**Título del proyecto:**

Sistema de análisis de sentimientos en comentarios de noticias online utilizando procesamiento de lenguaje natural (NLP) e Inteligencia Artificial

**Contexto:**

En la actualidad, los portales de noticias vía online generan un alto nivel de interacción por parte de los usuarios, quienes manifiestan sus **opiniones y emociones** en los comentarios o interacciones de cada publicación. Las **interacciones** representan una valiosa fuente de información para conocer la percepción de la ciudadanía respecto a hechos de c**ontingencia policial, política, económica, deportiva, etc.**

El análisis manual de los comentarios y/o interacciones, que reflejan las emociones de los usuarios, es **inviable** debido al gran volumen de datos. Sin embargo, el uso de técnicas NLP y Machine Learning permite **automatizar** la clasificación de sentimientos, entregando resultados objetivos y rápidos.

**Planteamiento del problema:**

Actualmente en Chile, no existe una herramienta automatizada y de libre acceso que permita recolectar, procesar y analizar los sentimientos expresados por los usuarios en los portales de noticias online (Meganoticias, t13, TVN, Chilevisión, ADN, CNN, Cooperativa, etc). Esto genera una pérdida de oportunidades para investigadores, periodistas, entidades sociales que buscan **conocer de manera ágil la opinión pública**

**¿Qué hacer? (objetivo general)**

**Desarrollar un sistema web**, capaz de **recolectar información** sobre comentarios de sitios de noticias online y aplicar modelos de análisis de sentimientos, **clasificando las interacciones en positivas, negativas o neutras,** con el fin de visualizar métricas y tendencias para **obtener conclusiones sobre la percepción ciudadana.**

**Objetivos específicos:**

1.- **Implementar un módulo de recolección de datos** desde portales de noticias online mediante API’S o una Web Scrapping autorizado

2.- **Diseñar** un proceso de **preprocesamiento y estandarización de texto** para comentarios en línea.

3.- **Entrenar y evaluar** modelos de análisis de sentimientos utilizando técnicas de **Machine Learning y Deep Learning** (LSTM, BERT, etc).

4.- **Desarrollar un Dashboard Web Interactivo** que muestre los resultados de todo el proceso anterior en tiempo real, incluyendo:

* Métricas
* Gráficas
* “Nubes” de palabras clave

5.- **Evaluar el desempeño** del sistema mediante métricas de **precisión, recall, F1-Score y exactitud**

**Alcance del Proyecto:**

Definición de las cosas que se van a desarrollar.

**Incluye**

* **Recolección** de datos desde un conjunto acotado (muestra) de sitios de medios de comunicación masivo (que tenga que ver tanto con la televisión, diarios online, radio, etc)
* **Entrenamiento** de un modelo de análisis de sentimiento.
* Implementación de un prototipo funcional de la página web junto con el dashboard interactivo.
* **Visualización** de resultados agregados, como por ejemplo:

- **Promedio de reacciones** por parte de los usuarios.

- **Tendencias de sentimientos y palabras claves más mencionadas** (de esta última idea viene la “Nube” de palabras claves)

**No incluye**

* El sistema no contiene cobertura de todas las plataformas sociales (redes sociales) como lo puede ser Facebook, Instagram, X, etc.
* Análisis **masivo** en tiempo real; solo prototipo con capacidad de escalabilidad futura.
* Interpretación sociológica/política de los resultados obtenidos; solo se entregará la herramienta técnica

**Metodología de Desarrollo:**

Metodología Ágil SCRUM, dividida en 5 fases:

* **Fase 1: Recolección de datos:** Implementar web scrapping/API y construir dataset inicial con la muestra.
* **Fase 2: Procesamiento y Entrenamiento:** Preprocesar el texto (Tokenizándolo, limpiando stopword, stemming/lemmatizacion).
* **Fase 3: Desarrollo Web:** 
  + Backend: API REST **(Utilizando Flask de Python o FastAPI)**
  + Frontend: Interfaz amigable con el usuario **(React puede ser una buena alternativa).**
  + Base de datos: MongoDB/PostgreSQL
* **Fase 4: Visualización de resultados:** Dashboard interactivo con métricas, gráficos y nube de palabras más utilizadas por los usuarios en función a sus interacciones.
* **Fase 5: Evaluación de los resultados y conclusiones:** Medición de **rendimiento del modelo,** que contendrá también el informe de los resultados y sus potenciales aplicaciones.

**Herramientas a utilizar:**

* Lenguajes: Python, JS
* Frameworks Backend: Flask / FastAPI
* Procesamiento NLP:
  + **NLTK**: Biblioteca gratuita y de código abierto para el procesamiento avanzado del lenguaje natural en Python.
  + **SpiCY**: Librería; facilita tareas de procesamiento avanzado del lenguaje natural.
  + **HuggingFace Transformers** (BETO, DistilBERT: Enfoque para detectar texto generado por grandes modelos de lenguaje (LLM))
  + **Machine learning**: Scikit-learn, TensorFlow/Keras, PyTorch
  + **Bases de datos**: PostgreSQl o MongoDB.
  + **Infraestructura**: En Docker (Despliegue en la nube, como por ejemplo, AWS o Azure)
  + **Visualización de resultados**: Plotly/Dash, PowerBi

**Resultados Esperados:**

Para partir, se necesita que el **prototipo web sea funcional**, y que recolecte los comentarios de las noticias, o reacciones, en los portales online dedicados; clasifiquen los sentimientos en positivos, neutro o negativos; y que presente los resultados visuales.

Se requiere que las **métricas cuantitativas** del modelo a evaluar cuenten con un correcta precisión, recall y que el F1-Score sea mayor al 80%.

El informe final cuente con las conclusiones esperadas sobre la **percepción ciudadana frente a un conjunto de noticias de contingencia.**

**Ventajas e Impacto del Proyecto:**

**Impacto tecnológico**: Este proyecto combina NLP, Machine Learning y desarrollo web en un solo sistema.

**Académico**: Con el enfoque puesto en lo que sería la simulación de un caso real, este proyecto aporta con un caso de estudio real sobre análisis de sentimientos, que recopila comentarios en español sobre las interacciones y reacciones de los usuarios sobre noticias de contingencia tanto nacional como internacional, se procesa y devuelve información concreta para su posterior análisis para, finalmente, culminar con una pequeña tesis sobre lo observado.

**Social**: El proyecto brinda una herramienta útil y amigable para investigadores, periodistas y/o instituciones que deseen medir la opinión pública.

**Escalable**: Podría adaptarse a redes sociales u otros portales de interacción masiva.

**Glosario:**

**Stopwords**: Las palabras vacías son palabras de uso común, como artículos, pronombres y preposiciones.

**Stemming:** El stemming elimina prefijos y sufijos de las palabras para obtener su raíz. Esta raíz puede no ser una palabra real en el idioma, pero ayuda a identificar la palabra base.

**Lematizar:** Proceso de reducir una palabra a su forma base o “lema”, que es su forma canónica en el diccionario