🧠 Informe de Evaluación del Modelo de Moderación de Chat IA

---

1. Introducción

Este documento detalla el proceso de evaluación y el rendimiento de nuestro modelo de Inteligencia Artificial para la moderación de chat. El objetivo principal de este modelo es clasificar mensajes de texto en categorías de toxicidad para fomentar entornos online más seguros y saludables.

---

2. Metodología de Evaluación

El modelo, basado en la arquitectura **DistilBERT**, fue evaluado en un conjunto de datos de prueba independiente que no fue utilizado durante el entrenamiento. Se utilizaron métricas estándar de clasificación para evaluar su desempeño, incluyendo:

- **Precisión (Accuracy):** Proporción de predicciones correctas sobre el total de predicciones.
- **Precisión por Clase (Precisión):** Proporción de verdaderos positivos sobre todos los resultados positivos predichos por el modelo.
- **Exhaustividad por Clase (Recall):** Proporción de verdaderos positivos sobre todos los casos reales positivos.
- **Puntuación F1 por Clase (F1-score):** Media armónica entre precisión y exhaustividad.
- **Soporte (Support):** Número de instancias reales de cada clase.
- **Pérdida (Loss):** Medida de la discrepancia entre predicciones y etiquetas verdaderas.

---

3. Resultados de la Evaluación

Métricas Generales:

Métrica	Valor (aproximado)
Precisión General (Accuracy)	91.74%
Pérdida (Loss)	0.602
F1-score Ponderado (Weighted Avg)	91.75%
Precisión Ponderada (Weighted Avg)	91.83%
Recall Ponderado (Weighted Avg)	91.74%

Desempeño por Clase:

Clase	Precisión	Recall	F1-score	Soporte
Acción/Juego	0.739	0.844	0.788	450
Gravemente Tóxico	0.873	0.886	0.879	1199
Levemente Tóxico	0.790	0.718	0.753	572
No Tóxico	0.956	0.951	0.954	5469

Promedios Generales:

Promedio	Precisión	Recall	F1-score	Soporte
Macro Avg	0.840	0.850	0.844	7690
Weighted Avg	0.918	0.917	0.918	7690

---

4. Análisis y Conclusiones

Los resultados demuestran que el modelo de moderación de chat es **altamente eficaz y fiable** para la tarea de clasificación de toxicidad.

- ✅ **Excelencia en Clases Críticas:**
  El modelo exhibe un desempeño sobresaliente en la identificación de mensajes **No Tóxicos** y, especialmente, de **Gravemente Tóxicos**. Su F1-score de ~0.879 y alto recall reducen el riesgo de que mensajes peligrosos pasen desapercibidos.

- 🧩 **Manejo de la Ambigüedad:**
  Las clases con subjetividad (como **Levemente Tóxico**) tienen un rendimiento más modesto (F1-score de ~0.753). Aunque aceptable, es una oportunidad clara para seguir entrenando el modelo con más ejemplos. La clase **Acción/Juego** también

muestra precisión menor, lo que indica posibles confusiones con otras categorías.

- 💪 **Robustez General:**
  Una precisión global del **91.74%**, junto con métricas ponderadas altas, confirma la capacidad del modelo para generalizar bien en datos no vistos y ofrecer resultados confiables.

---

5. Próximos Pasos y Mejoras Futuras

Para seguir mejorando el sistema de moderación, se proponen las siguientes acciones:

- 📈 **Expansión del Dataset:**
  Aumentar y diversificar el volumen de datos etiquetados, especialmente en categorías ambiguas como “Levemente Tóxico”.

- 🔍 **Análisis de Errores:**
  Estudiar los errores más frecuentes (falsos positivos/negativos) para identificar patrones y ajustar el entrenamiento.

- 🧠 **Feedback Humano Integrado:**
  Implementar retroalimentación por parte de moderadores humanos, incorporando esas correcciones para reentrenamiento periódico.

- 🏷️ **Clasificación Multi-Etiqueta:**
  Explorar modelos que permitan que un mismo mensaje tenga múltiples etiquetas si es necesario.