🚀 **NUEVA VERSIÓN DEL PROYECTO INTEGRADOR**

Título:

“Detección de ciberacoso en publicaciones en español usando redes neuronales recurrentes (LSTM)”

🧭 **1. PLANTEAMIENTO DEL PROBLEMA**

1. Contexto

En el Perú, el ciberacoso se ha consolidado como una de las principales formas de violencia digital, afectando especialmente a adolescentes y jóvenes. Según el Ministerio de la Mujer y Poblaciones Vulnerables (MIMP, 2024) y el INEI (2023), cerca del **70% de los jóvenes entre 12 y 24 años** ha presenciado o sufrido algún tipo de agresión en línea, mientras que **1 de cada 3** ha sido víctima directa. Las plataformas más asociadas son **Facebook (58%)**, **Instagram (23%)** y **TikTok (14%)**, donde predominan los insultos, la difusión de imágenes sin consentimiento y la suplantación de identidad.
Esta realidad genera graves consecuencias emocionales y sociales —como ansiedad, aislamiento y deserción escolar— y evidencia la necesidad de herramientas que permitan una detección oportuna y confiable de comportamientos abusivos en entornos digitales.

---

2. Problemática detectada

Actualmente, la identificación del ciberacoso en redes sociales se realiza de manera **manual y reactiva**, dependiendo del reporte de usuarios o de equipos humanos de moderación. Este proceso es **lento, subjetivo y costoso**, lo que impide responder a tiempo ante conductas dañinas.
A pesar del avance de la inteligencia artificial, **no existen herramientas automatizadas y accesibles en español** que identifiquen con precisión el ciberacoso en textos digitales.
Los grandes modelos de lenguaje (LLMs), como ChatGPT o LLaMA, han demostrado capacidad para interpretar el contexto y la intención de los mensajes, pero aún presentan **limitaciones en el manejo del español local, ambigüedades lingüísticas y sesgos culturales**, lo que afecta su desempeño en escenarios reales.
Esta brecha tecnológica impide contar con soluciones efectivas para **monitorear, clasificar y alertar de manera temprana** sobre comportamientos de ciberacoso en redes sociales peruanas.

---

3. Hipótesis

La implementación de un modelo de detección de ciberacoso en español, basado en técnicas de aprendizaje automático y modelos de lenguaje de gran escala (LLMs), **permitirá identificar de manera precisa y automática mensajes con contenido**

abusivo en redes sociales, reduciendo la dependencia del análisis manual y mejorando la capacidad de respuesta ante situaciones de violencia digital.

---

4. Objetivo general

Desarrollar e implementar un sistema automatizado de detección de ciberacoso en español, basado en modelos de lenguaje y aprendizaje automático, que permita **identificar y clasificar mensajes abusivos en redes sociales de manera rápida, precisa y ética**, contribuyendo a la prevención y mitigación de la violencia digital en el Perú.

🔹 e. Objetivos específicos

1. Recolectar y preparar datos etiquetados en español de redes sociales con presencia o ausencia de lenguaje abusivo.
2. Aplicar técnicas de limpieza y preprocesamiento de texto (tokenización, lematización, stopwords).
3. Entrenar una **red neuronal LSTM** con TensorFlow para clasificar mensajes según su nivel de agresividad.
4. Evaluar el rendimiento del modelo mediante métricas de clasificación (accuracy, recall, F1-score).
5. Implementar una **API web con FastAPI** para analizar mensajes nuevos en tiempo real.

🧠 2. ACCESO A DATOS

🔹 a. Datasets seleccionados

Se utilizarán fuentes públicas y académicas que contienen **textos en español etiquetados** por tipo de agresión o discurso de odio:

Fuente	Descripción	Tipo de texto	Enlace
sp_tweets_cyberbullying	Tweets en español con etiquetas de ciberacoso	Twitter	GitHub – ximenamar/sp_tweets_cyberbullying
Hate Speech Spanish Superset	Dataset unificado de hate speech en español	Twitter, foros	Hugging Face – manueltonneau/spanish-hate-speech-superset

Fuente	Descripción	Tipo de texto	Enlace
MDPI Dataset (Appl. Sci. 2021)	Dataset académico validado sobre detección de ciberacoso en español	Twitter	MDPI Paper – Appl. Sci. 2021, 11(22), 10706

🔹 **b. Tipo de datos**

- **No estructurados (texto corto).**
- Variables:
  - *text* → contenido del tweet o mensaje
  - *label* → 0 = no ofensivo / 1 = ofensivo

🔹 **c. Volumen estimado**

Entre **20,000 y 60,000 ejemplos**, suficientes para entrenar un modelo de deep learning.

🔹 **d. Ética y confidencialidad**

Los datasets son públicos y anonimizados. No se usa información personal.
El proyecto tiene **fines académicos y sociales** (no comerciales).

⚙️ **3. TIPO DE SOLUCIÓN A ELABORAR**

🔹 **a. Técnica de IA**

- **Red neuronal recurrente (LSTM)** implementada con TensorFlow/Keras.
- Arquitectura:
  1. **Embedding Layer:** conversión de palabras en vectores.
  2. **LSTM Layer:** captura de dependencias semánticas.
  3. **Dropout:** reducción de sobreajuste.
  4. **Dense Layer (sigmoid):** salida binaria (ciberacoso o no).

🔹 **b. Flujo del modelo**

🔹 **1. Recolección de datos**

Obtención de textos en español desde datasets abiertos (*sp_tweets_cyberbullying*, *Spanish Hate Speech Superset*, *MDPI*).
→ Se conforma un dataset etiquetado (ofensivo / no ofensivo).

🔹 **2. Preprocesamiento**

Limpieza y normalización de texto:

- **Eliminación de símbolos, emojis y URLs**
- **Tokenización y lematización**
- **Eliminación de *stopwords***
  Resultado: textos convertidos a secuencias numéricas.

🔹 **3. Vectorización**

Cada palabra se transforma en vectores mediante una capa Embedding, capturando el significado semántico.

🔹 **4. Entrenamiento del modelo LSTM**

- **Arquitectura: *Embedding* → *LSTM* → *Dropout* → *Dense (sigmoid)***
- **Entrenamiento con *binary_crossentropy* y optimizador *Adam*.**
- **Validación y ajuste de hiperparámetros.**

🔹 **5. Evaluación**

Se calculan métricas de desempeño:
***accuracy*, *precision*, *recall*, *F1-score* y matriz de confusión.**

🔹 **6. Despliegue**

El modelo se guarda (.h5) y se despliega:

- **Local: API con FastAPI para análisis en tiempo real.**
- **En la nube: Azure ML Studio (endpoint REST para integración externa).**

🔹 **c. Herramientas**

Etapa	**Herramienta**
Limpieza de texto	Python, NLTK, spaCy
Modelado	TensorFlow / Keras
Evaluación	Scikit-learn
API Web	FastAPI
Despliegue	Azure Machine Learning Studio o Azure Functions

📊 **4. CRONOGRAMA DE TRABAJO (ajustado desde 8 de setiembre)**

Semana	Fechas	Actividades	Entregables
1	8–14 set	Definición del problema, búsqueda de datasets y objetivos del proyecto	Documento de planteamiento
2	15–21 set	Descarga y exploración de datasets seleccionados	CSV unificado
3	22–28 set	Limpieza, tokenización y vectorización del texto	Dataset preprocesado
4	29 set–5 oct	Entrenamiento inicial del modelo LSTM	Notebook con resultados preliminares
5	6–12 oct	Evaluación y optimización del modelo	Reporte con métricas finales
6	13–19 oct	Implementación de API con FastAPI	API funcional local
7	20–26 oct	Despliegue en Azure Machine Learning Studio	Endpoint o demostración
8	27 oct–2 nov	Redacción del informe técnico final	Documento Word/PDF
9	3–9 nov	Preparación de presentación ejecutiva	PowerPoint ejecutiva
10	10–20 nov	Sustentación final	Presentación y código final

📈 5. RESULTADOS ESPERADOS

- Un modelo capaz de **detectar mensajes de ciberacoso** con una precisión >85%.
- API REST accesible para analizar texto en español.
- Posibilidad de integración en aplicaciones educativas o de redes sociales.
- Contribución ética: promover un **uso responsable y seguro del lenguaje digital**.

☁️ 6. DESPLIEGUE EN AZURE

🔹 **Plan de despliegue:**

1. Exportar el modelo entrenado como archivo .h5.
2. Subirlo a **Azure Machine Learning Studio** y registrarlo.
3. Crear un **inference pipeline** y **endpoint REST**.
4. Conectarlo a FastAPI o Postman para pruebas.

🔹 **Alternativa:**

Si el tiempo lo impide, usar **Azure Cognitive Services (Language)** para mostrar la integración de un servicio IA en la nube con textos en español.

🎯 7. CONCLUSIÓN EJECUTIVA

El proyecto demuestra la **aplicación práctica de la inteligencia artificial en un problema social relevante**: la detección de ciberacoso en redes sociales.
Usando **procesamiento de lenguaje natural (NLP)** y **redes neuronales recurrentes (LSTM)**, se propone una herramienta escalable capaz de analizar mensajes en español, ofreciendo una solución con **valor real, impacto social y despliegue profesional en Azure**.

📎 Formato de Presentación Ejecutiva (recomendado para PowerPoint)

Diapositiva	Contenido	Duración
1	Título, autor, diplomado, docente	30 seg
2	Contexto y problemática	1 min
3	Hipótesis y objetivos	1 min
4	Dataset y fuentes	1 min
5	Metodología (pipeline LSTM)	1.5 min
6	Resultados esperados	1 min
7	Despliegue en Azure	1 min
8	Conclusiones e impacto	1 min
