Nombre: Detector de Ciberacoso “Seguras”

Autora: María Angelica Barreto Cifuentes

Programa: Talento Tech Bootcamp en Inteligencia Artificial. Bogotá. Nivel Básico

Proyecto: Sistema de Prevención de Violencia de Género en Línea “Seguras”

Utilizar inteligencia artificial para monitorear redes sociales y plataformas digitales en busca de señales de acoso y violencia de género, alertando a las autoridades pertinentes y proporcionando recursos a las víctimas.

Problema

El acceso no supervisado a redes sociales expone especialmente a los jóvenes a riesgos como extorsión, sextorsión y grooming, siendo el ciberacoso uno de los delitos más comunes. Según el INEGI (2023), el 33% de los jóvenes a nivel mundial ha experimentado ciberacoso, y las mujeres tienen 1.3 veces más probabilidades de ser víctimas que los hombres. En Colombia, el 21% de las mujeres trabajadoras ha enfrentado acoso digital vinculado a su entorno laboral, lo que afecta su desempeño profesional (OEM, 2024). Además, el 75.6% de los casos reportados de violencia de género en 2024 han sido contra mujeres, y 1 de cada 3 mujeres en la región experimenta algún tipo de violencia en línea (UNFPA-Colombia).

El ciberacoso también impacta gravemente a niños y adolescentes, siendo el 27% de ellos víctimas, con un 40% acosados por compañeros de sus instituciones educativas (Chacón, 2023). En el caso de mujeres y niñas, el 37% ha sufrido acoso a través de actos como humillaciones, difamaciones y agresiones sexuales en línea. Según el Observatorio para la Equidad de las Mujeres (2024), el 77% de las víctimas de ciberacoso ha enfrentado también otras formas de violencia, muchas veces por parte de sus parejas, lo que provoca que abandonen las redes sociales, afectando sus oportunidades laborales y sociales. Estas cifras reflejan la urgente necesidad de proteger a las personas vulnerables en entornos digitales.

Descripción de la idea (solución propuesta)

“SEGURAS” tiene como objetivo utilizar tecnologías de inteligencia artificial (IA) para monitorear redes sociales, foros, y otras plataformas digitales en busca de señales de acoso y violencia de género. El sistema no solo identificaría comportamientos violentos o amenazas, sino que también alertaría a las autoridades competentes, proporcionaría recursos y ayuda a las víctimas, y fomentaría la concienciación social sobre la violencia de género en línea.

Esta herramienta permite identificar lenguaje de ciberacoso en textos, utilizando un modelo de Machine Learning que clasifica los textos como "acoso" o "no acoso". La interfaz se construye en Gradio

Impacto

Crear un entorno en línea más seguro para mujeres y jóvenes, permitiendo respuestas rápidas y efectivas ante situaciones de violencia, así como incrementar la conciencia sobre el problema

Población beneficiaria

Mujeres entre 8 y 70 años, con acceso a internet, con redes sociales propias, cuyo conocimiento o experiencia en la convivencia en redes sociales es poco, y por lo mismo tienen una probabilidad alta de vulnerabilidad. Si bien se tomó esta población para el estudio, la herramienta no solo se limita a esta población.

Versión

Seguras, se encuentra en su versión alfa la cual permite hacer pruebas, para ingresar texto y que el modelo determine si es o no es acoso, el texto respectivamente, clasificando el texto ingresado. El código usa lenguaje Python; el modelo de machine learning Logistic Regression para la clasificación, y para la interfaz gráfica se usa Gradio.

Lenguaje utilizado en el proyecto

- ✓ Python

Librerías de Python y la razón de estas

- Pandas: Permite la analizar y trabajar con los datos ingresados desde la base de datos disponible.
- Re: Su fin es facilitar la manipulación de patrones de texto. En este caso, se utiliza en el presente código para limpiar los textos de la base de datos, y así el modelo se puede centrar en los textos importantes.
- Sklearn: Esta librería ofrece herramientas para preprocesar datos, entrenar y evaluar modelos. En este código se utiliza con las siguientes herramientas:
    - TfidfVectorizer: permite que el modelo procese las palabras como datos numéricos, esta herramienta ayudo a convertir ACOSO en 0 y NOACOSO en 1
    - Model_selection: Para este caso, se usó un modelo 80%-20%

    - LogisticRegression: Se usa este modelo de machine learning para clasificar los datos ingresados, entre acoso (0) y no acoso (1)
    - Metrics: Evalúa el rendimiento del modelo usando métricas como la precisión
    - Accuracy_score, classification_report: Da los resultados del modelo utilizado
- Nltk (Natural Language Toolkit): Permite manipular el texto en español. Para este código se usó "stop words" para delimitar palabras repetidas.
- Joblib: Se usa para almacenar el presente modelo.
- Gradio: Se usa esta interfaz ya que permite conectarse con Google Collab.
- String

Archivo de datos:

El archivo de datos escogido para este ejercicio se llama "Dataset-Acoso-Twitter-Es". Este archivo de datos se escogió por ser de las pocas bases de datos que reúnen frases de acoso en el idioma español.

La fuente es la comunidad "I Hackathon Somos NLP: PLN en Español" del entorno de alojamiento de modelos de desarrollo "Hugging Face". Desde esta comunidad se ofrecen bases de datos para el uso libre, con el fin de realizar modelos que ayuden al desarrollo y cumplimiento de los Objetivos de Desarrollo Sostenible.

Esta base de datos fue creada por la Universidad Nacional de Loja (Perú); y los profesionales que hicieron arte de este proyecto son Anderson Quizhpe; Luis Negrón; David Pacheco; Bryan Requenes, y Paul Pasaca.

En el código del modelo se integro esta base de datos por medio de la librería de Pandas, de la siguiente manera.

```
dataset = pd.read_csv("hf://datasets/somosnlp-hackathon-2022/Dataset-Acoso-Twitter-Es/datasetfinal.csv")
dataset
```

La base de datos consta de 5 columnas y 474 filas, para un total de 2370 casillas de datos. Las columnas son las siguientes "['id', 'source', 'languaje', 'text', 'task1'". *Id* se refiere al numero de la fila; *source* es la fuente en este caso todos los datos provienen de Twitter; *languaje* es el idioma usado en la tabla es decir españo; *text* es el texto recogido en los comentarios twits retwist en esta respectiva red social; finalmente, *task1* es la identificación de "acoso" o "noAcoso" para cada texto alojado en la base de datos.

Para el fin del modelo de usa como variable dependiente “task1” y variable independiente “text”, las demás columnas se retiran del ejercicio ya que por la naturaleza del mismo se sabe que es una tabla que solo maneja el idioma “Español” y que la fuente de todos los datos es Twitter, asimismo el Id solo el es el orden de entrada de cada dato.

La siguiente es la información de la base de datos rescatada durante el código

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 474 entries, 0 to 473
Data columns (total 5 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   id        474 non-null    int64
 1   source    474 non-null    object
 2   languaje  474 non-null    object
 3   text      474 non-null    object
 4   task1     474 non-null    object
dtypes: int64(1), object(4)
memory usage: 18.6+ KB
```

Hugging face (2022) Dataset-Acoso-Twitter-Es. Referenciado de somosnlp-hackathon-2022/Dataset-Acoso-Twitter-Es · Datasets at Hugging Face

Estructura de Archivos

!pip install gradio #instalación de gradio como dependencia de python

Detección de ciberacoso/

Seguras.py #codigo para entrena y guardar el modelo
cibeacoso_vectorizador.pkl # Vectorizador TF-IDF guardado
cibeacoso_model.pkl # Modelo guardado
gr.Interface # Lógica de interfaz para conectar con Gradio

Se copia el link de la interface de Gradio

Descripción del Código

a. Entrenamiento del Modelo (Seguras1.py)

El archivo Seguras1.py incluye:

- Carga de datos: Carga y preprocesa un conjunto de datos de frases con etiquetas de ciberacoso. Fuente “Dataset-Acoso-Twitter-Es”.
- Limpieza de texto: Elimina URL, menciones, hashtags y puntuación, y convierte el texto a minúsculas.

- Vectorización de texto: Utiliza TF-IDF para transformar el texto en vectores numéricos.
- Entrenamiento: Entrena un modelo de regresión logística.
- Guardado de modelo: Guarda el modelo y el vectorizador para usarlos en la extensión.

Ejemplos de fragmentos de código:

Importaciones

```
import pandas as pd
import re
import nltk
import string
```

Preprocesamiento de texto

```
def clean_text(text):
  return text
```

Cargar y limpiar datos

```
#limpio el dato/trato de limpiarlo pero no me deja #ACA ESTOY CON
PROBLEMAS
dataset.loc[:, 'text_clean'] = dataset['text'].apply(clean_text)
#imprimirlas
dataset
```

Vectorización

```
vectorizador = TfidfVectorizer(max_features=5000)
```

Entrenar modelo y Rsultados

```
#voy a entrenar con logisti regression
model = LogisticRegression(random_state=42, max_iter=1000)
model.fit(X_train_vect, y_train)
```

```
Accuracy: 0.8210526315789474

Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.67      0.77        42
           1       0.78      0.94      0.85        53

    accuracy                           0.82        95
   macro avg       0.84      0.81      0.81        95
weighted avg       0.84      0.82      0.82        95
```

Guardar el modelo

```
joblib.dump(model, 'cibeacoso_model.pkl') # Guardar el modelo para usarlo en la extensión
```

Input/Output

```
def predict_cibeacoso(text):
```

Interfaz con Gradio

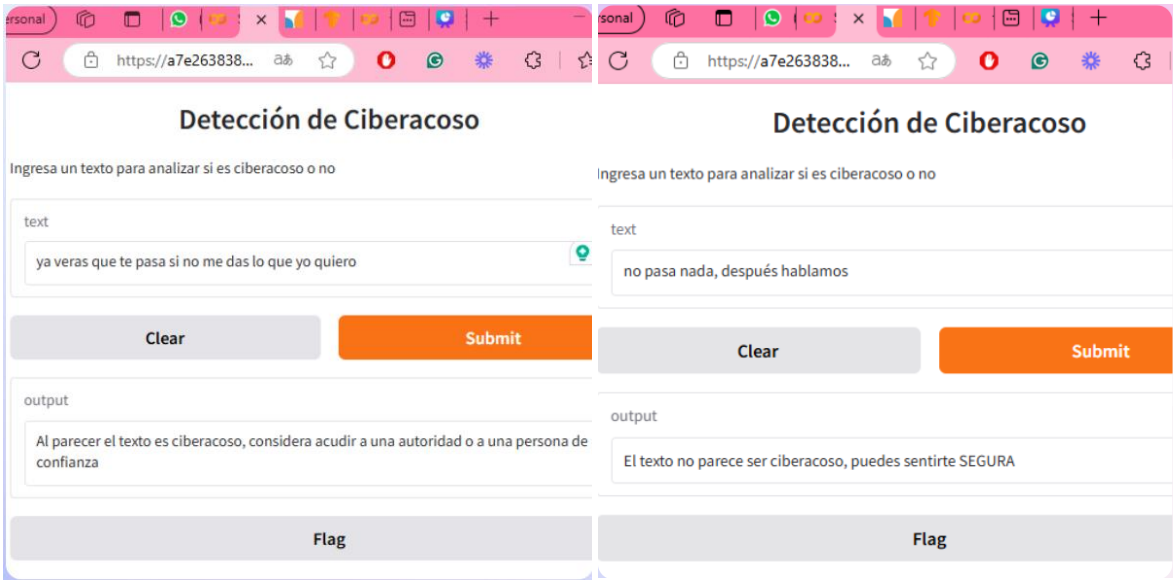
```
interface = gr.Interface ()
```

Instrucciones de Ejecución

a. Entrenamiento del Modelo

Ejecuta Seguras1.py para crear y guardar el modelo y el vectorizador.
b. Lanzar la Interfaz de Gradio
Ejecuta Seguras1.py para abrir la interfaz Gradio.

Resultados



Mantenimiento y Actualización

- ✓ **Actualizar el modelo:** Actualizar el archivo Seguras1.py para mejorar la precisión del modelo.
- ✓ **Mejoras del modelo:** Avanzar en el desarrollo para que el modelo pueda usar técnicas de procesamiento de lenguaje natural profesionalizándolo con Transformers u otras herramientas. Agregar técnicas de lectura generativa tipo GPT, para respuestas mas acertadas.
- ✓ **Actualización de datos:** Reentrenar el modelo cuando se tenga acceso a nuevos datos de ciberacoso, o a una base de datos más amplia con mas datos, incluyendo expresiones regionalistas, de distintas partes de Hispanoamérica.
- ✓ **Migrar a otros entornos:** Migrar a otros entornos de desarrollo, para así integrar archivos .html .json para generar la extensión
- ✓ **Automatizar:** Hacer que la maquina agregue a la base de datos nuevas entradas de texto para ser leídas y procesadas.
- ✓ **Redes neuronales profundas:** Seria útil para identificar patrones complejos de interacción y que el modelo aprenda.

Conclusión

El proyecto "Seguras" representa un avance significativo en el uso de inteligencia artificial para abordar el problema del ciberacoso y la violencia de género en línea. A través del desarrollo de un sistema que identifica de manera eficiente lenguaje de acoso en textos, este proyecto no solo ofrece una herramienta práctica para la detección y prevención de estas conductas, sino que también sienta las bases para

una mayor concienciación sobre un problema global que afecta de manera desproporcionada a mujeres y niñas.

Con la implementación de tecnologías como Gradio, Logistic Regression, y herramientas de procesamiento de lenguaje natural como NLTK, "Seguras" logra proporcionar una solución accesible y funcional en su versión alfa. Sin embargo, el éxito del modelo no se limita a su capacidad técnica; también radica en su enfoque ético y su potencial de impacto positivo en comunidades vulnerables.

A medida que se avanza hacia futuras actualizaciones, como la incorporación de redes neuronales profundas, la ampliación de bases de datos y la profesionalización del modelo con técnicas como Transformers, "Seguras" puede convertirse en una solución robusta que no solo detecta ciberacoso, sino que también fomenta un entorno digital más seguro y equitativo.

En última instancia, este proyecto destaca el poder de la inteligencia artificial como herramienta de cambio social, invitando a la reflexión y a la acción conjunta para erradicar la violencia digital y promover un uso responsable y seguro de las plataformas en línea.

Referencias:

Chacón, P (2023). Ciberbullying: ONU reveló los altos índices de este delito que se presentan en Colombia. Infobae. Referenciado de: https://www.infobae.com/colombia/2023/03/08/ciberbullying-onu-revelo-los-altos-indices-de-este-delito-que-se-presentan-en-colombia/

El País (2024). Conectadas pero vulnerables: esto revelan las cifras sobre violencia digital. *7 de enero de 2024*. Referenciado de: https://www.elpais.com.co/judicial/conectadas-pero-vulnerables-esto-revelan-las-cifras-sobre-violencia-digital-0725.html

Observatorio nacional de violencias de genero (2024) Sistema integrado de información de violencias de género-SIVIGE. Referenciado de: https://www.sispro.gov.co/observatorios/onviolenciasgenero/Paginas/home.aspx

Realidades y Desafíos en la Lucha Contra la Violencias de Género en Colombia (2023). Fondo de Población de las Naciones Unidas. Referenciado de: https://colombia.unfpa.org/es/publications/realidades-y-desafios-en-la-lucha-contra-la-violencias-de-genero-en-colombia

Universidad Nacional de Loja (2022) Somosnlp-hackathon-2022/Dataset-Acoso-Twitter-Es [Conjunto de datos]. Hungging Face. Referenciado de: somosnlp-hackathon-2022/Dataset-Acoso-Twitter-Es · Datasets at Hugging Face