Uso de IA para la despersonalización de documentos

Juan Alejandro Herrera López

www.alejoherrera.com

https://www.linkedin.com/in/alejandro-herreracr/

Agenda

1. Entendemos nuestro derecho a la intimidad
2. Inteligencia artifical(aprendizaje profundo)
3. Evolución de los modelos procesamiento de lenguaje natural (llegada de los transformers)
4. Uso de LLM (grandes modelos de lenguaje natural) para la despersonalización

- Bert
- GPT

1. Tipos de errores
2. Algunas recomendaciones

"Una vida privada es una vida feliz" Anónimo

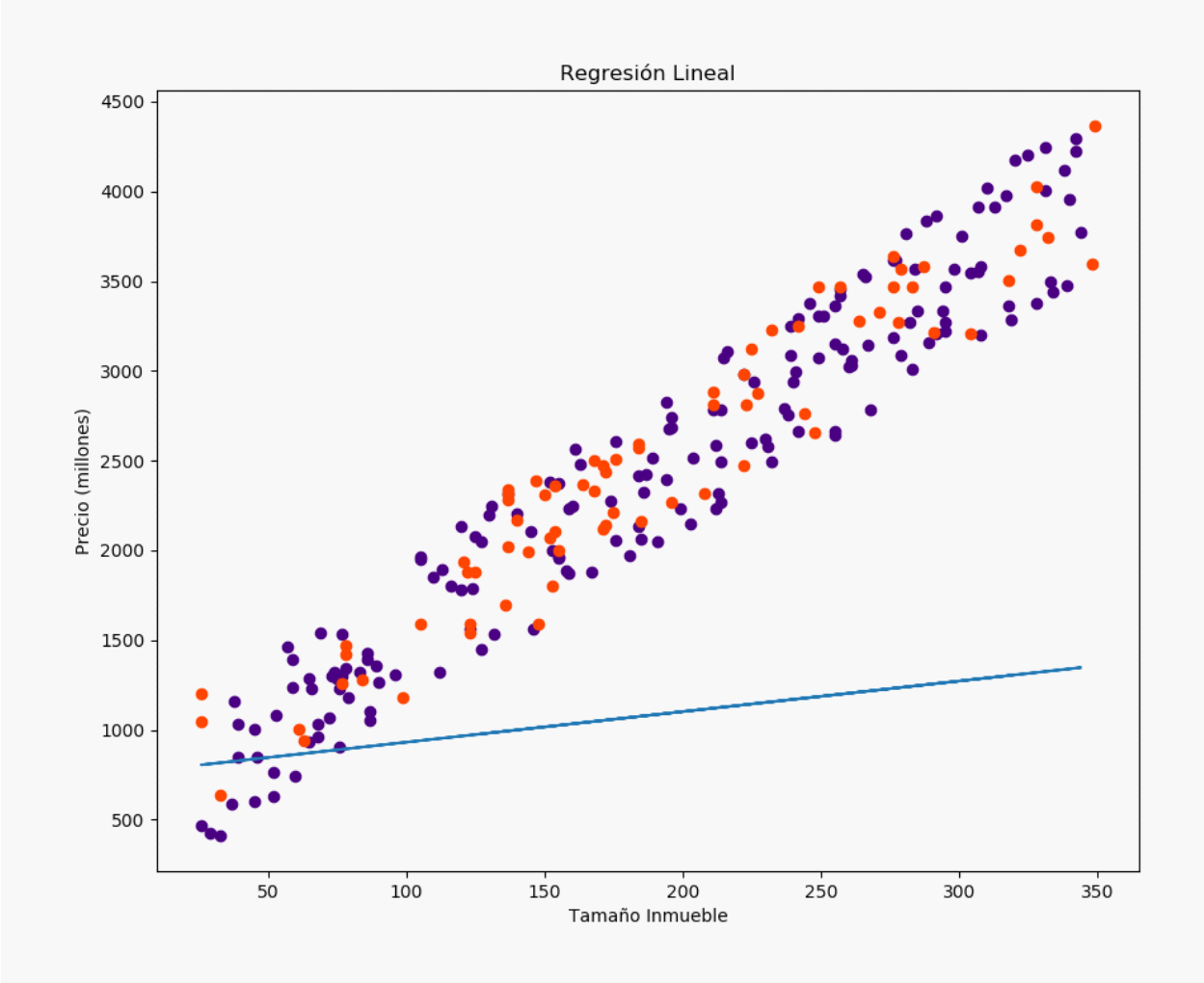
Entendemos las implicaciones de nuestro derecho a la intimidad

-No es un derecho aislado, se contrapone con un derecho de acceso a la información pública que es un pilar de nuestras democracias (No tenemos regulación legal de publicidad de la información pública)

-No es un derecho consolidado actualmente, y resulta complejo por el estado del arte tecnológico.

-Su marco de tratamiento está en creación, por lo que hay contradicciones en su aplicación (Caso Costeja)

#IA.Aprendizaje automático/Apendizaje profundo



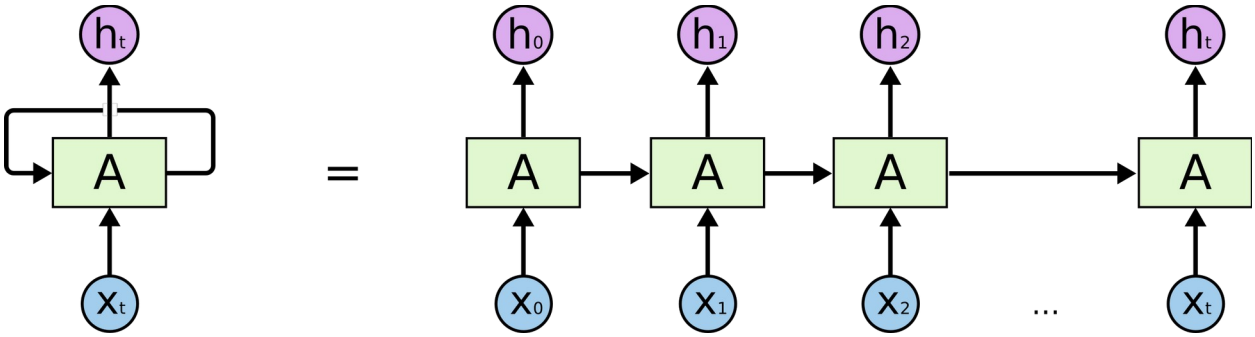
Playground

Evolución de los modelos de NLP(Llegada de los transformers)

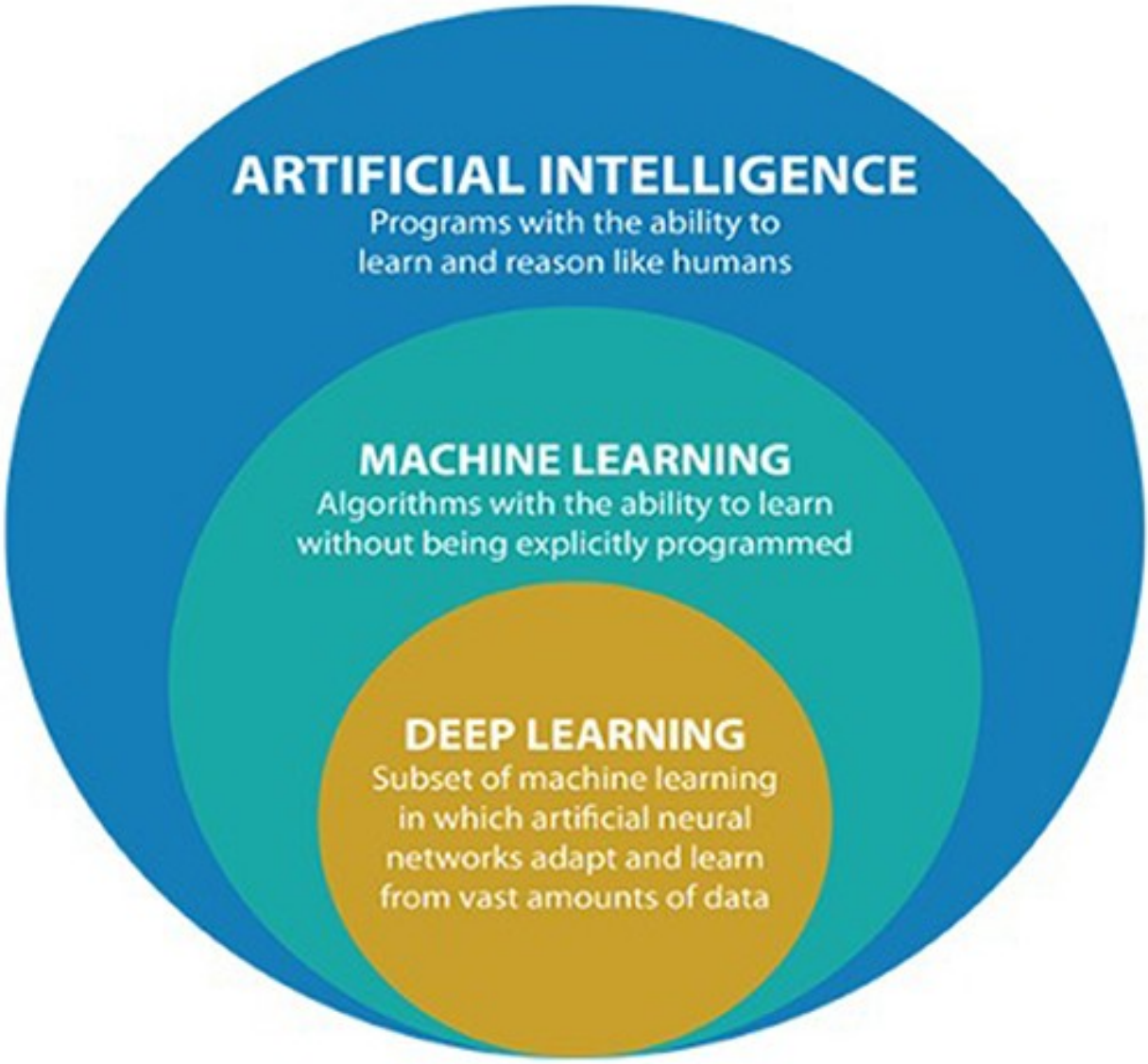
Al principio las redes procesaban el texto como los seres humanos, de izquierda a derecha -si era español- y de forma secuencial...esta forma de procesamiento venía con ciertos problemas importantes:

-Tendencia al olvido de las primeras palabra o problemas con oraciones muy largas. -Se perdía el contexto. -El entrenamiento era complejo pues no era factible mecanismos de paralelización.

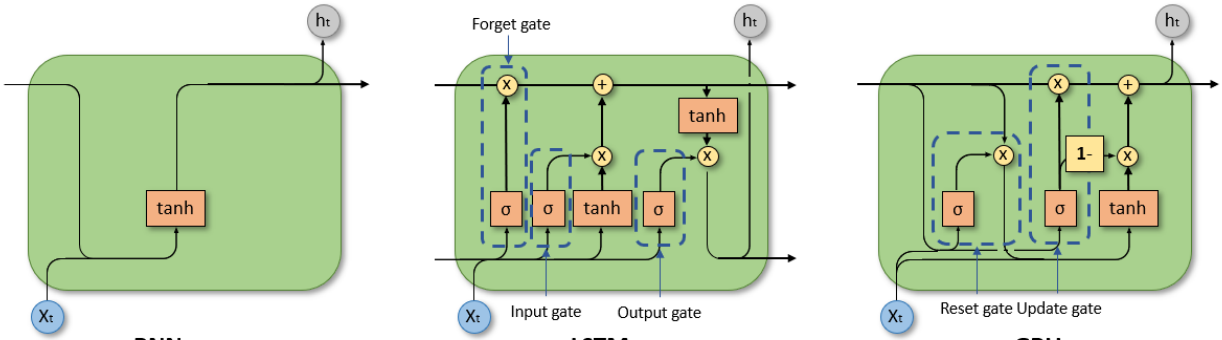
Redes neuronales recurrentes



Con el tiempo se trato de solucionar estos problemas, incorporando en las redes mecanismos de atención que evitan el olvido o disminuyen la pérdida de contexto, sin embargo mantienen un alto diseño secuencial.



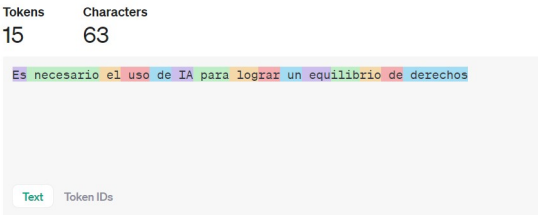
LSTM (Long short term memory)



RNN LSTM GRU

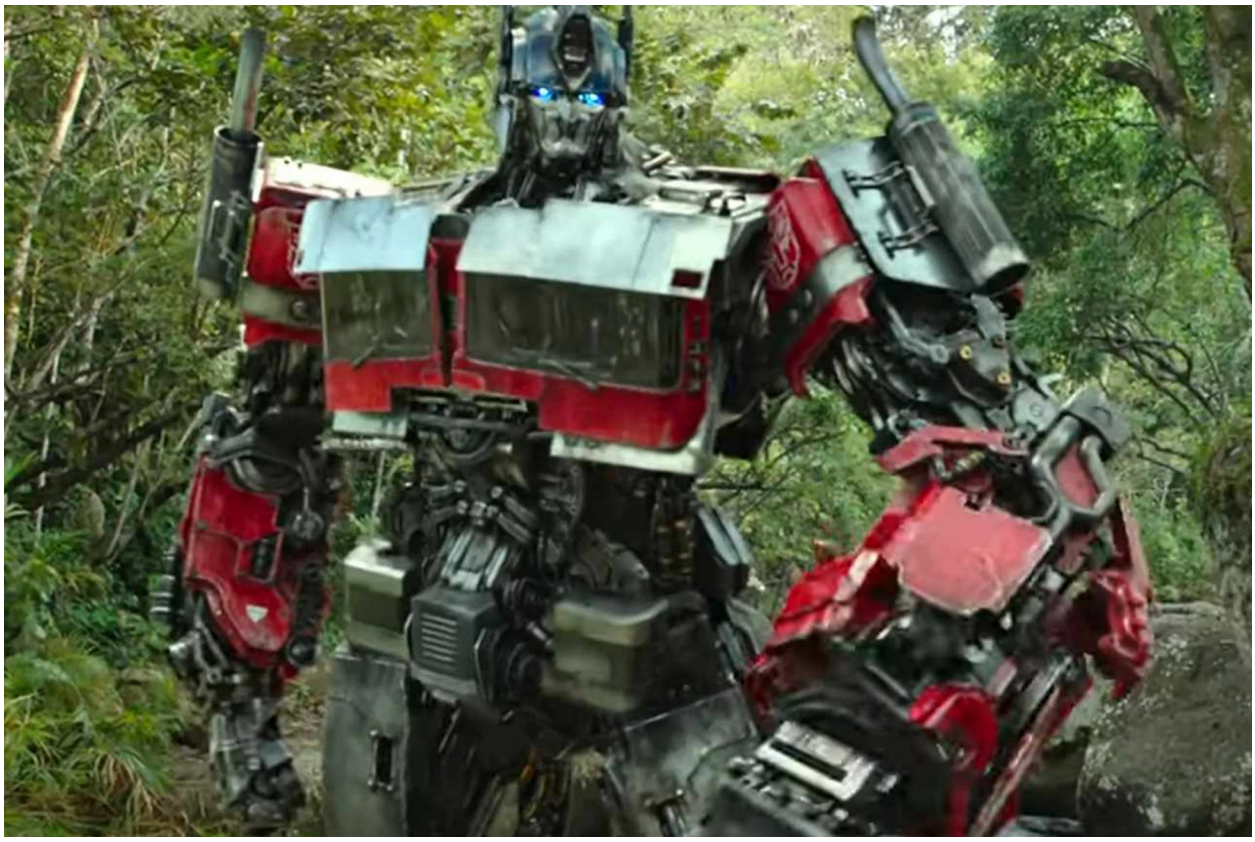
Ejemplo

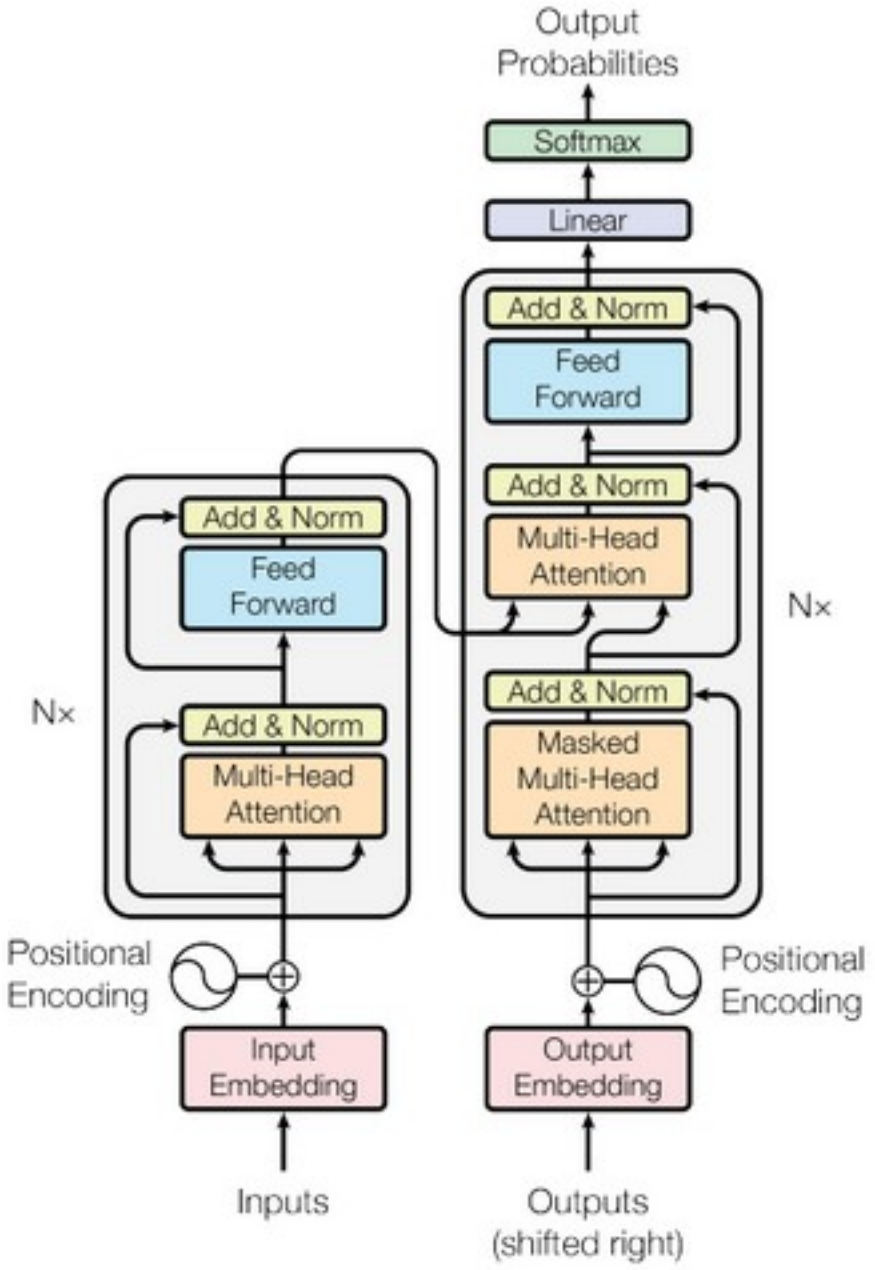
"Es necesario el uso de IA para lograr un equilibrio de derechos."



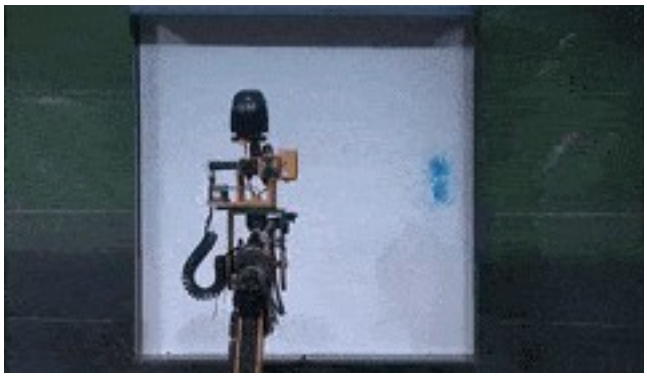
En el 2017 investigadores de Google junto con la Universidad de Toronto presentan una propuesta de arquitectura de redes, en un paper llamado "Atención es todo lo que necesitas" , generando una explosión en la generación de grandes modelos de lenguaje ya que permitía la paralelización del entrenamiento de forma plena, y reforzó los mecanismos de atención...hoy los grandes modelos de procesamiento de lenguaje, incluyendo GPT se basan en está arquitectura.

Llegada de los Transformers

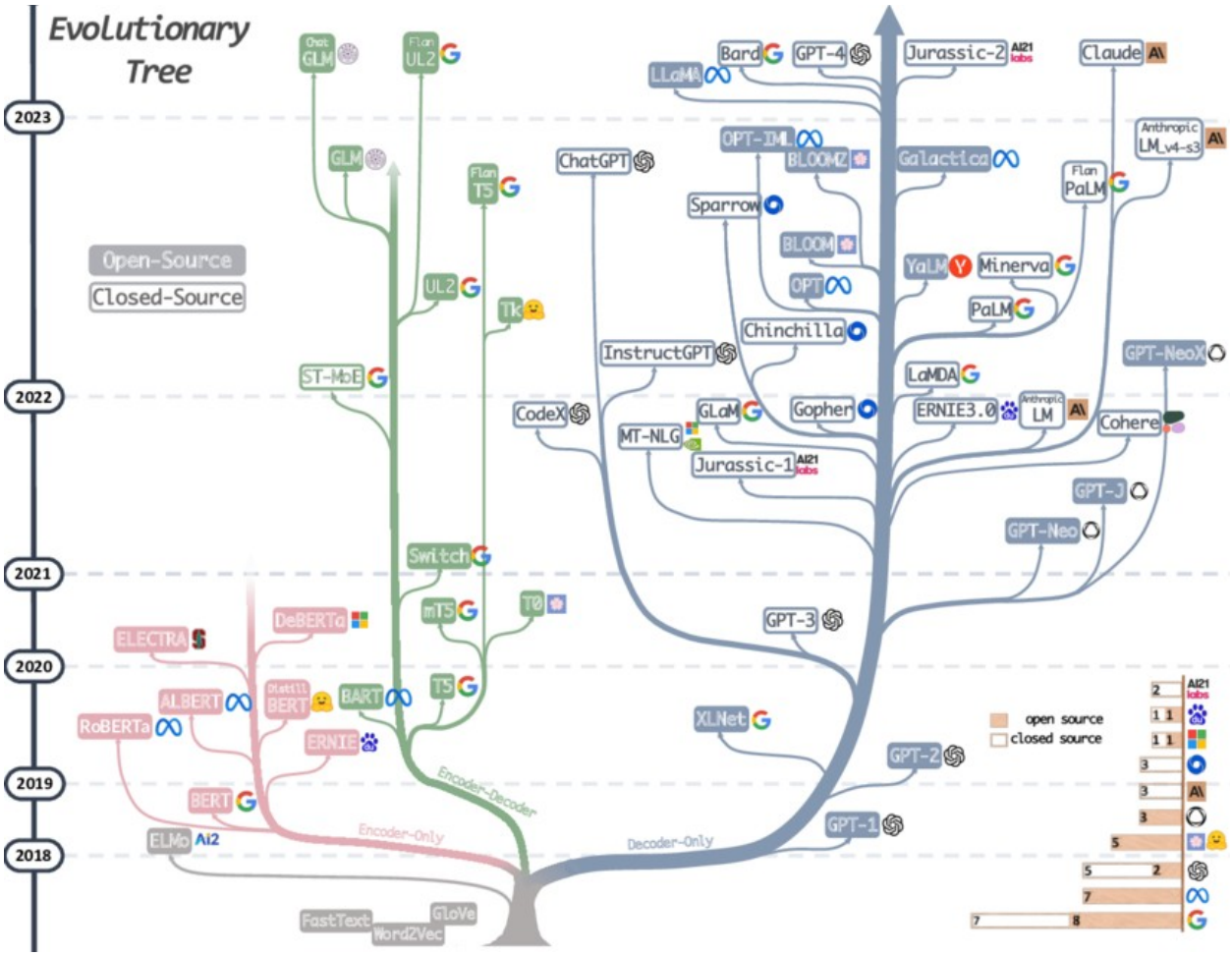




Diferencia entre procesamiento recurrente y paralelización....



A partir de este paper...



Uso de LLM para la despersonalización

Bert

Cómo podríamos incorporar los grandes modelos de lenguaje (LLM en inglés) en el proceso de fiscalización, algunos tips....

```
%%capture
!pip install transformers;

%%capture
from transformers import pipeline

nlp_ner = pipeline(
    "ner",
    model="mrm8488/bert-spanish-cased-finetuned-ner",
    tokenizer=(
        'mrm8488/bert-spanish-cased-finetuned-ner',
```

```
        {"use_fast": False}
))
```

Detección de entidades

```
from IPython import display
from tabulate import tabulate
text = 'Pedro Pérez fue denunciado ante el Ministerio Público por corrupción'
display.HTML(tabulate(nlp_ner(text), tablefmt='html',headers="keys"))
```

```
<IPython.core.display.HTML object>
```

```
from IPython import display
from tabulate import tabulate
text = 'El 5 de marzo de 2010, el Sr. Costeja González, de nacionalidad española y domiciliado en España, presentó ante la AEPD una reclamación contra La Vanguardia Ediciones, S.L., que publica un periódico de gran difusión, concretamente en Cataluña (en lo sucesivo, «La Vanguardia»), y contra Google Spain y Google Inc. Esta reclamación se basaba en que, cuando un internauta introducía el nombre del Sr. Costeja González en el motor de búsqueda de Google (en lo sucesivo, «Google Search»), obtenía como resultado vínculos hacia dos páginas del periódico La Vanguardia, del 19 de enero y del 9 de marzo de 1998, respectivamente, en las que figuraba un anuncio de una subasta de inmuebles relacionada con un embargo por deudas a la Seguridad Social, que mencionaba el nombre del Sr. Costeja González'
display.HTML(tabulate(nlp_ner(text), tablefmt='html',headers="keys"))
```

```
<IPython.core.display.HTML object>
```

Clasificación de texto

```
%%capture
from transformers import pipeline
classifier = pipeline('zero-shot-classification', model='roberta-large-mnli')

sequence_to_classify = "Los ingresos del señor Soto se han visto afectados por los años en los cuales no tuvo empleo"
candidate_labels = ['salario', 'edad']

resultado=classifier(sequence_to_classify, candidate_labels)
resultado =list(map(resultado.get, ["labels","scores"]))


display.HTML(tabulate(resultado,
tablefmt='html',headers=["clase1","clase2"]))
```

```
<IPython.core.display.HTML object>
```

```
sequence_to_classify = "El señor Soto tiene 60 años, lo que le
dificulta encontrar un trabajo bien remunerado."
candidate_labels = ['salario', 'edad']

resultado=classifier(sequence_to_classify, candidate_labels)
resultado =list(map(resultado.get, ["labels","scores"]))


display.HTML(tabulate(resultado,
tablefmt='html',headers=["clase1","clase2"]))
```

```
<IPython.core.display.HTML object>
```

```
sequence_to_classify = "El costo de la  licitación es irrazonable"
candidate_labels = ['tiempo', 'precio']
resultado=classifier(sequence_to_classify, candidate_labels)
resultado =list(map(resultado.get, ["labels","scores"]))


display.HTML(tabulate(resultado,
tablefmt='html',headers=["clase1","clase2"]))
```

```
<IPython.core.display.HTML object>
```

Análisis de sentimientos

```
%%capture
from transformers import pipeline
classifier = pipeline("sentiment-analysis")

resultado=classifier("Me encuentro frustrado por la forma como esa
institución maneja mi datos personales")
display.HTML(tabulate(resultado, tablefmt='html',headers="keys"))
```

```
<IPython.core.display.HTML object>
```

```
resultado=classifier("Las condiciones de manejo de datos personales en
el país han mejorado bastante con respecto al pasado.")
display.HTML(tabulate(resultado, tablefmt='html',headers="keys"))
```

```
<IPython.core.display.HTML object>
```

GPT

```
%%capture
!pip install openai==0.28
!pip install typing_extensions
!jupyter notebook restart

import openai
```

```
openai.api_key = "**************************"


def get_completion(prompt, model="gpt-3.5-turbo"):
    messages = [{"role": "user", "content": prompt}]
    response = openai.ChatCompletion.create(
        model=model,
        messages=messages,
        temperature=0, # this is the degree of randomness of the
model's output
    )
    return response.choices[0].message["content"]


text_principal = f"""
*El 5 de marzo de 2010, el Sr. Costeja González, de nacionalidad
española y domiciliado en España,
presentó ante la AEPD una reclamación contra La Vanguardia Ediciones,
S.L., que publica un periódico de gran difusión,
concretamente en Cataluña (en lo sucesivo, «La Vanguardia»), y contra
Google Spain y Google Inc. Esta reclamación se basaba en que,
cuando un internauta introducía el nombre del Sr. Costeja González en
el motor de búsqueda de Google (en lo sucesivo, «Google Search»),
 obtenía como resultado vínculos hacia dos páginas del periódico La
Vanguardia, del 19 de enero y del 9 de marzo de 1998, respectivamente,
  en las que figuraba un anuncio de una subasta de inmuebles
relacionada con un embargo por deudas a la Seguridad Social, que
mencionaba el nombre del Sr. Costeja González.
  Mediante esta reclamación, el Sr. Costeja González solicitaba, por
un lado, que se exigiese a La Vanguardia eliminar o modificar la
publicación para que no apareciesen
  sus datos personales, o utilizar las herramientas facilitadas por
los motores de búsqueda para proteger estos datos. Por otro lado,
solicitaba que se exigiese a Google
   Spain o a Google Inc. que eliminaran u ocultaran sus datos
personales para que dejaran de incluirse en sus resultados de búsqueda
y dejaran de estar ligados a
    los enlaces de La Vanguardia. En este marco, el Sr. Costeja
González afirmaba que el embargo al que se vio sometido en su día
estaba totalmente solucionado y
  resuelto desde hace años y carecía de relevancia actualmente.*
"""

prompt = f"""
Resume lo que está dentro de un asterisco * en una sola oración no
mayor a 15 palabras.
```{text_principal}```
"""
```
```

```
response = get_completion(prompt)
print(response)
```

```
El Sr. Costeja González presentó una reclamación contra La Vanguardia
y Google por la publicación de datos personales.
```

```
prompt = f"""
Del texto anterior  haz una lista de las personas ubicadas\
Si no encuentras ninguna, \
Escribe \"No encontradas.\"
\"\"\"{text_principal}\"\"\"
"""
response = get_completion(prompt)
print("Completion for text:")
print(response)
```

```
Completion for text:
- Sr. Costeja González
```

```
prompt = f"""
Del texto anterior * encuentra a personas, y sustituye en el texto los
nombres o apellidos de las personas por tres puntos suspensivos ... \
Si no encuentras ninguna, \
Escribe \"No encontradas.\"
\"\"\"{text_principal}\"\"\"
"""
response = get_completion(prompt)
print(response)
```

```
*El 5 de marzo de 2010, el ... ..., de nacionalidad española y
domiciliado en España,
presentó ante la AEPD una reclamación contra La Vanguardia Ediciones,
S.L., que publica un periódico de gran difusión,
concretamente en Cataluña (en lo sucesivo, «La Vanguardia»), y contra
Google Spain y Google Inc. Esta reclamación se basaba en que,
cuando un internauta introducía el nombre del ... ... en el motor de
búsqueda de Google (en lo sucesivo, «Google Search»),
 obtenía como resultado vínculos hacia dos páginas del periódico La
Vanguardia, del 19 de enero y del 9 de marzo de 1998, respectivamente,
  en las que figuraba un anuncio de una subasta de inmuebles
relacionada con un embargo por deudas a la Seguridad Social, que
mencionaba el nombre del ... ....
  Mediante esta reclamación, el ... ... solicitaba, por un lado, que
se exigiese a La Vanguardia eliminar o modificar la publicación para
que no apareciesen
  sus datos personales, o utilizar las herramientas facilitadas por
los motores de búsqueda para proteger estos datos. Por otro lado,
solicitaba que se exigiese a Google
   Spain o a Google Inc. que eliminaran u ocultaran sus datos
```

```
personales para que dejaran de incluirse en sus resultados de búsqueda y dejaran de estar ligados a
    los enlaces de La Vanguardia. En este marco, el ... ... afirmaba que el embargo al que se vio sometido en su día estaba totalmente solucionado y
  resuelto desde hace años y carecía de relevancia actualmente.*


prompt_1 = f"""
Toma el texto que se encuentra dentro de un asterico * :
1 - Detecta las personas.
2 - Genera un json con lo siguiente \
keys: resumen_sentencias, verbos_encontrado.

Genera las respuestas separadas por dos asteriscos * cada una.

Text:
```{text_principal}```
"""
response = get_completion(prompt_1)
print("Tareas requeridas:")
print(response)

Tareas requeridas:
Respuesta 1:
El Sr. Costeja González

Respuesta 2:
{
  "resumen_sentencias": "El 5 de marzo de 2010, el Sr. Costeja González, de nacionalidad española y domiciliado en España, presentó ante la AEPD una reclamación contra La Vanguardia Ediciones, S.L., que publica un periódico de gran difusión, concretamente en Cataluña (en lo sucesivo, «La Vanguardia»), y contra Google Spain y Google Inc. Esta reclamación se basaba en que, cuando un internauta introducía el nombre del Sr. Costeja González en el motor de búsqueda de Google (en lo sucesivo, «Google Search»), obtenía como resultado vínculos hacia dos páginas del periódico La Vanguardia, del 19 de enero y del 9 de marzo de 1998, respectivamente, en las que figuraba un anuncio de una subasta de inmuebles relacionada con un embargo por deudas a la Seguridad Social, que mencionaba el nombre del Sr. Costeja González. Mediante esta reclamación, el Sr. Costeja González solicitaba, por un lado, que se exigiese a La Vanguardia eliminar o modificar la publicación para que no apareciesen sus datos personales, o utilizar las herramientas facilitadas por los motores de búsqueda para proteger estos datos. Por otro lado, solicitaba que se exigiese a Google Spain o a Google Inc. que eliminaran u ocultaran sus datos personales para que dejaran de incluirse en sus resultados de búsqueda y dejaran de estar ligados a los enlaces de La Vanguardia. En este marco, el Sr. Costeja González afirmaba que el embargo al que se vio sometido en su
```
```

```
día estaba totalmente solucionado y resuelto desde hace años y carecía
de relevancia actualmente.",
  "verbos_encontrados": "detecta, genera"
}


!pip install -q openai
!pip install -q elevenlabs

import os
import openai
import tempfile
import requests
from IPython.display import Audio, clear_output
from elevenlabs import generate, play, set_api_key, voices, Models

#@title
openai_api_key     = "****************************************"
#@param {type:"string"}
eleven_api_key = "****************************************" #@param
{type:"string"}

# Configure GPT-4 and Text-to-speech API keys
openai.api_key = openai_api_key
set_api_key(eleven_api_key)

voice_list = voices()

━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 71.9/71.9 kB 3.3 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.0/1.0 MB 23.2 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 114.5/114.5 kB 10.9 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 268.8/268.8 kB 22.7 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 149.6/149.6 kB 12.0 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.6/1.6 MB 22.8 MB/s eta
0:00:00


#@title
import ipywidgets as widgets

voice_labels = [voice.category + " voice: " + voice.name for voice in
voice_list]

voice_id_dropdown = widgets.Dropdown(
    options=voice_labels,
    value=voice_labels[0],
```

```
    description="Selecciona una voz:",
)

display(voice_id_dropdown)
```

{"model_id":"105045e044944d8195681091d9642758","version_major":2,"version_minor":0}

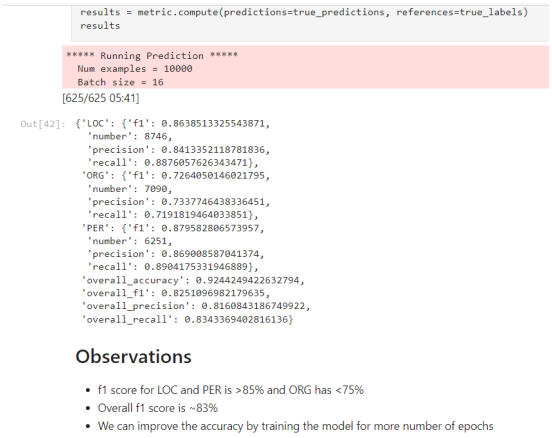
Tipos de errores

	Predicción negativo	Predicción positiva
Casos negativos	TN: 9760	FP: 140
Casos positivos	FN: 40	TP: 60

1. **TN / True Negative:** Caso cuando es negativo y es predecido negativo.
2. **TP / True Positive:** Caso cuando es positivo y es predecido positivo.
3. **FN / False Negative:** Caso cuando es positivo, pero es predecido negativo.
4. **FP / False Positive:** Caso cuando es negativo, pero es predecido positivo.

Error Tipo I : Falso positivo, eliminó información de interés público

Error Tipo II: Falso negativo, no eliminó el dato personal



```
results = metric.compute(predictions=true_predictions, references=true_labels)
results
```

```
***** Running Prediction *****
  Num examples = 10000
  Batch size = 16
[625/625 05:41]
```

```
Out[42]: {'LOC': {'f1': 0.8638513325543871,
  'number': 8746,
  'precision': 0.8413352118781836,
  'recall': 0.8876057626343471},
 'ORG': {'f1': 0.7264050146021795,
  'number': 7090,
  'precision': 0.7337746438336451,
  'recall': 0.7191819464033851},
 'PER': {'f1': 0.879582806573957,
  'number': 6251,
  'precision': 0.869008587041374,
  'recall': 0.8904175331946889},
 'overall_accuracy': 0.9244249422632794,
 'overall_f1': 0.8251096982179635,
 'overall_precision': 0.8160843186749922,
 'overall_recall': 0.8343369402816136}
```

Observations

- f1 score for LOC and PER is >85% and ORG has <75%
- Overall f1 score is ~83%
- We can improve the accuracy by training the model for more number of epochs

Algunas recomendaciones:

- A los órganos de control:

  Es importante valorar en los protocolos de actuación, no sólo la robustez de los controles y mecanismos de despersonalización, sino la velocidad y eficiencia con qué

los actores disponen de la información de interés público despersonalizada a la sociedad.

En los casos de uso de mecanismos de IA, se debe sopesar que es un mundo probabilístico y no deterministico, esto es, no se puede asegurar un 100% de efectividad (los proceso humanos tampoco son 100% fiables pero generalmente no se evaluan). Por lo tanto para la aplicación de la gravedad de un error, se debe sopesar no sólo el error concreto, sino el proceso de verificación y control de entrenamiento de los algoritmos o modelos.

- A los generadores de documentos que deben ser despersonalizados

La IA puede ser una parte de la solución, pero no es la única respuesta, se debe valorar la arquitectura de los documentos, entre más desperdigados y poco estándarizado se encuentren los datos personales, más probable van a ser los errores y los costos computacionales de correr los modelos.

Formatos como el PDF son poco amigables para procesos de despersonalización automatizadas.

#Muchas gracias por su atención...