



# AUTEXTIFICATION

## Identificación de autoría en textos

Sobre textos de IA vs Humanos



# OBJETIVO

Desarrollar un sistema que identifique (con un considerable porcentaje de acierto) si un texto fue escrito por un ser humano o un Modelo Generativo de Lenguaje

# OBJETIVOS GENERALES

x

x



## TAREA A

Clasificar un texto en inglés bajo alguna de las dos etiquetas:

- Humano
- Máquina

x



## TAREA B

Si un texto en inglés es generado por máquina, darle una etiqueta correspondiente al modelo que lo generó



Source/ Domain	Language	Total Human	Parallel Data						Total
			Human	Davinci003	ChatGPT	Cohere	Dolly-v2	BLOOMz	
Wikipedia	English	6,458,670	3,000	3,000	2,995	2,336	2,702	3,000	17,033
Reddit ELI5	English	558,669	3,000	3,000	3,000	3,000	3,000	3,000	18,000
WikiHow	English	31,102	3,000	3,000	3,000	3,000	3,000	3,000	18,000
PeerRead	English	5,798	5,798	2,344	2,344	2,344	2,344	2,344	17,518
arXiv abstract	English	2,219,423	3,000	3,000	3,000	3,000	3,000	3,000	18,000
Baike/Web QA	Chinese	113,313	3,000	3,000	3,000	–	–	–	9,000
RuATD	Russian	75,291	3,000	3,000	3,000	–	–	–	9,000
Urdu-news	Urdu	107,881	3,000	–	3,000	–	–	–	9,000
id_newspapers_2018	Indonesian	499,164	3,000	–	3,000	–	–	–	6,000
Arabic-Wikipedia	Arabic	1,209,042	3,000	–	3,000	–	–	–	6,000
True & Fake News	Bulgarian	94,000	3,000	3,000	3,000	–	–	–	9,000
<b>Total</b>			35,798	23,344	32,339	13,680	14,046	14,344	133,551



×

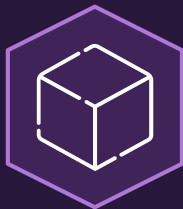
# HERRAMIENTAS

Se posee un conjunto de textos  
en inglés con sus respectivas  
etiquetas de autor

×

# FASES RELEVANTES

x



## PREPARACIÓN

Se aplicarán técnicas para extraer, limpiar y pre-procesar los textos de entrenamiento

+



## MODELOS

Se implementarán Modelos de clasificación basados en Supervisión



x

## EVALUACIÓN

Se calificará bajo distintas métricas la capacidad de predicción de los Modelos elegidos





01

# PREPARACIÓN

Aplicada sobre los datos de ejemplo

# PREPARACIÓN

×

×

De detalles  
irrelevantes

LIMPIEZA



SEPARACIÓN

De los textos  
individuales



ESTRUCTURACIÓN

Con base en  
estructuras de  
datos adaptadas



A un formato  
computable

ADAPTACIÓN



×



# SEPARACIÓN

1. Identificar los archivos del *dataset* con textos en inglés y separarlos conservando las etiquetas correspondientes a su autoría
2. Haciendo una exploración de cada archivo, identificar el formato en que se almacena el texto objetivo y separarlo
3. Se almacenará el texto individual anexo a un archivo *TXT* global que contendrá en cada línea cada uno de los textos
4. Se creará un archivo *TXT* para cada “*autor*”, cuyo nombre de archivo reflejará el nombre del modelo generador o bien de simplemente “*humano*”





# LIMPIEZA

1. Para cada archivo *TXT* generado, se almacenará en *RAM* y se someterá a una eliminación de detalles irrelevantes:
  1. Eliminación de caracteres especiales [\n, \t, \r, \&u..., ...]
  2. ~~Eliminación de símbolos de puntuación [, . ; ...]~~
  3. ~~Eliminación de *stopwords* ["a", "the", "is", "are", ...]~~



# ESTRUCTURACIÓN

x

1. Para cada archivo texto almacenado en *RAM* para el paso de *LIMPIEZA*, se almacenará en una estructura de datos que contenga la lista de textos junto a la etiqueta correspondiente a su autor
2. Se almacenarán todos los textos limpiados sobre un archivo tabular CSV que contenga en cada línea la etiqueta con el autor
3. ~~Se tokenizarán los textos para identificar vocabulario completo de cada autor~~



# ADAPTACIÓN

Alternativas para la adaptación de los textos en el *csv* completo:

1. **Embeddings:** Alternativas (tentativamente sólo uno):
  1. **CBOW**
  2. **Word2Vec**
  3. **Fine-Tuning con GoogleWord2Vec**
2. **N-Gramas:** Alternativas:
  1. **Trigramas de palabras**
  2. **Bigramas de frases**

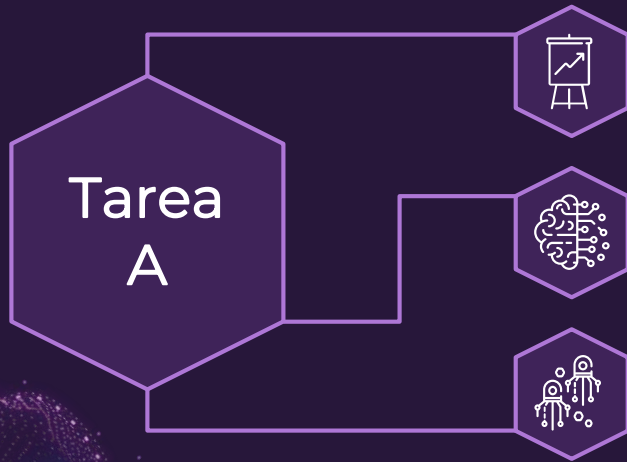


02

# MODELOS

Modelos de Clasificación de Aprendizaje Supervisado

# × CLASIFICACIÓN BINARIA ×



Regresión Logística

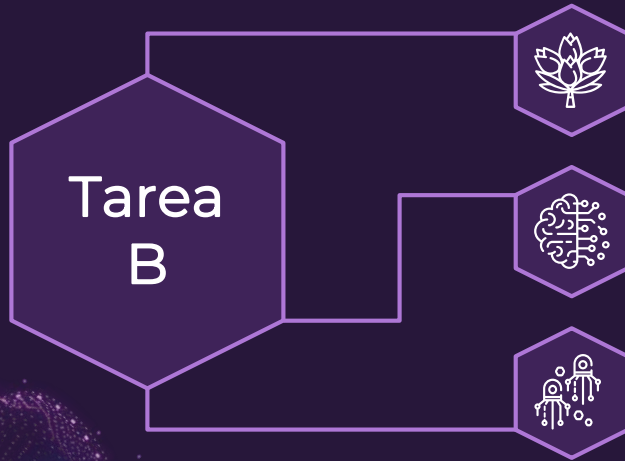
Red Neuronal Recurrente LSTM

Algoritmo XGBoost

×

# CLASIFICACIÓN MULTICLASE

x



Random Forest

Red Neuronal Recurrente LSTM

Algoritmo XGBoost

x



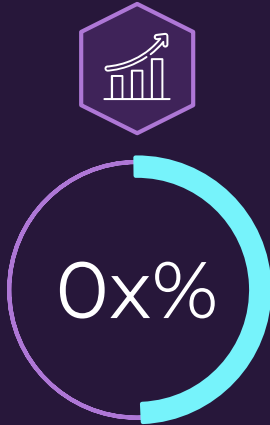
03

# EVALUACIÓN

Métricas de rendimiento de los modelos y métodos

# MÉTRICAS RELEVANTES

x



ACCURACY

Para cada clase (sin considerar desbalance en las etiquetas)



RECALL

Para cada clase



x



F1-Score

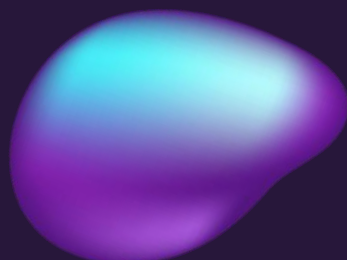
Para la clasificación general del Modelo

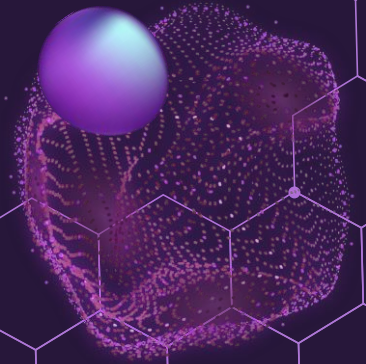
x



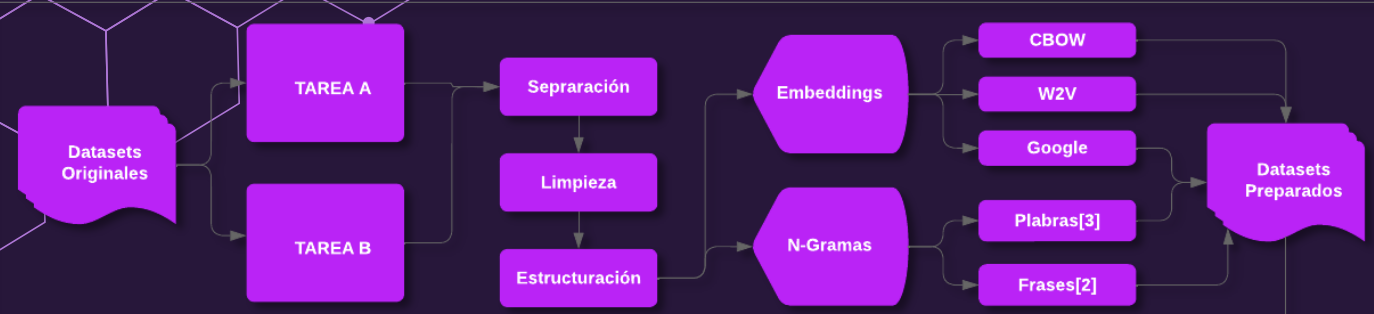


SE EVALUARÁN MÉTRICAS PARA:

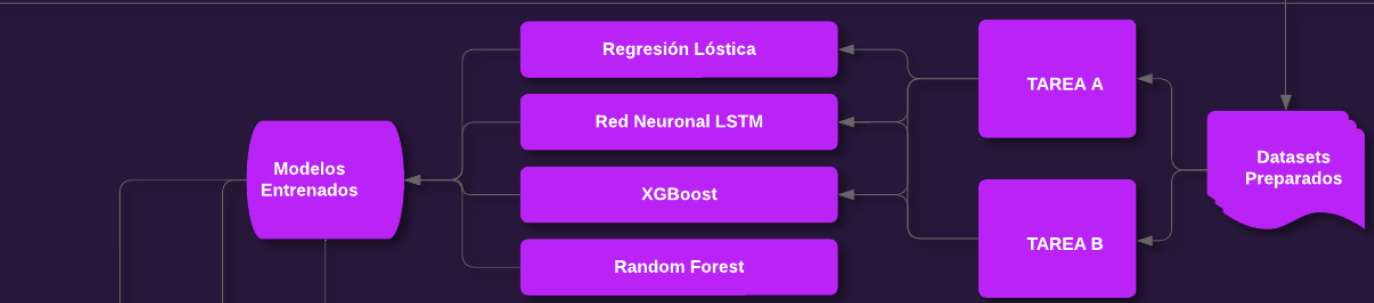
- \* CADA TAREA (2) =>
  - \* CADA PREPARACIÓN (2, 5) =>
  - \* CADA MODELO (3)
- 



PREPARACIÓN



MODELOS



EVALUACIÓN

