro avg       0.78      0.78      0.78    319096
weighted avg       0.78      0.78      0.78    319096
```

De manera global, el modelo logra una precisión cercana al 78%, lo que demuestra un desempeño aceptable considerando la variabilidad natural del lenguaje en los tweets. Además, la métrica F1-Score, que balancea precisión y recall, se mantiene consistente para ambas clases, evidenciando que el modelo no está desbalanceado hacia una sola categoría.

Analizando bien estas métricas, se observa que el modelo tiene un ligero mejor desempeño identificando tweets positivos con un recall del 80%, lo que significa que logra detectar la mayoría de los tweets positivos. En cambio, presenta un poco más de errores al identificar tweets negativos, aunque la diferencia no es significativa. Este comportamiento puede atribuirse a que, en redes sociales, los mensajes positivos suelen contener palabras más características o claras, mientras que los negativos presentan una mayor variedad lingüística o sarcasmo, que puede ser más difícil de capturar.

Además de las métricas globales, se analizó la matriz de confusión, que permite visualizar el desempeño del modelo en términos de aciertos y errores para cada clase:



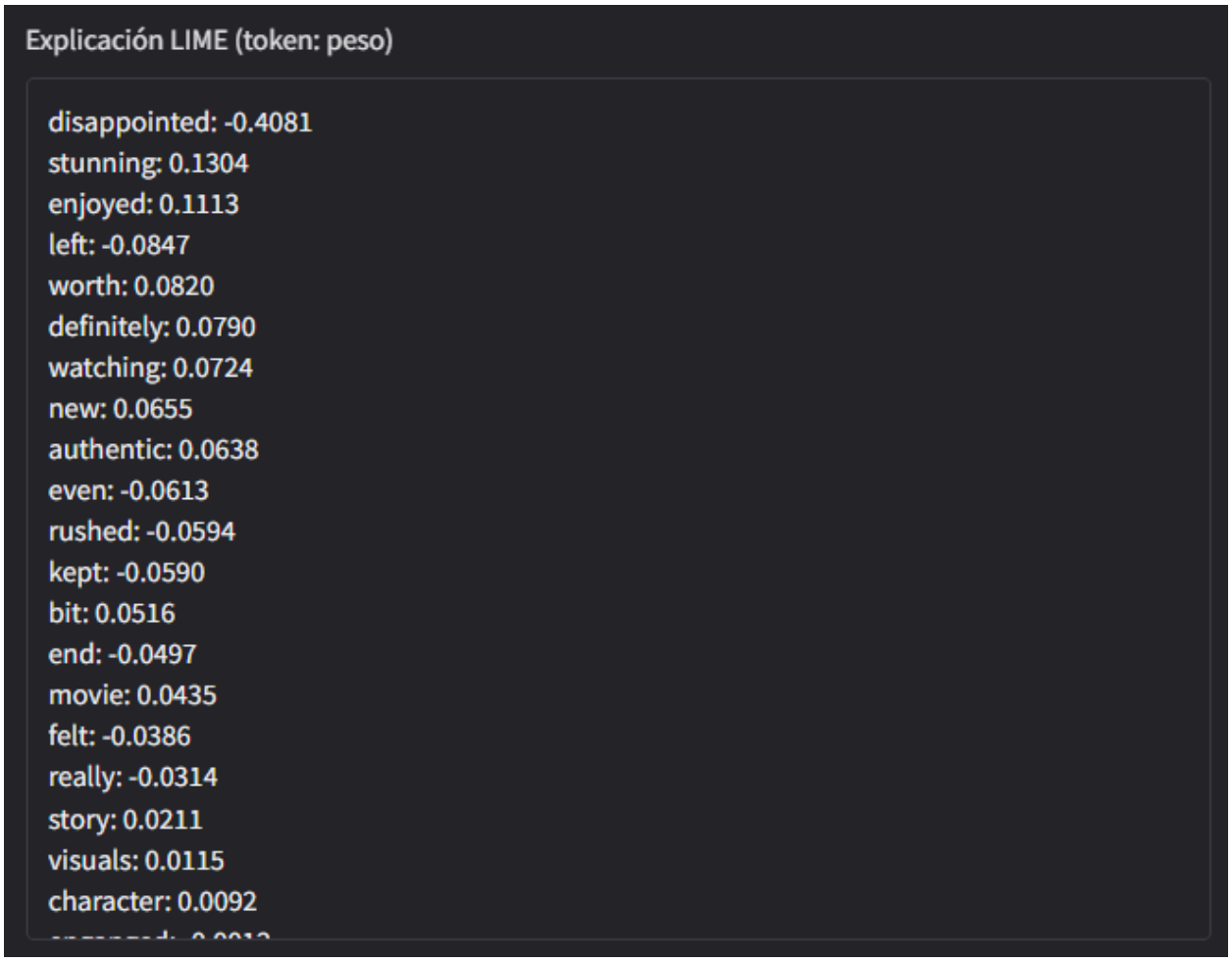
Con esto entendemos que el modelo tiende a acertar más en los tweets positivos (128,239) que en los negativos (120,399). Sin embargo, hay una cantidad considerable de falsos positivos (38,667), lo que indica que a veces interpreta tweets negativos como positivos. Además, la cantidad de falsos negativos (31,791) es un poco menor, mostrando un mejor recall para la clase positiva.

Estos resultados son coherentes con las métricas de precisión y recall calculadas anteriormente y sugieren que el modelo tiene un rendimiento equilibrado, aunque podría optimizarse más para reducir los falsos positivos.

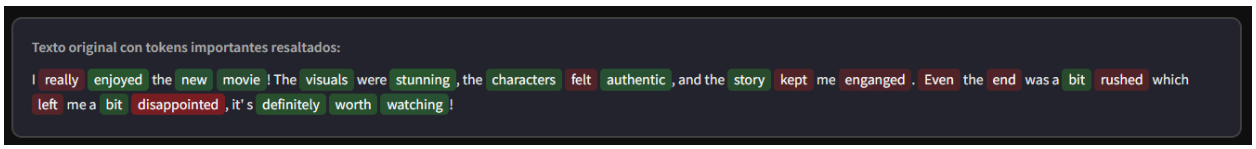
Análisis de Explicabilidad

Con el fin de interpretar el comportamiento del modelo de regresión logística entrenado sobre características TF-IDF, se utilizó LIME (Local Interpretable Model-agnostic Explanations), una técnica que permite explicar las predicciones identificando qué tokens dentro del texto contribuyen de manera positiva o negativa a la clasificación.

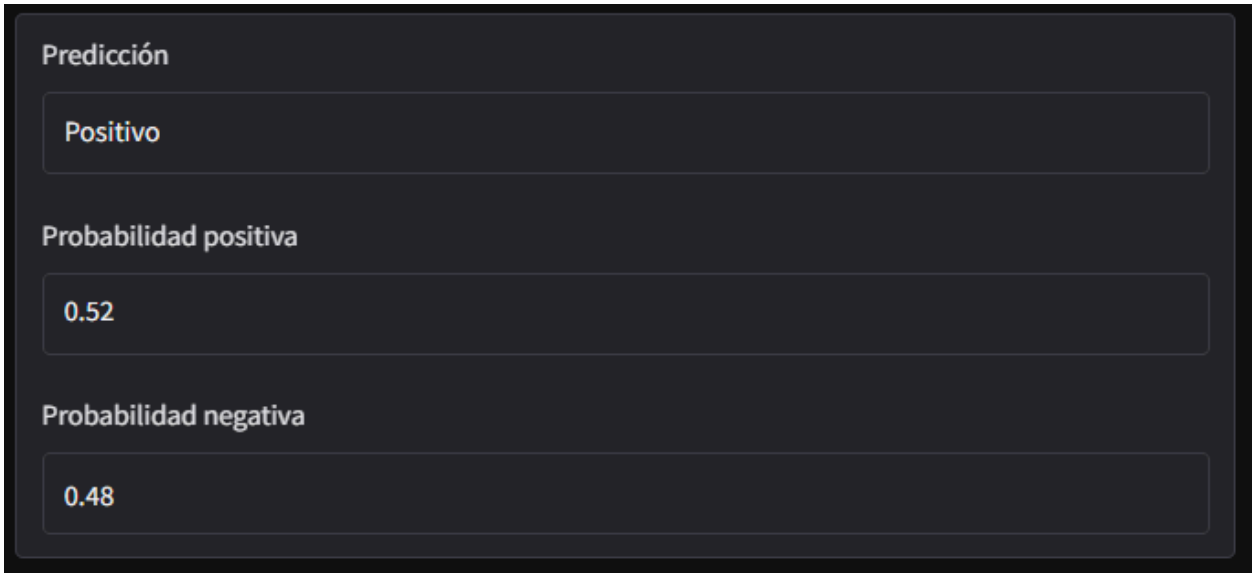
A partir del análisis, se identificaron tokens con alta influencia en las predicciones, que tan positivas o negativas son:



Adicionalmente, se implementó una visualización interactiva mediante **Gradio**, lo cual permitió resaltar los tokens más influyentes en cada tweet, facilitando así la comprensión del modelo por parte de usuarios no expertos.



La interpretación de los resultados mostró que el modelo sigue un comportamiento coherente con el lenguaje humano, atribuyendo peso positivo a palabras asociadas con emociones agradables y peso negativo a términos relacionados con el desagrado o la crítica.



Conclusiones

- En el presente trabajo se implementó un modelo de análisis de sentimientos aplicado a tweets del conjunto de datos Sentiment140, logrando clasificar los mensajes como positivos o negativos. El enfoque basado en la vectorización mediante TF-IDF y clasificación con regresión logística permitió obtener un rendimiento global satisfactorio, alcanzando una precisión cercana al 78%, con métricas equilibradas entre ambas clases.
- El análisis de explicabilidad realizado con LIME permitió identificar los tokens más influyentes en las decisiones del modelo, mostrando coherencia con el contexto semántico habitual de las redes sociales. Palabras con carga emocional positiva como “love” y “enjoyed” contribuyeron a predicciones positivas, mientras que términos como “hate” o “disappointed” inclinaron las predicciones hacia lo negativo.
- Aunque el modelo presentó buenos resultados iniciales, se identificaron oportunidades de mejora, especialmente en la detección de mensajes negativos, donde la presencia de sarcasmo o ambigüedad dificulta la clasificación precisa.
- Para una próxima oportunidad, ya sería oportuno explorar el entrenamiento de modelos de lenguaje más avanzados, como BERT, que permiten captar mejor el contexto y matices del lenguaje coloquial, así como ajustar hiperparámetros y evaluar nuevas técnicas de preprocesamiento que reduzcan la ambigüedad en los datos.

Referencias bibliográficas

Center for Technology and Society. (2023, 24 de mayo). *Threads of Hate: How Twitter's Content Moderation Misses the Mark.* ADL. https://www.adl.org/resources/article/threads-hate-how-twitters-content-moderation-misses-mark

Dreißigacker, A., Müller, P., Isenhardt, A., & Schemmel, J. (2024). *Online hate speech victimization: consequences for victims' feelings of insecurity.* Crime Science, 13(1). https://doi.org/10.1186/s40163-024-00204-y

Elmitwalli, S., & Mehegan, J. (2024). *Sentiment analysis of COP9-related tweets: a comparative study of pre-trained models and traditional techniques.* Frontiers In Big Data, 7. https://doi.org/10.3389/fdata.2024.1357926

Go, A., Bhayani, R., Huang, L. (2009). *Twitter Sentiment Classification using Distant Supervision.* Stanford University. https://www-cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf

Imran, A. S., Daudpota, S. M., Kastrati, Z., & Batra, R. (2020). *Cross-Cultural Polarity and Emotion Detection Using Sentiment Analysis and Deep Learning on COVID-19 Related Tweets.* IEEE Access, 8, 181074-181090. https://doi.org/10.1109/access.2020.3027350

Kaggle. (2017, 13 de septiembre). *Sentiment140 dataset with 1.6 million tweets.* Kaggle. https://www.kaggle.com/datasets/kazanova/sentiment140

Lohiya, H. (2023, 30 de agosto). *Sentiment analysis with Sentiment140 - Himanshu Lohiya - Medium.* Medium. https://himanshulohiya.medium.com/sentiment-analysis-with-sentiment140-e6b0c789e0ce

Lutkevich, B. y Hashemi-Pour, C. (s.f.). *BERT language model*. TechTarget. https://www.techtarget.com/searchenterpriseai/definition/BERT-language-model

Pujari, A., & Malik, P. (2024, 24 de diciembre). *NLP in Social Media: Impact and Use Cases.* https://www.sprinklr.com/blog/nlp-in-social-media/

Saha, K., Chandrasekharan, E., & De Choudhury, M. (2019). *Prevalence and Psychological Effects of Hateful Speech in Online College Communities.* Proceedings of the ... ACM Web Science Conference. ACM Web Science Conference, 2019, 255–264. https://doi.org/10.1145/3292522.3326032

Strait, E. (2023, 22 de mayo). *Leveraging NLP Techniques for Effective Content Moderation.* Lettria. https://www.lettria.com/blogpost/nlp-techniques-for-content-moderation

TensorFlow. (s. f.). *Sentiment140*. TensorFlow. https://www.tensorflow.org/datasets/catalog/sentiment140