Trabajo de Fin de Máster

Comparación de técnicas de optimización de modelos aplicadas al análisis de sentimientos.

Autor: Martín Iglesias Nalerio.

Tutor: Antonio García Cabot.

Cotutor: Pablo Trull Báguena.

Analítica de Negocio y Big Data.

Escuela Politécnica Superior.

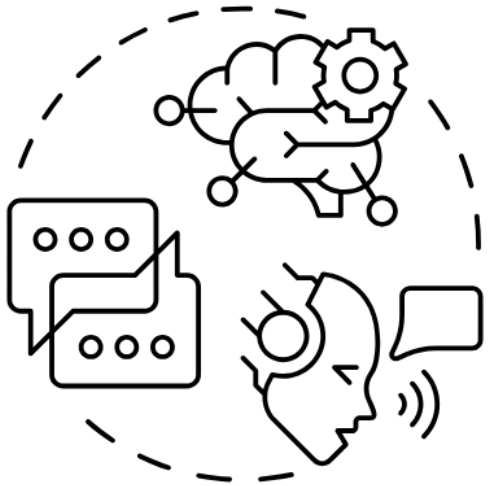
Curso Académico 2024/2025.

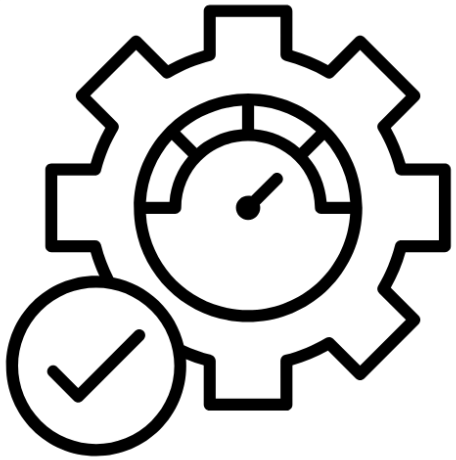
Índice

- Introducción y objetivos del proyecto.
- Estudio teórico de las técnicas.
- Desarrollo e implementación práctica.
- Resultados y conclusiones finales.
- Líneas de trabajo futuras.

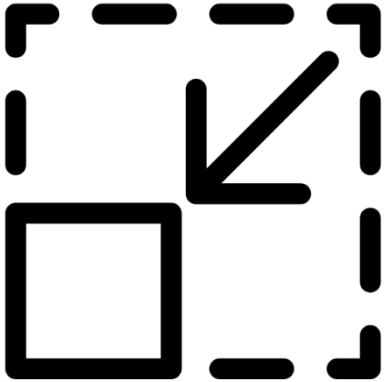
Introducción y objetivos del proyecto.

- La Inteligencia Artificial ha llegado para quedarse.
- Modelos grandes, con limitaciones de despliegue,
- Existen técnicas que optimizan estos modelos.
- ¿Objetivo del proyecto?
  - Modelo base que permita analizar el sentimiento de distintas opiniones.
  - Optimizar ese modelo base con varias técnicas.
  - Comparar sus resultados y obtener conclusiones.



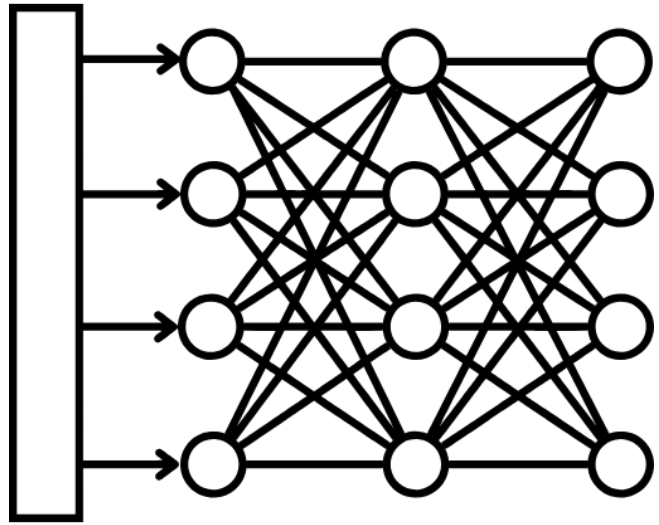


Estudio teórico: Cuantización.



- Propósito: Reducir la precisión del tipo de datos de una red neuronal, tratando de no perder rendimiento.

- Clasificación:
  - Según lo que cuantizan: Principalmente pesos de la red.
  - Según cuándo se realiza: Post Training VS Aware Training.
    - Dentro de Post Training destaca Dynamic Range Quantization.
    - Dentro de Aware Training destaca QAT con cuantizaciones falsas.



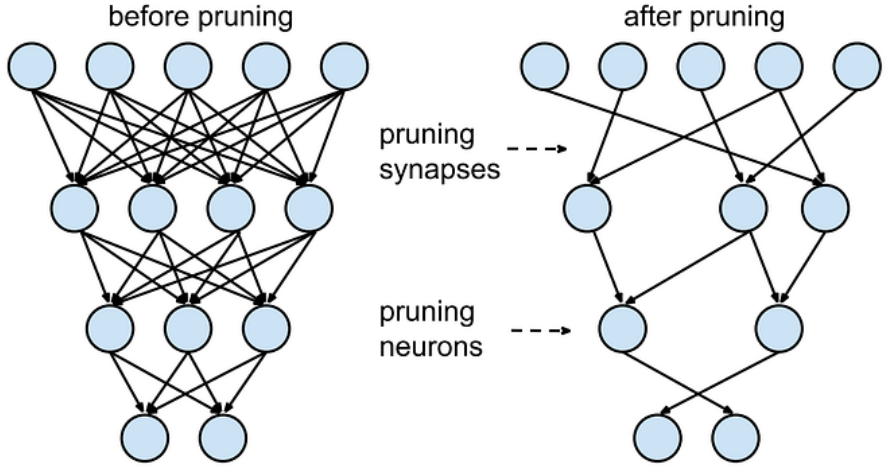
- Equilibrio a la hora de elegir el tipo de datos, según tarea a realizar.

Estudio teórico: Podado.

- Propósito: Eliminar partes de la red neuronal cuyo aporte al rendimiento final de la misma es muy reducido, simplificando su estructura sin perder rendimiento.

- Clasificación:
  - Según lo que se poda: Principalmente pesos o activaciones.
  - Según cómo se poda: No Estructurado VS Estructurado.

- Según cómo se determina si se contribuye al rendimiento:
  - Basado en Magnitudes, Gradientes, % de ceros…etc.

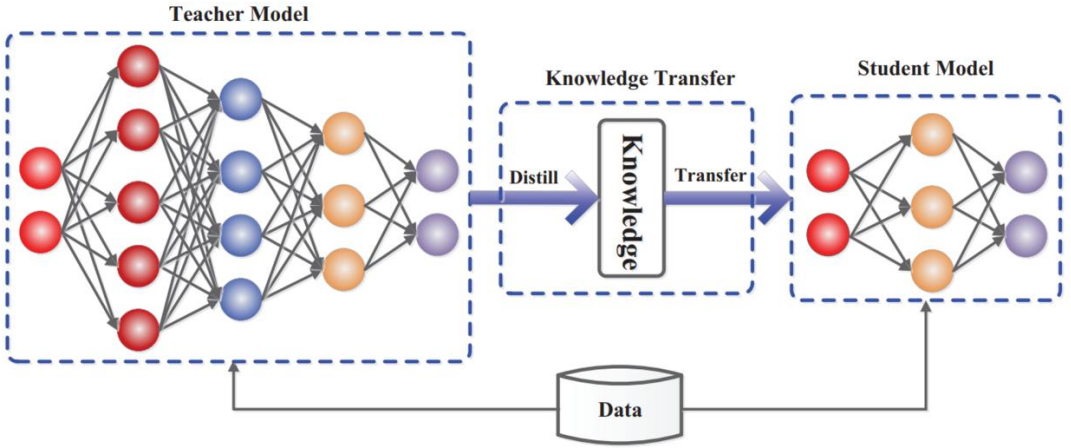


- Al contar los parámetros de la red podada, es probable que salgan los mismos que en la red original - - > Máscara.

Estudio teórico: Destilación de Conocimiento.

- Propósito: Dado un modelo de grandes capacidades y muy buen rendimiento, crear otro modelo reducido y que aprenda del modelo original.

- Clasificación:
  - Según lo que puede aprender: Respuestas, Features o relaciones.
  - Según cómo se entrene los modelos: Offline VS Online VS Self.

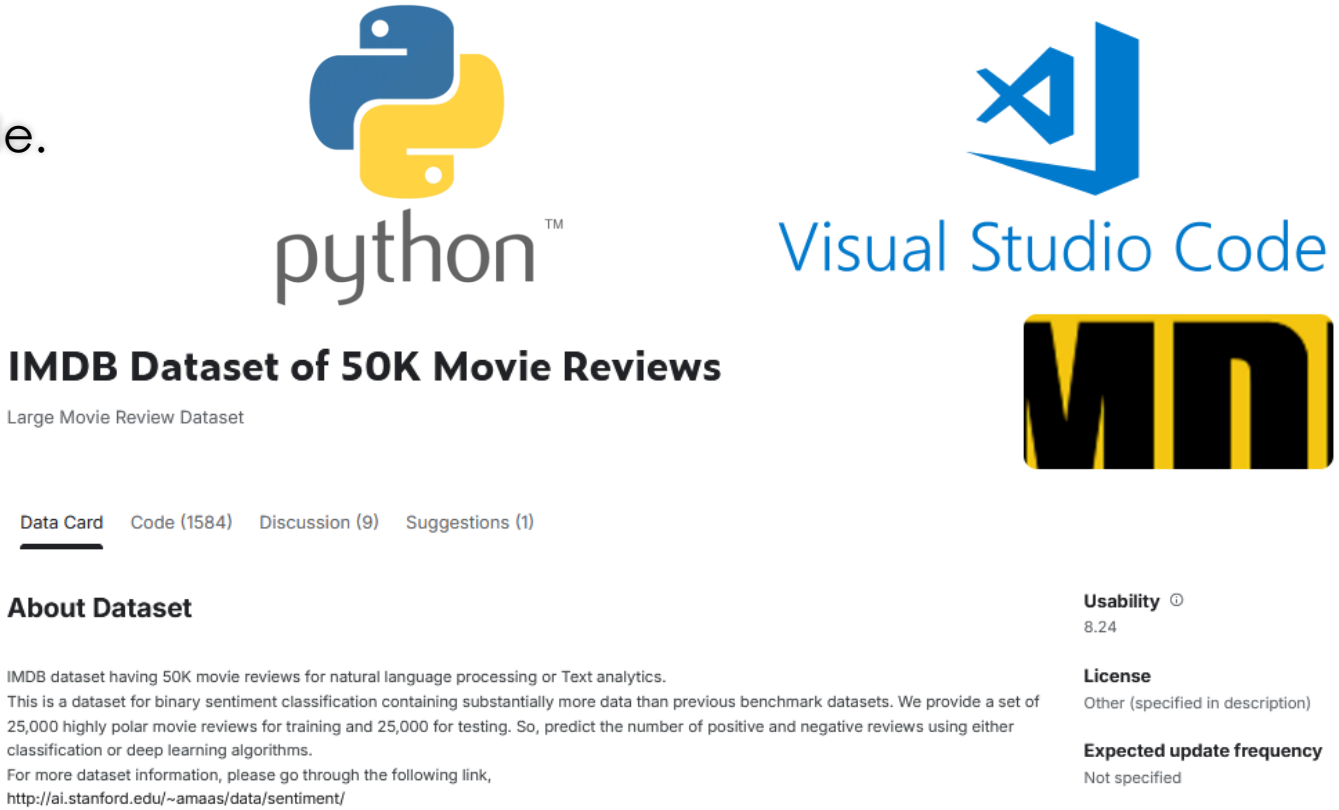


- No se garantiza que el estudiante mantenga la calidad del maestro, pero sí que puede aprender todos los sesgos negativos que haya aprendido el maestro.

Desarrollo e implementación práctica.

- Realizado en Python 3.9 y Visual Studio Code.

- Proyecto estructurado en cinco scripts:
  - 1-Data Collection and Preprocessing.py
  - 2-LSTM Model.py
  - 3-Quantization.py
  - 4-Pruning.py
  - 5-Knowledge Distillation.py



- Dataset: IMDB Reviews, con datos de 50000 críticas a películas y su sentimiento positivo o negativo.

EDA: Análisis de Datos Exploratorio.

```
A BIT OF EDA
CUANTOS NULOS
review       0
sentiment    0
dtype: int64
DESCRIPTIVO DATOS
                                                  review sentiment
count                                              50000     50000
unique                                             49579         2
top     loved today's show variety solely cooking whic...  positive
freq                                                   5     25000
```

```
PALABRAS EN REVIEW POR CADA SENTIMENT
                mean  min   max
sentiment
negative   125.08472    3   875
positive   127.81848    6  1465
```

```
CARACTERES EN REVIEW POR CADA SENTIMENT
                 mean  min   max
sentiment
negative   835.48688   18  5951
positive   869.16684   37  9394
```

```
CARGAR DATOS
                                              review sentiment
0  One of the other reviewers has mentioned that ...  positive
1  A wonderful little production. <br /><br />The...  positive
2  I thought this was a wonderful way to spend ti...  positive
3  Basically there's a family where a little boy ...  negative
4  Petter Mattei's "Love in the Time of Money" is...  positive
```

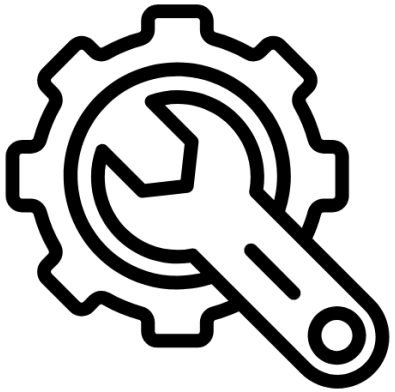
```
NORMALIZAR DATOS
                                              review sentiment
0  one reviewers mentioned watching 1 oz episode ...  positive
1  wonderful little production the filming techni...  positive
2  thought wonderful way spend time hot summer we...  positive
3  basically there's family little boy jake think...  negative
4  petter mattei's love time money visually stunn...  positive
```

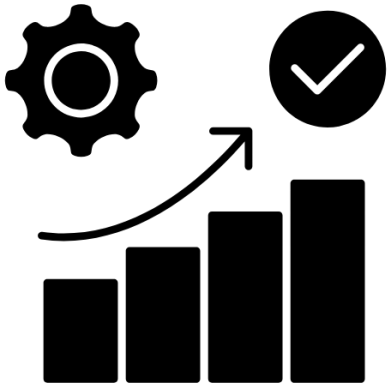
Resultados obtenidos.

Optimización	Precisión máxima	Tamaño en disco	Parámetros entrenables	Tiempo de entrenamiento	Complejidad de Implementación
Modelo base	86.43%	31578 Kilobytes	Aprox. 2.7 millones	30 minutos	Base
Dynamic Range Quantization	86.57%	2644 Kilobytes	Aprox. 2.7 millones	6-8 minutos	Media
FLOAT16 Quantization	86.43%	5271 Kilobytes	Aprox. 2.7 millones	4-6 minutos	Media
Magnitude Based Pruning	86.4%	10535 Kilobytes	80% del total, 543847 parámetros	12-15 minutos	Sencilla
Response-Based Offline Knowledge Distillation	88.4%	5149 Kilobytes	50% del total, 1.31 millones	25 minutos	Alta

Conclusiones obtenidas.

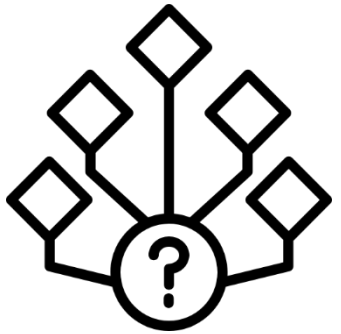
1. Con un buen ajuste, la pérdida de rendimiento es marginal.
2. No se trata de ganar rendimiento, sino de reducir otros aspectos y no perderlo.
3. Cuantización y Destilación de Conocimiento destacan, para beneficios diferentes.
4. Procesos de entrenamiento e inferencia más rápidos.
5. Optimización, en general, es requerida, y es utilizada en productos actuales.





Principales líneas de trabajo futuras.

1. Explorar formas alternativas de implementar las técnicas de optimización utilizadas en el proyecto.



2. Implementar nuevas técnicas de optimización: LOw Rank Adaptation.

3. Crear y evaluar rendimiento de modelos optimizados con multiples técnicas de optimización.