Proyecto MML

Clasificación Automática de Discursos de Odio

GRUPO 1

** Esta presentación contiene ejemplos con lenguaje potencialmente ofensivo, usados solo con fines académicos para ilustrar un modelo de detección de discurso de odio.

Agenda

DEFINICIÓN DEL PROBLEMA

Definir claramente el problema o pregunta de negocio

DISEÑO DE LA SOLUCIÓN

Modelos de lenguaje para atender este problema (modelos de embedding, modelos entrenados para tareas específicas

CONJUNTO DE DATOS

Identificar el conjunto de datos a emplear.

EVALUACIÓN DE LA SOLUCIÓN

Respuesta a preguntas de la A a la F

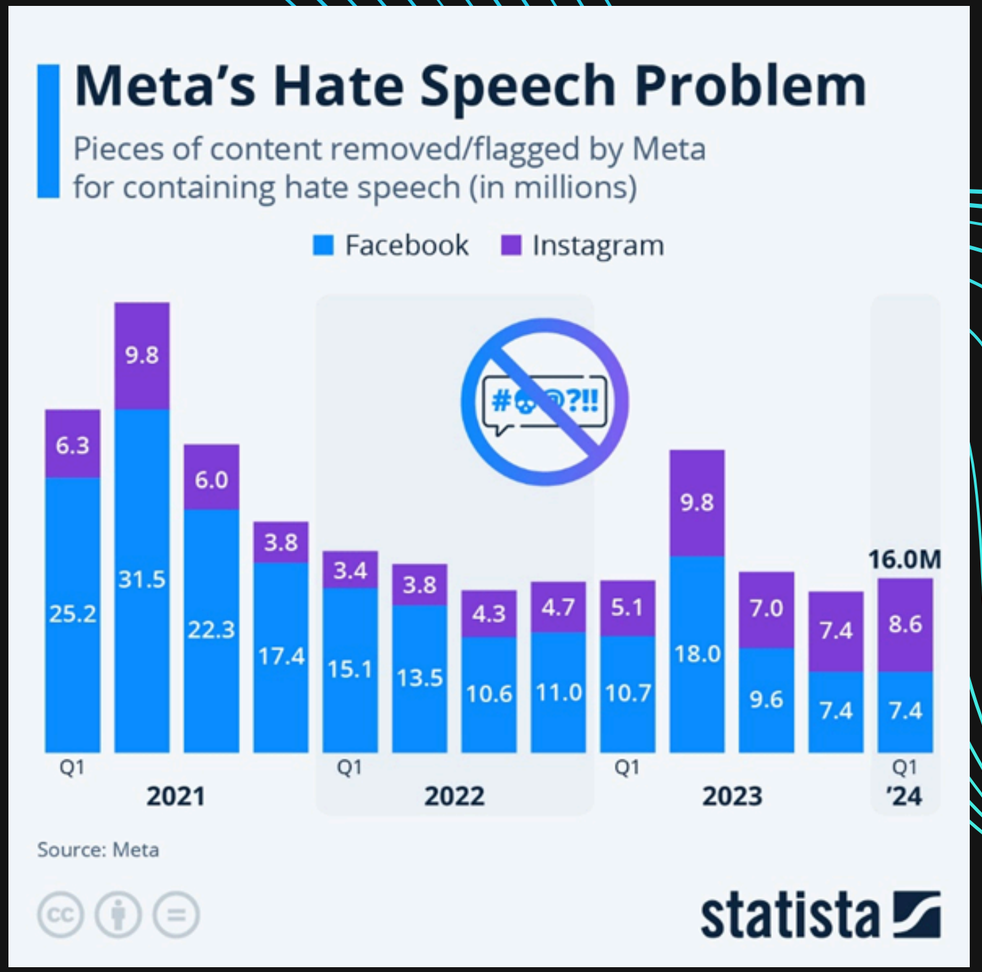
RESULTADOS DE LA SOLUCIÓN

Explorar los resultados de la soluci´on, su alcance y limitaciones.

Definición del Problema

Las redes sociales han facilitado la difusión global de información, pero también han impulsado la propagación del discurso de odio.

- Casos como el asalto al Capitolio (2021) y los ataques en Brasil (2023) evidencian su impacto real, al haber estado precedidos por actividad en redes sociales.
- En plataformas como X:
  - El discurso de odio ha aumentado un 50%.
  - Comentarios transfóbicos, homofóbicos y racistas han subido un 260%, 30% y 42%, respectivamente.
- La moderación manual no escala ante el volumen de contenido generado diariamente.
- El lenguaje ofensivo es ambiguo, irónico o codificado, dificultando su detección automática.



Meta´s Hate Speech Problem. Tomado de: (Zandt, 2024)

Solución

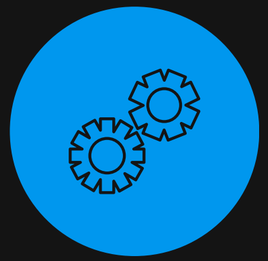
- Utilizamos **redes neuronales** para identificar automáticamente si un mensaje es tóxico.
- Cada tweet es transformado en una representación numérica usando **embeddings contextuales**, que capturan el significado de las palabras en su contexto.
- Esto permite al sistema detectar formas sutiles o codificadas de discurso de odio.
- El resultado es un sistema automático, escalable y adaptable.

Etapas de la Solución

Recolección y Preparación de Datos

Extracción de Características

Modelado, Entrenamiento y Finetunning



Aplicación

Conjunto de Datos

DAVIDSON ET AL.'S HATE SPEECH AND OFFENSIVE LANGUAGE DATASET:

- Tweets etiquetados como hate speech, offensive language o neither.
- Origen: https://github.com/t-davidson/hate-speech-and-offensive-language
- Ejemplo:

@user "Shut up nigga, go away" → hate speech

HATE SPEECH AND OFFENSIVE CONTENT IDENTIFICATION IN INDO-EUROPEAN LANGUAGESDATASET:

- Competencia internacional que incluye miles de publicaciones en inglés (y otros idiomas) clasificadas en hate/offensive vs non-offensive.
- Origen: https://hasocfire.github.io/hasoc/2019/dataset.html
- Ejemplo:

"You are such a retard" → offensive

ZEERAK TALAT'S HATE SPEECH DATASET:

- IDs de tuits anotados para detallar discursos de odio (hate vs non-hate)
- Origen: https://github.com/zeeraktalat/hatespeech
- Ejemplo (reconstruido del ID):

"All Muslims are terrorists" → hate

Roboto

RECOLECCIÓN & PREPROCESAMIENTO DE DATOS

- **Concatenación y Homogeneización**
  - Se unen los datasets (Davidson, HASOC, Talat) unificando etiquetas a 0 = non-toxic y 1 = toxic. Al finalizar este proceso se termino con ~45000 tweets
- **Limpieza del Texto**
  - Conversión de emojis a palabras.
  - Eliminación de URLs, menciones, símbolos innecesarios.
  - Corrección ortográfica y remoción de duplicados y elementos nulos.
- **Balanceo con EDA**
  - Se aplica Easy Data Augmentation: sinónimos, inserción/ intercambio/ eliminación aleatoria de palabras con alpha = 0.25 para equilibrar clases minoritarias

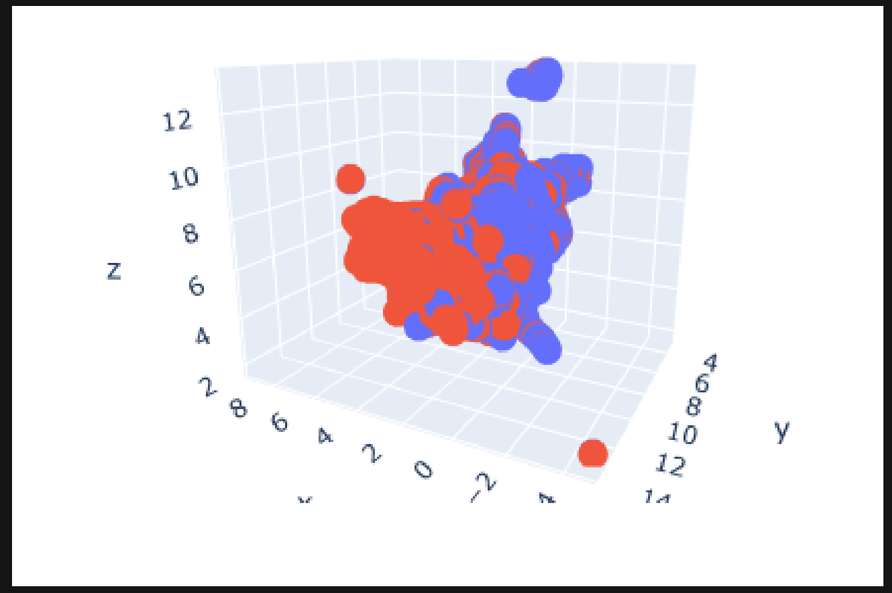
MODELOS PROPUESTOS

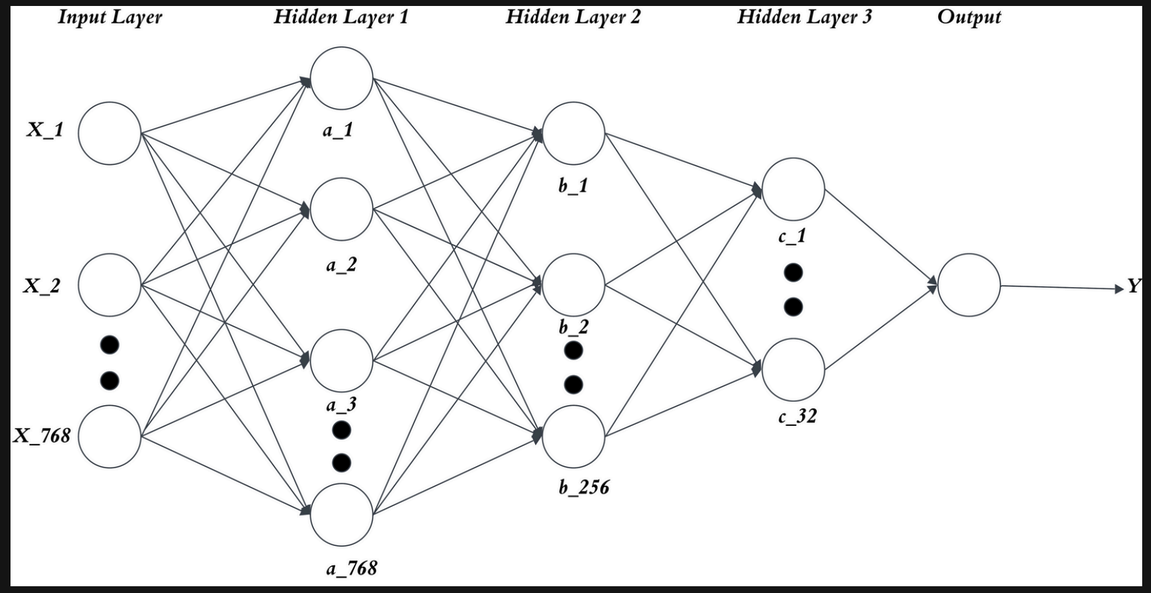
1. RoBERTa embeddings + MLP (Baseline)

- Se genera la representación de los tweets preprocesados usando un modelo **RoBERTa-base** entrenado sobre aproximadamente 58 millones de tuits, a esta se le junto una MLP con ReLU para clasificar los tweets la cual fue entrenada sobre los datos usando **Binary Cross Entropy** y **ADAM.**
- Mejor modelo con **4 capas** de **32 neuronas** y **100 épocas**

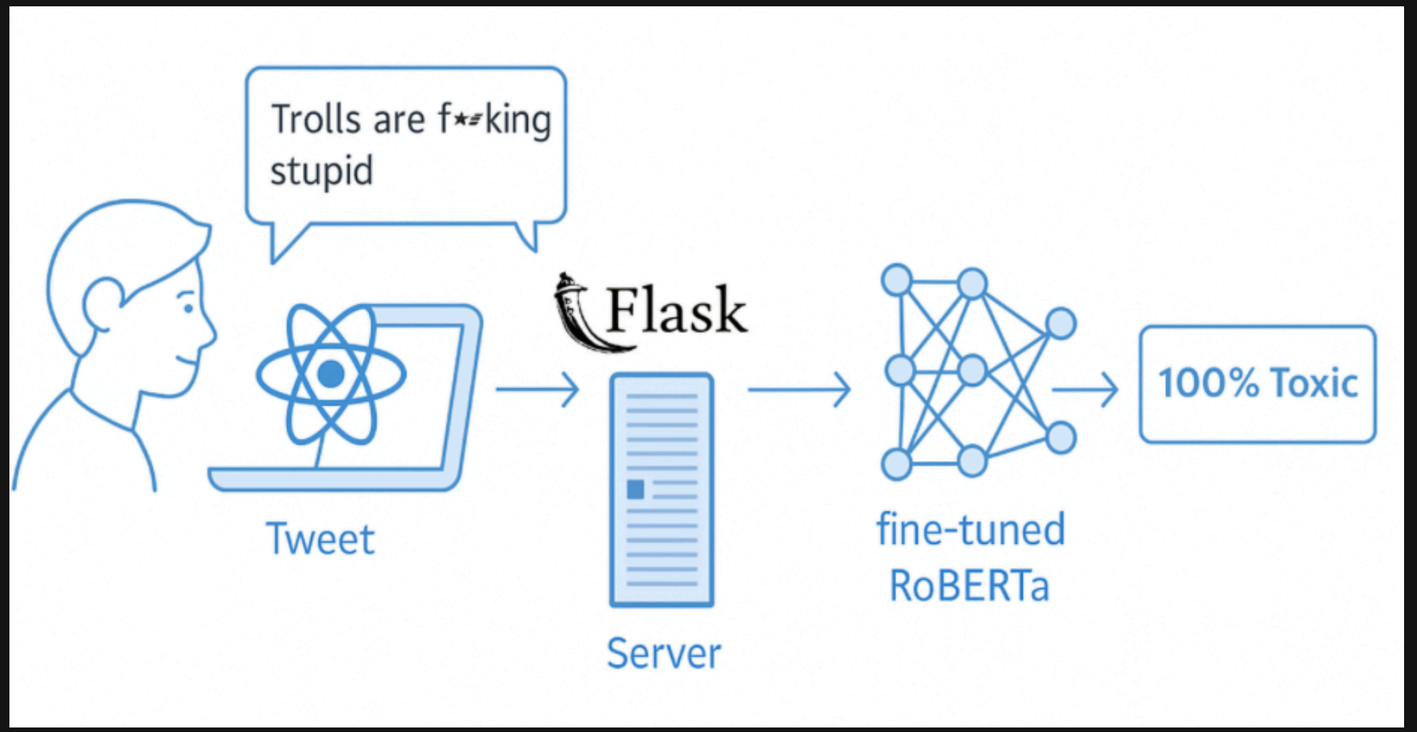
2. Fine-tuning de RoBERTa con cabeza clasificadora

- Se hace Fine-tuning de **RoBERTa** haciendo uso de una cabeza clasificadora, con **dropout** y **ReLU.** se utiliza **Binary Cross Entropy** + **Sigmoid, ADAM** y **100 épocas**





Aplicación


Trolls are f*ing stupid
Flask
100% Toxic
Tweet
Server
fine-tuned RoBERTa

Evaluación de la Solución

ASPECTOS A EVALUAR

- Precisión en la clasificación de tweets como tóxicos o no tóxicos.
- Robustez ante ambigüedad lingüística (sarcasmo, ironía, eufemismos).
- Balance entre detección y sobre-identificación (evitar falsos positivos).
- Interfaz funcional y amigable con el usuario.

PRUEBAS A REALIZAR

- Evaluación sobre conjunto de prueba separado y representativo.
- Pruebas con ejemplos ambiguos o adversariales (creados manualmente).
- Testeo en tiempo real desde la interfaz web, con casos reales y controlados.

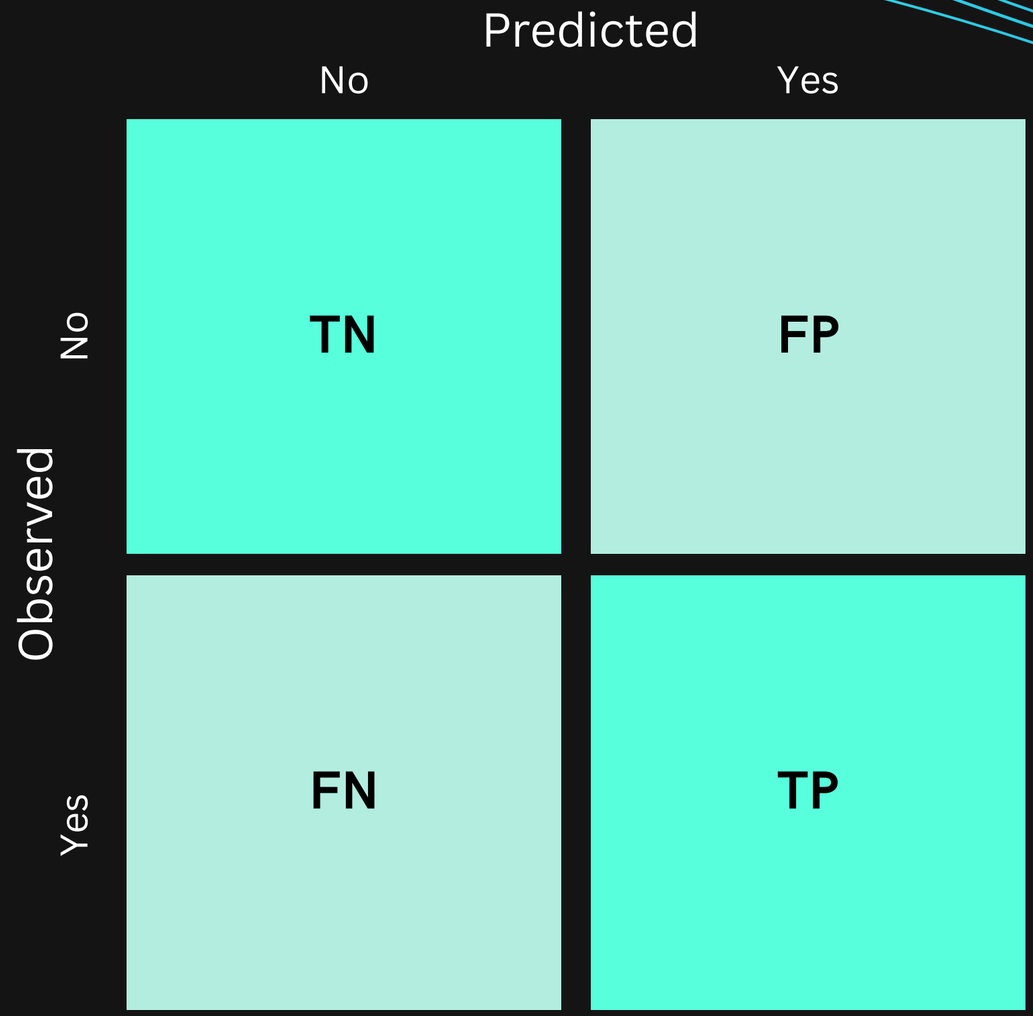
Evaluación de la Solución

MÉTRICAS A EVALUAR

- Accuracy: proporción de predicciones correctas.
- Precision: proporción de verdaderos positivos entre los detectados como tóxicos.
- Recall (sensibilidad): proporción de verdaderos tóxicos correctamente detectados.
- F1-Score: balance entre precision y recall.
- Matriz de confusión: permite visualizar claramente los verdaderos y falsos positivos y negativos.

¿POR QUÉ SELECCIONA ESTAS MÉTRICAS?

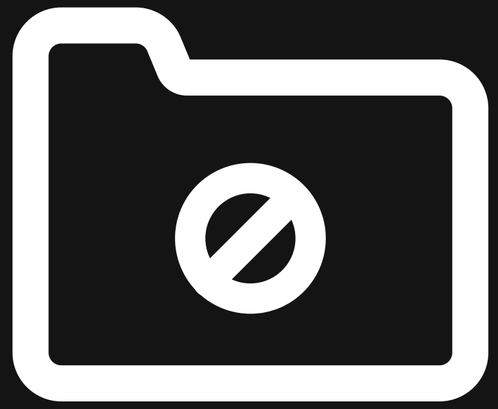
- Estas métricas permiten evaluar de manera integral la calidad del modelo, considerando tanto exactitud general como comportamiento en casos límite.
- El dataset balanceado justifica el uso de accuracy, que de otro modo podría ser engañosa.
- F1-Score da una medida robusta del rendimiento cuando hay riesgo de comprometer precisión o recall.
- Incluir ejemplos ambiguos permite validar si el modelo entiende el contexto más allá de palabras aisladas.



Evaluación de la Solución

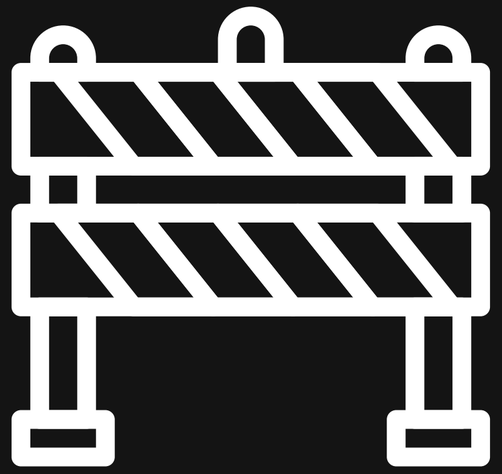
¿QUÉ LIMITACIONES TIENEN ESTAS PRUEBAS Y MÉTRICAS?

- El análisis de ejemplos ambiguos es manual y poco escalable.
- No se ha probado el modelo con tráfico real de usuarios.
- El conjunto de prueba puede no reflejar todos los contextos y variaciones lingüísticas.



¿CÓMO ESTAS PRUEBAS LE PERMITEN TOMAR DECISIONES SOBRE SU SOLUCIÓN?

- Las métricas permiten seleccionar el modelo con mejor balance entre precisión y recall.
- Los resultados sobre ejemplos ambiguos indican si se requiere más entrenamiento o refuerzo de contexto.
- Las pruebas permiten definir si el sistema es suficientemente confiable para su uso en producción o necesita ajustes adicionales

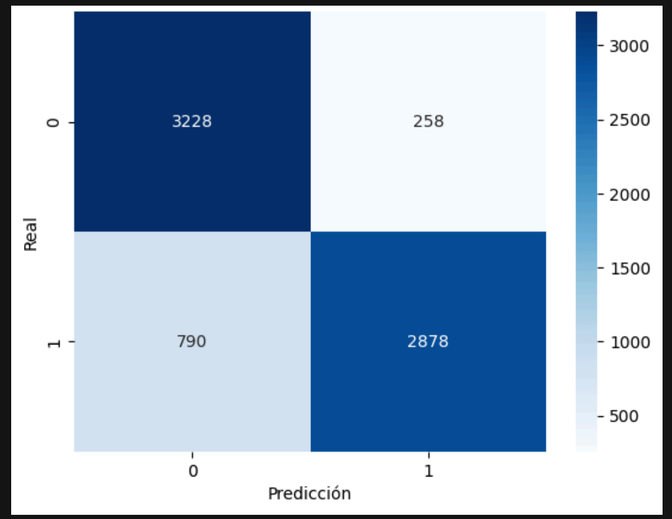


Resultados de la solución

MLP

Accuracy: 0.85

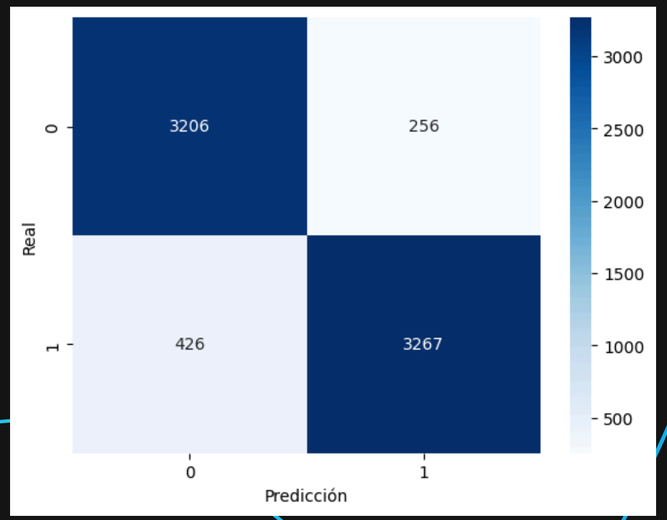
Categoría	Precision	Recall	F1
Not Hate Speech	0.80	0.93	0.86
Hate Speech	0.92	0.78	0.85



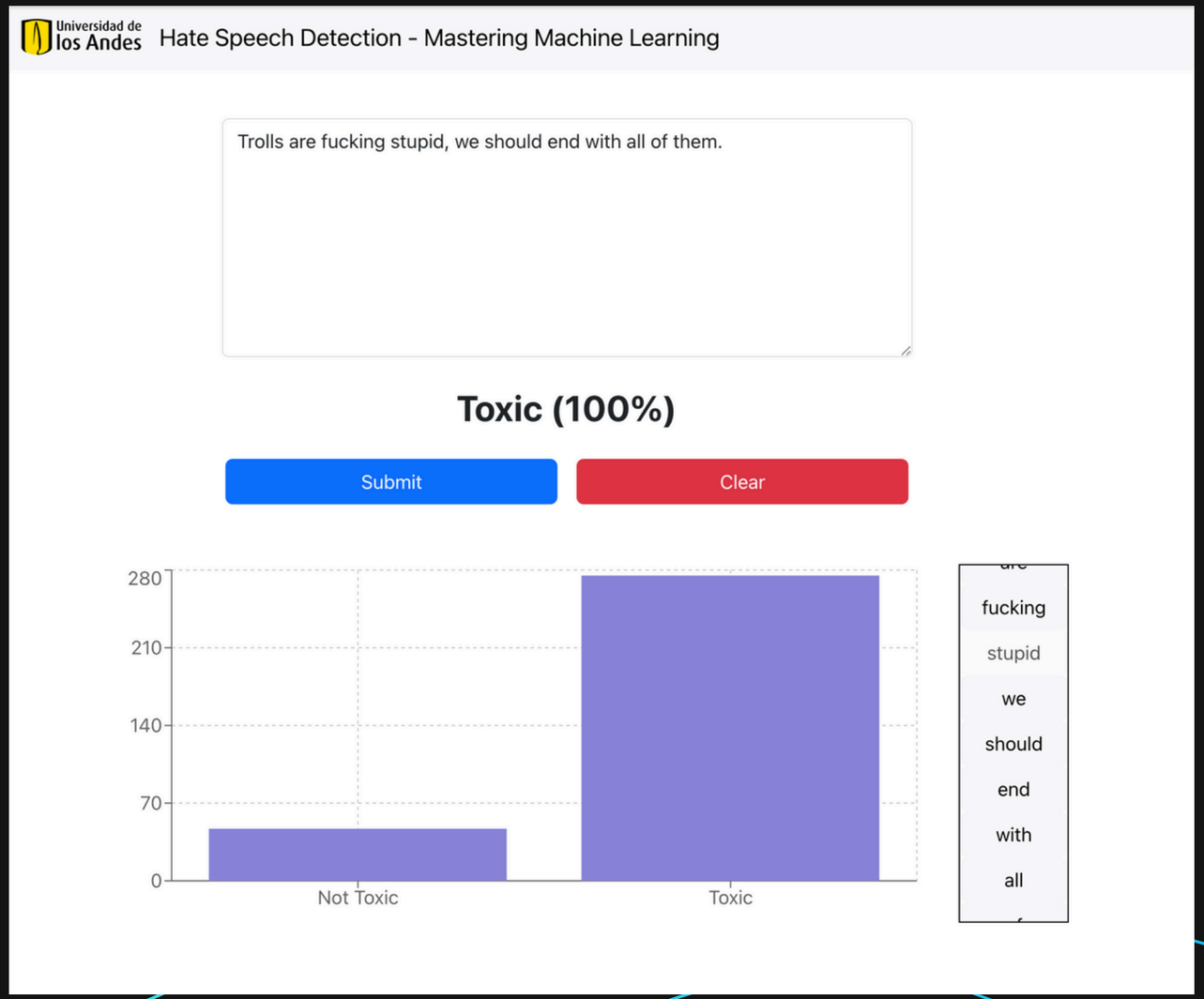
ROBERTA FINETUNE

Accuracy: 0.90

Categoría	Precision	Recall	F1
Not Hate Speech	0.88	0.93	0.90
Hate Speech	0.93	0.88	0.91



Prototipo


Universidad de los Andes
Hate Speech Detection - Mastering Machine Learning
Trolls are fucking stupid, we should end with all of them.
Toxic (100%)
Submit
Clear
280
210
140
70
0
Not Toxic
Toxic
fucking
stupid
we
should
end
with
all

SI HICIERA EL PROYECTO NUEVAMENTE, ¿QUÉ HABRÍA HECHO DIFERENTE?

Gracias por su Atención

NICOLAS BEDOYA

SANTIAGO ARENAS

SANTIAGO REYES

RAFAEL TEJÓN

Bibliografia:

- B. Fieri and D. Suhartono, “Offensive Language Detection Using Soft Voting Ensemble Model,” Mendel, vol. 29, no. 1, 2023, doi: 10.13164/mendel.2023.1.001.
- A. Bonetti, M. Martínez-Sober, J. C. Torres, J. M. Vega, S. Pellerin, and J. Vila-Francés, “Comparison between Machine Learning and Deep Learning Approaches for the Detection of Toxic Comments on Social Networks,” 2023. doi: 10.3390/app13106038.
- A. Toktarova et al., “Hate Speech Detection in Social Networks using Machine Learning and Deep Learning Methods,” International Journal of Advanced Computer Science and Applications, vol. 14, no. 5, 2023, doi: 10.14569/IJACSA.2023.0140542.
- Amina Saleh Omar, K. O. (2024). The Impact of Social Media on Society: A Systematic Literature Review. The International Journal of Engineering and Science (IJES), 96-106.
- Cohen, J. (12 de February de 2025). A Platform Problem: Hate Speech and Bots Still Thriving on X. Obtenido de USC Viterbi School of Engineering: https://viterbischool.usc.edu/news/2025/02/a-platform-problem-hate-speech-and-bots-still-thriving-on-x/
- Geuens, R. (19 de Enero de 2025). What are the top social media platforms in 2024? Obtenido de SOAX: https://soax.com/research/top-social-media-platforms
- United Nations. (28 de Enero de 2023). Hate speech: A growing, international threat. Obtenido de UN News: https://news.un.org/en/story/2023/01/1132597
- Zandt, F. (18 de Junio de 2024). Meta's Hate Speech Problem. Obtenido de Statista: https://www.statista.com/chart/21704/hate-speech-content-removed-by-facebook/

SUSTENTO DETALLADO

Sustento 1

En el año del 2024, este crecimiento continuó. Facebook lideró el número de usuarios activos mensuales, llegando a acumular más de 3 mil millones de usuarios, algo que ninguna otra red social ha logrado hasta la fecha. YouTube, por otro lado, tiene 2 mil 500 millones de usuarios activos, mientras que Instagram y WhatsApp se ubican en el tercer puesto con 2 mil millones de usuarios. Lo anterior también genera que las redes sociales tengan una alta cantidad de ingresos, por ejemplo, Facebook genera más de 80 mil millones de dólares anualmente. De lo anterior es posible observar la relevancia que las plataformas de redes sociales poseen, estas tienen un gran número de usuarios, lo que a su vez contribuye a una alta generación de ingresos. (Geuens, 2025)

Sustento 2

Así, lo que se buscará con este proyecto es desarrollar modelos de Machine Learning y Deep Learning para clasificar de forma efectiva comentarios en inglés, puesto que este es uno de los más usados, por ejemplo, para X, el 55% de tuits publicados en este lenguaje (Semiocast, 2024).

Sustento 3

Determinar lo siguiente en la clasificación de mensajes inapropiados y comparar contra los estudios de referencia resaltados (Al hacer uso de diferentes modelos).
Accurancy: Su valor nominal debe estar entre: 92.5% <= x <= 97%
Precision: Su valor nominal debe estar entre: 85% <= x <= 97%
Recall: Su valor nominal debe estar entre: 92.5% <= x <= 90%
F1-score: Su valor nominal debe estar entre: 88% <= x <= 93%

Reducción de falsos positivos y falsos negativos a menos del 5%.

Mejora en el manejo de mensajes complejos (sarcasmo, ironía y lenguaje ambiguo).

Sustento 4

Marco de Referencia para realizar las comparaciones pertinentes

Table 2: EDA Dataset – Machine Learning Perfor-mance.				
Model	**Accuracy**	**Precision**	**Recall**	**F1**
Random Forest (RF)	95.518	89.610	96.890	92.685
Decision Tree (DT)	95.357	89.385	96.545	92.414
Logistic Regression (LR)	94.422	87.696	95.949	91.050
AdaBoost (AB)	93.572	86.355	94.943	89.773
Na¨ıve Bayes (NB)	92.808	85.456	92.355	88.32
k-NN	89.494	80.515	90.566	84.030
Soft Voting – Top 3 (RF, DT, LR)	95.433	89.445	96.850	92.560
Soft Voting – Top 5 (RF, DT, LR, AB, NB)	95.571	89.743	96.820	**92.750**

Pagina 5 https://mendel-journal.org/index.php/mendel/article/view/211/194

Sustento 4

Marco de Referencia para realizar las comparaciones pertinentes

Model	Dataset	Cleaning	Accuracy	F1-Score	Precision	Recall	AUC
LR	1	Lemma	0.9210	0.9073	0.9384	0.8783	0.9165
RF	2	Lemma	0.9188	0.9011	0.9578	0.8507	0.9110
LSA+SVM	1	Lemma	0.9248	0.9112	0.9486	0.8767	/*
BERTweet	1	Cleaned	0.9238	0.9140	0.9084	0.9197	0.9248
LDA+LR	1	Lemma	0.9190	0.9046	0.9391	0.8726	0.9140
LDA+RF	2	Lemma	0.9144	0.8973	0.9367	0.8611	0.9085

Pagina 9 - Table 6.Summary results. * The AUC value is missing due to computational capacity issue, but itmust be similar to the BERTweet's one

Sustento 4

Marco de Referencia para realizar las comparaciones pertinentes

Dataset	Approach	Model	Accuracy	Precision	Recall	F-score	ROC
Hate Speech and Offensive Language	Machine Learning Models	SVM	0.873	0.852	0.862	0.851	0.78
		KNN	0.856	0.839	0.831	0.837	0.92
		NB	0.874	0.832	0.863	0.851	0.80
		DT	0.602	0.524	0.585	0.642	0.65
		RF	0.851	0.854	0.822	0.856	0.77
		LR	0.862	0.853	0.837	0.858	0.78
	Deep Learning Models	CNN	0.892	0.895	0.898	0.896	0.93
		LSTM	0.901	0.896	0.91	0.898	0.93
		BiLSTM	0.902	0.916	0.904	0.899	0.94
Twitter Hate Speech	Machine Learning Models	SVM	0.873	0.852	0.862	0.851	0.75
		KNN	0.856	0.839	0.831	0.837	0.90
		NB	0.874	0.832	0.863	0.851	0.76
		DT	0.602	0.524	0.585	0.642	0.68
		RF	0.851	0.854	0.822	0.856	0.77
		LR	0.862	0.853	0.837	0.858	0.78
	Deep Learning Models	CNN	0.892	0.895	0.898	0.896	0.92
		LSTM	0.901	0.896	0.91	0.898	0.92
		BiLSTM	0.902	0.916	0.904	0.899	0.93
Cyberbullying	Machine Learning Models	SVM	0.873	0.852	0.862	0.851	0.75
		KNN	0.856	0.839	0.831	0.837	0.80
		NB	0.874	0.832	0.863	0.851	0.79
		DT	0.602	0.524	0.585	0.642	0.67
		RF	0.851	0.854	0.822	0.856	0.78
		LR	0.862	0.853	0.837	0.858	0.78
	Deep Learning Models	CNN	0.892	0.895	0.898	0.896	0.91
		LSTM	0.901	0.896	0.92	0.898	0.91
		BiLSTM	0.902	0.916	0.904	0.899	0.93

Pagina 403 - TABLE I. COMPARISON OF THE OBTAINED RESULTS