Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Estado de México

Evidencia 1 - Fase 2 - Parte B: Implementación usando NLP

Miguel González Mendoza

Raúl Monroy Borja Raúl Monroy Borja

Ariel Ortíz Ramírez

Jorge Adolfo Ramírez Uresti

Adolfo Sebastián González Mora	A01754412
Jorge Daniel Rea Prado	A01747327
Marco Antonio Caudillo Morales	A01753729
Oswaldo Daniel Hernandez de Luna	A01753911

Introducción

El presente proyecto tiene como finalidad desarrollar un sistema automatizado capaz de identificar publicaciones en redes sociales que estén asociadas a trastornos de la conducta alimentaria, con un enfoque particular en la detección de contenidos vinculados a la anorexia. Para ello, se emplean técnicas avanzadas de procesamiento de lenguaje natural (NLP, por sus siglas en inglés) junto con algoritmos de aprendizaje automático (Machine Learning).

El objetivo principal es construir un modelo que pueda analizar textos, tales como tweets u otros tipos de mensajes breves, y discernir si estos presentan indicios relacionados con conductas propias de la anorexia. Este proceso no solo implica la clasificación binaria de textos (anorexia vs. control), sino también una comprensión más profunda del lenguaje y del contexto en el que se emplean ciertas palabras o expresiones relacionadas con el tema.

Motivación

La motivación detrás de este trabajo radica en la necesidad de contar con herramientas automatizadas que ayuden en la detección temprana de señales de alerta en contenidos digitales, con el fin de apoyar esfuerzos de prevención, monitoreo y eventual intervención en el ámbito de la salud mental. Al aplicar NLP y Machine Learning, se busca ir más allá del análisis superficial de palabras clave, incorporando técnicas que permitan capturar el significado contextual, la semántica y las relaciones entre términos, logrando así una identificación más precisa y confiable.

Flujo del Proyecto

Carga y limpieza de datos (tweets)

Se parte de un conjunto de datos compuesto por tweets etiquetados, los cuales pueden estar relacionados o no con trastornos alimenticios como la anorexia. Estos datos son cargados desde archivos CSV y posteriormente se realiza un proceso de limpieza para asegurar su calidad antes de aplicar técnicas de análisis.

La limpieza incluye:

- Eliminación de URLs, menciones a otros usuarios y hashtags.
- Conversión del texto a minúsculas.
- Eliminación de caracteres no alfabéticos y números innecesarios.
- Reducción de ruido textual, como puntuación y palabras irrelevantes (stopwords).

Tokenización y eliminación de ruido textual

Una vez limpio el texto, se procede a tokenizar, es decir, a dividir cada texto en palabras individuales o “tokens”. Esta es una etapa esencial en el procesamiento de lenguaje natural, ya que transforma el texto en una forma manejable por los modelos de aprendizaje automático.

Durante esta etapa también se eliminan:

- Stopwords en español (palabras comunes como “el”, “de”, “y” que no aportan mucho significado).
- Tokens con menos de tres caracteres.
- Palabras poco frecuentes o con errores ortográficos que no contribuyen al análisis.

Vectorización mediante TF-IDF

Los textos procesados se convierten en representaciones numéricas mediante TF-IDF (Term Frequency - Inverse Document Frequency). Este método permite:

- Asignar un peso a cada palabra según su frecuencia relativa dentro de un documento (tweet), y su rareza en el conjunto total de documentos.
- Capturar no solo la presencia de palabras clave, sino también su importancia contextual.

Se consideran n-gramas (palabras individuales, pares o tríos consecutivos) para capturar patrones semánticos más ricos que los unigramas.

Entrenamiento con Random Forest optimizado vía GridSearchCV

Se utiliza el algoritmo Random Forest, un ensamble de árboles de decisión, por su capacidad de manejar datos ruidosos y alta dimensionalidad (como es común en texto vectorizado).

Para maximizar su rendimiento, se aplica GridSearchCV, que permite:

- Explorar múltiples combinaciones de hiperparámetros (como profundidad máxima, número de árboles, y tamaño mínimo de hojas).
- Realizar validación cruzada para evitar overfitting y obtener un modelo más robusto.

Evaluación con métricas AUC, precisión, recall, curva ROC y matriz de confusión

Se evalúa el desempeño del modelo con varias métricas clave:

- AUC (Área bajo la curva ROC): mide la capacidad del modelo para distinguir entre clases.
- Precisión: proporción de predicciones positivas correctas (cuántos de los que predijo como "anorexia" lo eran realmente).
- Recall: capacidad del modelo para identificar correctamente todos los casos positivos (anorexia).
- Curva ROC: representa gráficamente la relación entre la tasa de verdaderos positivos y falsos positivos.
- Matriz de confusión: permite observar cómo se distribuyen los errores del modelo entre las clases.

Funcionalidades del Código

1. Carga de Datos
    - Lee archivos .csv con textos (tweets) y etiquetas de clase (anorexia / control).
2. Limpieza de Texto
    - Elimina URLs, menciones, hashtags, números y signos de puntuación.
    - Convierte a minúsculas y filtra palabras vacías (stopwords) en español.
3. Tokenización
    - Divide el texto en palabras (tokens) y filtra aquellas con menos de tres caracteres o sin valor semántico.
4. Generación de Texto Limpio
    - Reconstruye los textos limpios a partir de los tokens filtrados, para su posterior análisis.
5. Vectorización con TF-IDF
    - Convierte los textos a una matriz numérica utilizando n-gramas (1 a 3) y un límite de 5000 características.
    - Pondera las palabras por frecuencia y rareza.

6. Etiquetado Binario

- Transforma las clases en etiquetas numéricas (1 para anorexia, 0 para control).

7. División de Datos

- Separa los datos en conjuntos de entrenamiento y prueba usando train_test_split.

8. Entrenamiento del Modelo

- Entrena un clasificador RandomForestClassifier con hiperparámetros optimizados mediante GridSearchCV.

9. Evaluación del Modelo

- Calcula métricas de rendimiento: AUC, precisión, recall, F1-score y accuracy.
- Genera y muestra la curva ROC.
- Construye la matriz de confusión.

10. Análisis de Importancia de Características

- Identifica los n-gramas más relevantes según su importancia en el modelo Random Forest.

11. Visualización de Resultados

- Presenta gráficas de curva ROC y de los n-gramas más influyentes en la predicción.

12. Serialización del Modelo

- Guarda el modelo entrenado y el vectorizador TF-IDF en archivos .pkl para su reutilización.

13. Validación Externa

- Carga un nuevo conjunto de datos, realiza limpieza, vectorización, predicción y evaluación externa del modelo.

Resultados

El modelo desarrollado fue evaluado utilizando dos conjuntos de datos: uno de prueba (extraído del conjunto de entrenamiento) y otro de validación externa (datos completamente nuevos que el modelo no había visto antes). Esto permite medir tanto su capacidad de aprendizaje como su generalización a datos reales.

Evaluación sobre datos de prueba

Accuracy (Precisión Global): 0.90

El modelo clasifica correctamente el 90% de los tweets. Este valor es alto y refleja un buen desempeño general.

AUC (Área bajo la curva ROC): 0.96

Este valor representa la capacidad del modelo para distinguir entre clases (anorexia vs. control). Un AUC cercano a 1.0 indica una excelente discriminación. Con 0.96, el modelo demuestra ser muy eficaz en esta tarea.

Precisión (clase Anorexia): 0.87

De todos los tweets que el modelo clasificó como relacionados con anorexia, el 87% realmente lo eran. Esto sugiere que el modelo tiene una baja tasa de falsos positivos.

Recall (clase Anorexia): 0.94

El modelo detecta correctamente el 94% de los casos reales de anorexia. Esta alta tasa de verdaderos positivos indica que es poco probable que pase por alto señales importantes.

Evaluación sobre datos de validación externa

Accuracy: 0.84

En un conjunto completamente nuevo, el modelo logra un 84% de acierto. Aunque ligeramente menor al conjunto de prueba, sigue siendo un resultado sólido que indica buena capacidad de generalización.

AUC: 0.93

Un valor de 0.93 en datos externos sigue siendo excelente y reafirma que el modelo mantiene un buen desempeño discriminativo incluso en textos desconocidos.

Precisión (Anorexia): 0.83

En validación externa, el modelo clasifica correctamente el 83% de los casos positivos detectados como anorexia, lo que implica una ligera reducción pero mantiene su fiabilidad.

Recall (Anorexia): 0.90

Detecta el 90% de los casos reales de anorexia en el conjunto externo, lo cual es crítico para evitar que señales de alerta pasen desapercibidas.

Visualizaciones Clave

Curva ROC:

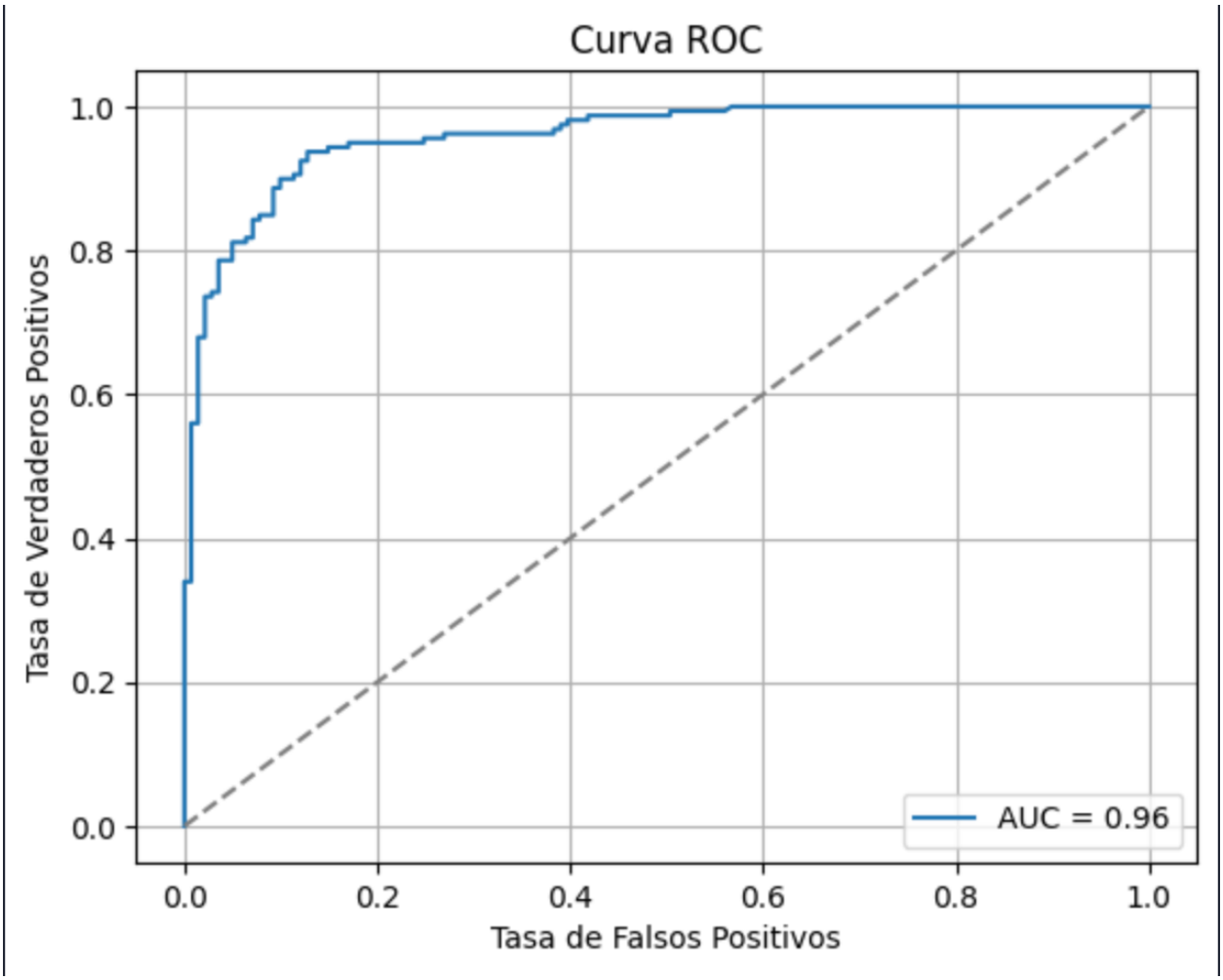
Las gráficas ROC mostraron una clara separación entre las clases, reforzando el alto valor del AUC.

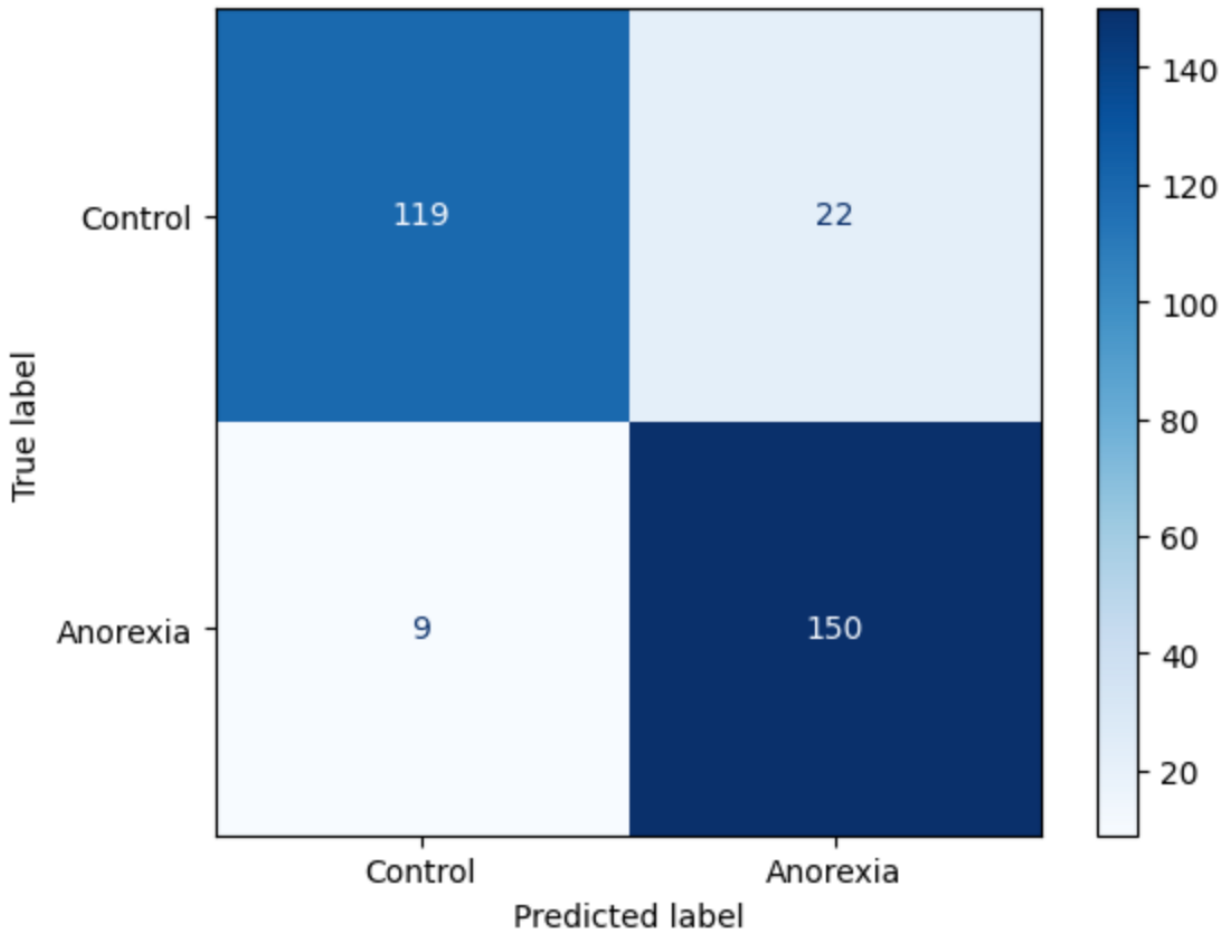
Matriz de confusión:

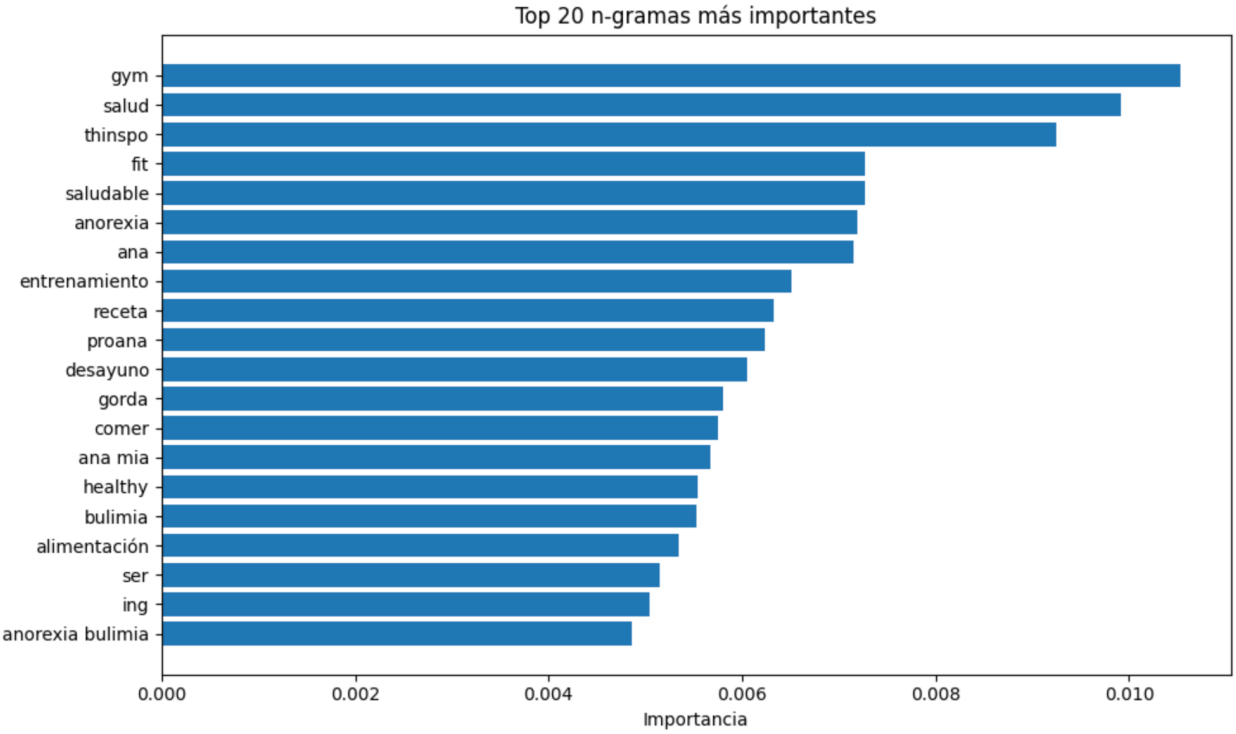
Ayudó a identificar el número de falsos positivos y falsos negativos, mostrando que el modelo comete pocos errores de clasificación.

Importancia de n-gramas:

Se identificaron los n-gramas más relevantes para la detección, lo que aporta interpretabilidad al modelo, permitiendo entender qué expresiones tienen mayor peso en la predicción.


Curva ROC
Tasa de Verdaderos Positivos
Tasa de Falsos Positivos
0.0
0.2
0.4
0.6
0.8
1.0
AUC = 0.96


Control
Anorexia
True label
119
22
9
150
Control
Anorexia
Predicted label
140
120
100
80
60
40
20



Conclusión sobre los resultados:

El modelo logra un equilibrio efectivo entre precisión y sensibilidad, lo que es fundamental para una aplicación en contextos sensibles como la salud mental. Puede detectar publicaciones relacionadas con anorexia con alta eficacia, tanto en datos conocidos como en nuevos, lo que lo hace apto para tareas de monitoreo automatizado o apoyo a intervenciones tempranas.

Conclusiones

Este proyecto demuestra la viabilidad y efectividad del uso de técnicas de procesamiento de lenguaje natural (NLP) y aprendizaje automático para identificar publicaciones relacionadas con trastornos alimenticios, específicamente anorexia, a partir de textos en redes sociales. A través de un enfoque basado en la limpieza de datos, vectorización con TF-IDF y clasificación mediante Random Forest optimizado, se logró construir un modelo con alto rendimiento predictivo, obteniendo métricas destacadas como un AUC de hasta 0.96 y un recall de 0.94, lo que indica una excelente capacidad para detectar casos relevantes sin omitir señales críticas. Estos resultados evidencian el potencial del modelo como una herramienta de apoyo para monitoreo automatizado y prevención en el ámbito de la salud mental, con posibilidades de mejora futura mediante técnicas más avanzadas como embeddings semánticos o modelos de lenguaje más sofisticados.

Apéndice

Tecnologías Usadas

Python, pandas, NumPy, scikit-learn, matplotlib, gensim, NLTK, joblib.