UNIVERSIDAD DE CONCEPCIÓN

FACULTAD DE INGENIERÍA

DEPARTAMENTO DE INGENIERÍA INFORMÁTICA Y CIENCIAS DE LA COMPUTACIÓN

Análisis De Datos

Proyecto semestral*:*
ChatSentinel

Javier Cadagán Parra
Gabriel Huerta Torres
Diego Oyarzo Navia

Fecha: 1 Junio, 2025

Descripción del enfoque analítico/modelo elegido y justificación técnica de la elección.

El enfoque analítico adoptado para este proyecto es la clasificación con aprendizaje supervisado ya que objetivo principal es construir un modelo capaz de distinguir entre dos o más categorías predefinidas ('grooming' y 'no grooming') basándose en las características extraídas de los chats de texto, que son nuestros datos de entrada.

Después, el modelo específico elegido es la regresión logística (logistic regression). Se seleccionó este modelo al ser útil para predecir la probabilidad de un resultado binario, ajustándose bien a problemas de clasificación con dos clases como lo tenemos en este caso. Por otro lado es relativamente rápido de entrenar y eficiente con conjuntos de datos de tamaño considerable, proporcionando un rendimiento sólido y también sirviendo como un buen punto de partida para problemas de clasificación.

En cuanto al proceso analítico para la implementación y evaluación del modelo, el primer paso consistió en trabajar con base de datos original que contenía únicamente conversaciones etiquetadas como grooming. Estas conversaciones estaban organizadas por archivos, por lo que fue necesario agrupar los mensajes por conversación mediante la función groupby, concatenar sus textos y asignarles un identificador único. A cada conversación se le asignó la etiqueta ‘1’ para saber que pertenecen a la clase de grooming. Sin embargo, para la clasificación de los chats se requería también una nueva base de datos que contuviera chats que estén etiquetados como ‘no grooming’.

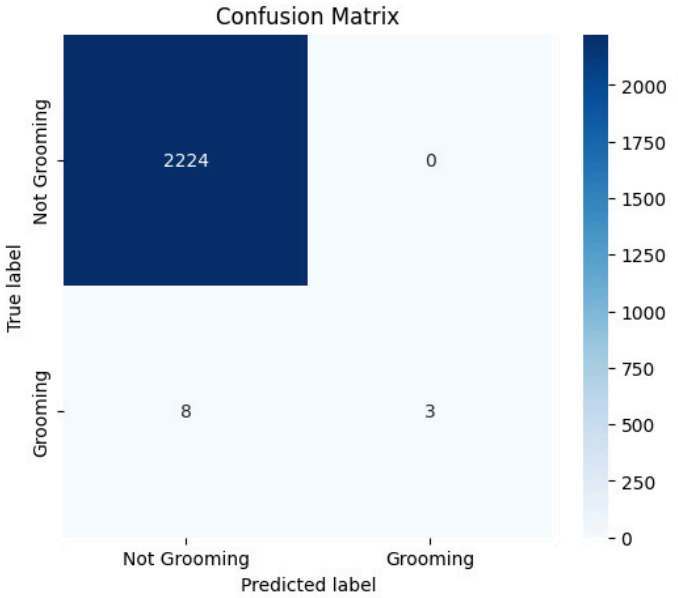
Después, se incorporó una segunda base de datos del conjunto de datos “DailyDialog”, disponible en la plataforma “Hugging Face”. Contiene diálogos cotidianos que no tienen que ver con grooming, por lo que se utilizó como representación de la clase de no grooming. Cada uno de los diálogos fueron convertidos en una única cadena de texto, y se eliminaron columnas irrelevantes, posteriormente se realizó un análisis de las longitudes de los chats en ambos conjuntos con el fin de identificar posibles desequilibrios extremos, eliminando aquellos que fueran outliers con respecto a la longitud de caracteres por chat con respecto al resto de mensajes, ambos conjuntos fueron unificados en un único dataframe asegurando que las columnas y formatos fueran consistentes.

Consecutivamente, una vez consolidado el conjunto de datos, se procedió a la ingeniería de características. Se utilizó una vectorización TF IDF (Term Frequency Inverse Document Frequency) para convertir los textos en representaciones numéricas adecuadas para el entrenamiento del modelo. Este proceso consideró un máximo de 10.000 características, la eliminación de palabras vacías utilizando stopwords y la aplicación del ajuste sublineal, donde “sublinear_tf=True” para moderar la influencia de términos frecuentes.

Posteriormente, el conjunto de datos fue dividido en un conjunto de entrenamiento y otro de prueba, utilizando una proporción 80/20. Por último se utilizó la implementación de Logistic Regression de la librería scikit-learn, configurada con un número máximo de iteraciones igual a 1000.

Resultados preliminares con métricas básicas de desempeño.

Finalmente, se evaluó el desempeño del modelo sobre el conjunto de prueba utilizando métricas como lo son la matriz de confusión como también valores de precisión, recall, F1-score y soporte para cada clase, obteniendo los siguientes resultados:



Classification Report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	2224
1	1.00	0.27	0.43	11
accuracy			1.00	2235
macro avg	1.00	0.64	0.71	2235
weighted avg	1.00	1.00	1.00	2235

Reflexión inicial sobre fortalezas y debilidades del modelo.

Los resultados obtenidos tras evaluar el modelo muestran una alta precisión general, con una “accuracy” de 1.00. Existe un desbalance en las clases del conjunto de prueba, donde 2224 ejemplos pertenecen a la clase no grooming (0), mientras que solo 11 pertenecen a la clase grooming (1). Esta proporción influye en las métricas asociadas a la clase minoritaria.

En particular, el modelo alcanza una “precision” de 1.00 para ambas clases, lo cual indica que los ejemplos etiquetados como grooming son efectivamente positivos. Sin embargo, el recall para la clase grooming es solo de 0.27, lo que significa que el modelo detecta correctamente solo el 27% de los casos reales de grooming. Esto también se refleja en el bajo valor del “F1-score” (0.43) para esta clase, lo que indica un rendimiento bajo en cuanto al equilibrio entre precisión y recall.

El análisis de estas métricas sugiere que, aunque el modelo es muy eficaz para reconocer la clase no grooming, tiene dificultades para identificar correctamente los casos de grooming. Este comportamiento es debido a que las clases aún están desbalanceadas, y resalta la necesidad de aplicar algún re-muestreo con otra base de datos para alguno de los 2 casos, un ajuste de umbrales de decisión para aumentar la sensibilidad y así el recall, o bien, explorar modelos más robustos ante desequilibrios en los datos.