

Proyecto Final: Detección de Estrés

NATURAL LANGUAGE PROCESSING
7CM2

Ortega Prado Mauricio
Palmerin Garcia Diego
Pérez Gómez Andres

Contenido

Introducción y
Objetivos

03

Modelos LLM

07

Corpus y Datos

04

Comparación de
Resultados

08

Datos y
Preprocesamiento

05

Conclusiones

09

Modelos
Tradicionales

06

Trabajo a Futuro

11



Introducción

- El estrés universitario es una problemática creciente.
- Puede afectar el rendimiento y la salud mental.
- La detección temprana permite intervenir a tiempo.

Objetivo

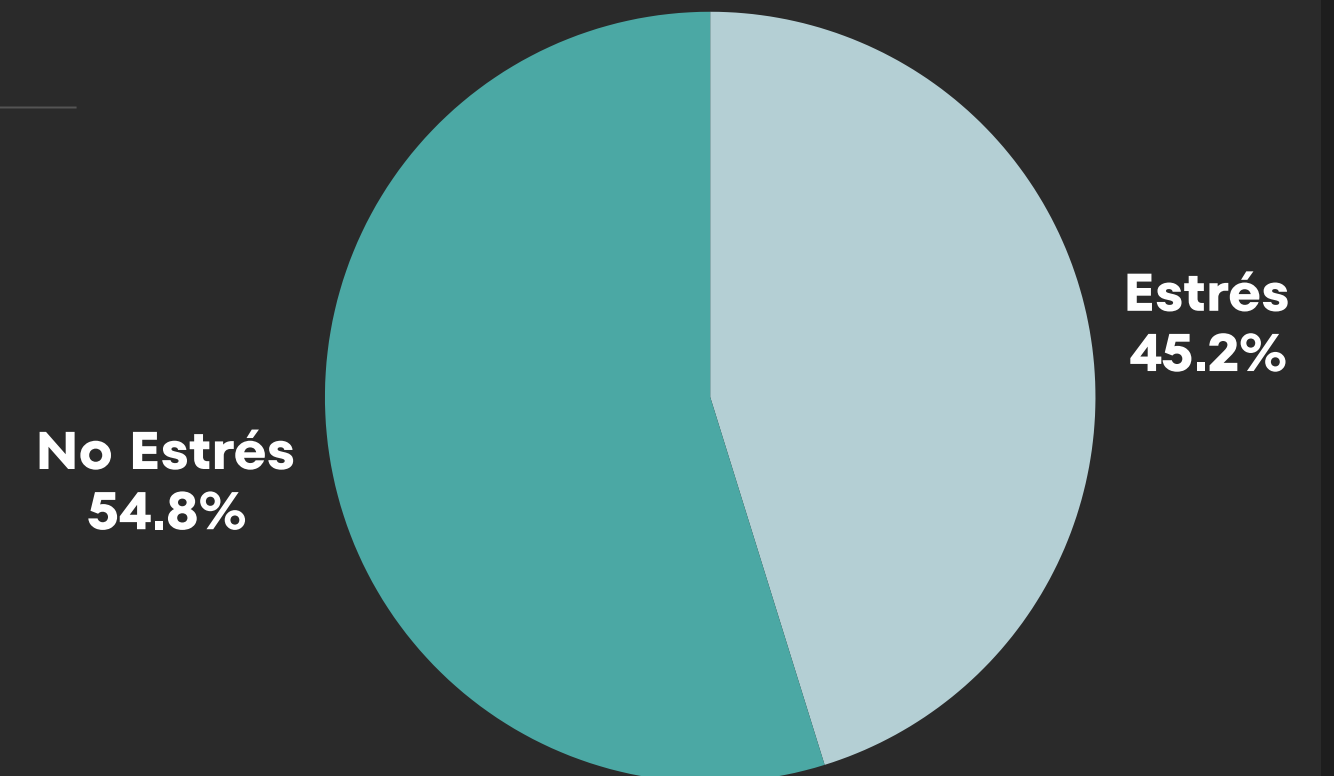
- Estimar la presencia de estrés en estudiantes a partir de textos breves.
- Datos de respuestas abiertas escritas por estudiantes.
- Clasificación automática de textos en "estrés / no estrés".
- Técnicas de PLN y modelos de lenguaje (tradicionales y LLMs).

Corpus y Datos

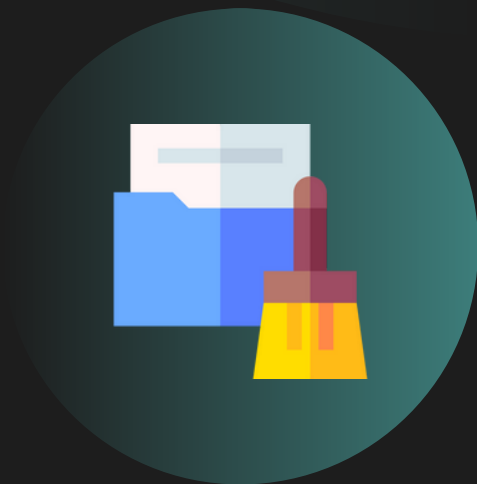
Características del Corpus

Se utilizó un corpus propio compuesto por transcripciones de respuestas abiertas generadas por estudiantes universitarios.

- ✓ Respuestas **abiertas** sobre su **experiencia emocional**.
- ✓ Tamaño del corpus de **608 textos individuales**.
- ✓ **Estrés: 275 (45.2%)** **No estrés: 333 (54.8%)**

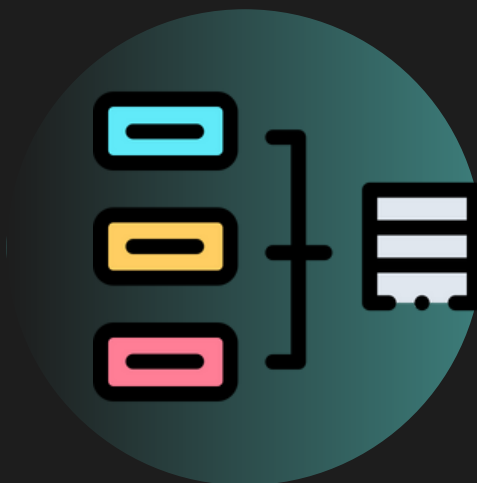


Dataset y Preprocesamiento



Limpieza de los Datos

- Conversión a minúsculas (spaCy)
- Eliminación de stopwords
- Lematización



Estructura del Dataset

ID	Pregunta	Transcripción	Etiqueta
ID01	Pregunta 1	"Me cuesta concentrarme cuando tengo demasiadas tareas."	<input checked="" type="checkbox"/> si <input type="checkbox"/>
ID02	Pregunta 2	"Últimamente me siento agotado y frustrado todo el tiempo."	<input checked="" type="checkbox"/> si <input type="checkbox"/>
ID03	Pregunta 3	"Hacer ejercicio y salir con amigos me ayuda a relajarme."	<input type="checkbox"/> no <input checked="" type="checkbox"/>

Modelos Tradicionales y Vectorización



Modelos Tradicionales

Naive Bayes
Support Vector Machine (SVM)
Regresión Logística



Vectorización del texto

TF-IDF
Frecuencia (CountVectorizer)
Binaria



N-Gramas

Unigramas (1,1)
Bigramas (2,2)
Trigramas (3,3)
Uni+Bi (1,2)
Uni+Tri (1,3)

Modelos LLM (Large Language Model)

BETO

Modelo basado en BERT, entrenado en españols

Se ajustó con Trainer de Hugging Face
Tokenización contextual

Accuracy: 83.6%
F1 macro: 83.4%



RoBERTa + LoRA

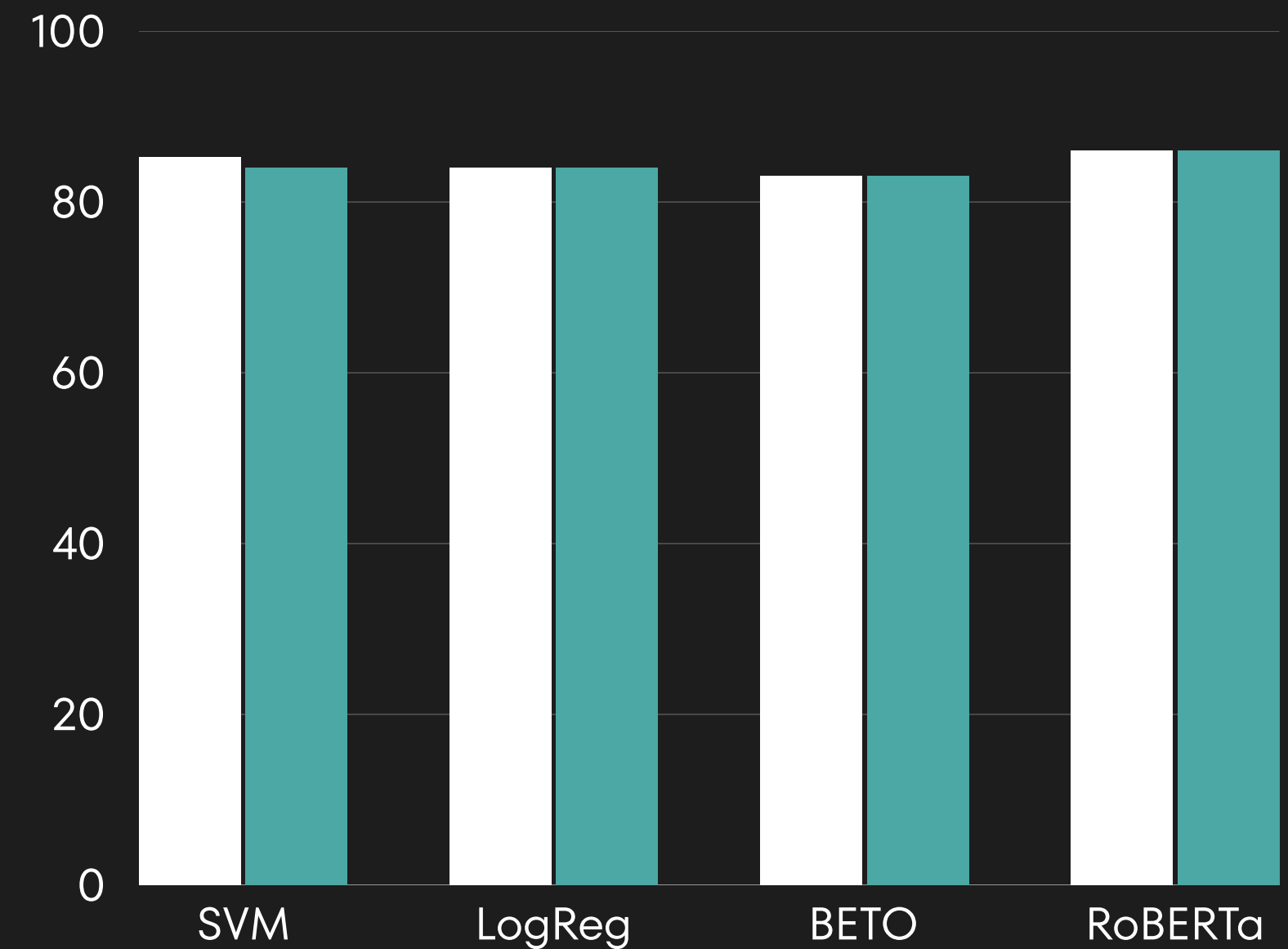
Modelo robusto preentrenado en español (roberta-base-bne)

Menor costo computacional y tiempo

Accuracy: 86.1%
F1 macro: 86.0%

Comparación de resultados

Modelo	Vectorizador	N-gramas	Accuracy 	F1_macro 
SVM	Frecuencia	Unigrama	0.8525	0.8484
Regresión Logística	Frecuencia	Unigrama	0.8443	0.8412
BETO	Tokenización BERT	—	0.8361	0.8345
RoBERTa + LoRA	Tokenización Roberta	—	0.8607	0.8602



Conclusiones

- Se logró detectar estrés en textos escritos por estudiantes con una precisión considerable, incluso con un corpus reducido.
- Los modelos tradicionales ofrecieron un rendimiento razonable, pero presentaron limitaciones para captar matices emocionales.
- El mejor desempeño lo obtuvo RoBERTa con LoRA, alcanzando un F1 macro de 0.86, superando a modelos clásicos y a BETO.
- Las técnicas de adaptación eficiente como LoRA demostraron ser útiles para trabajar con LLMs sin necesidad de grandes recursos.



Conclusiones y elección de modelos

Criterio	Modelo tradicional	LLM
Datos disponibles	Pocos	Medianos o grandes
Recursos computacionales	Limitados (sin GPU)	GPU recomendado
Tiempo de entrenamiento	Rápido	Lento
Profundidad semántica	Limitada	Alta



Trabajo a Futuro

- ✓ **Ampliar el dataset:** Más muestras mejorarían la generalización del modelo.
- ✓ **Etiquetado más detallado:** Clasificación por niveles de estrés (leve, moderado, severo) o tipo (académico, emocional, entre otros).
- ✓ **Análisis de emociones complementarias:** Detectar ansiedad, frustración, tristeza, etc.
- ✓ **Explorar nuevos modelos:** LLMs más recientes o ajustados al dominio emocional.
- ✓ **Implementación práctica:** Desarrollar una app o plataforma para monitoreo emocional en estudiantes.



Gracias por su
Atención