DistilBert

La primera arquitectura describe un sistema de clasificación de secuencias basado en el modelo distilbert-base-uncased para procesar texto, con un enfoque particular en la modificación de texto mediante la técnica de enmascaramiento selectivo. La idea principal detrás del enmascaramiento es ocultar deliberadamente ciertas palabras que podrían ser indicativas de sesgos o temas específicos, como género o roles sociales, para luego permitir que el modelo maneje la entrada censurada de manera que la clasificación final no se vea sesgada o influida indebidamente por estas. Por último, el texto enmascarado se tokeniza utilizando el tokenizer propio del modelo.

En este enfoque se ha utilizado el modelo preentrenado y se ha continuado su entrenamiento con los datos proporcionados para especializar su funcionamiento. Gracias a la clase AutoModel de HuggingFace permite adaptar el modelo para la tarea de clasificación.

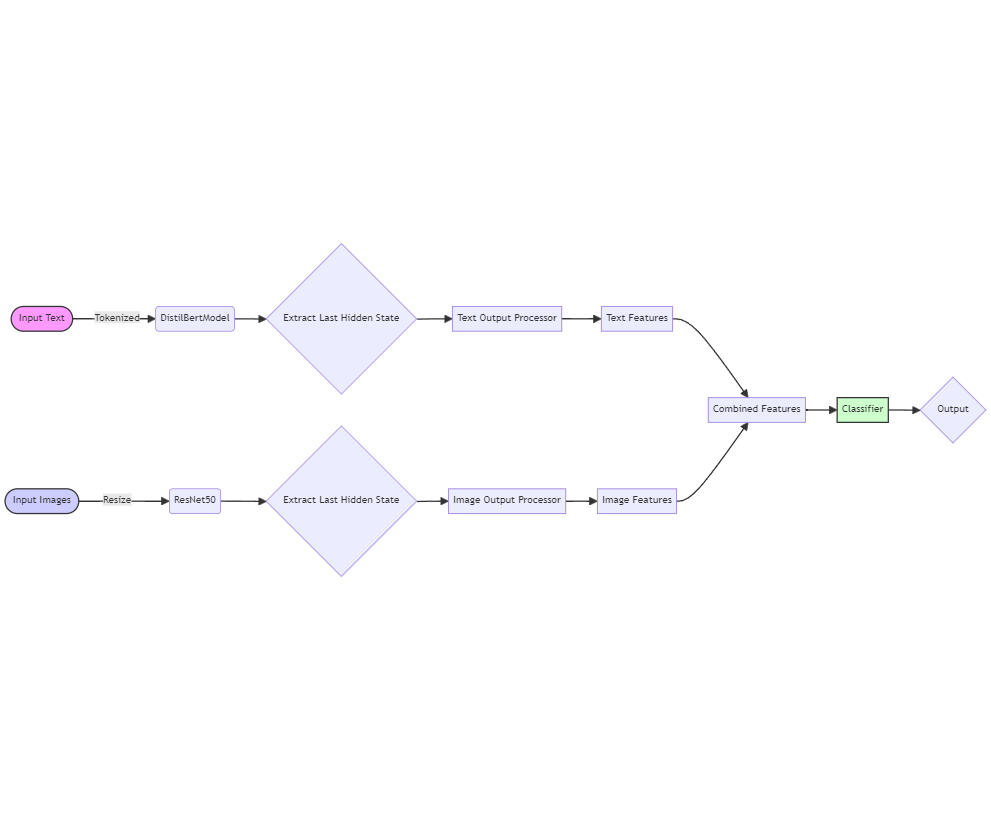
ResNet

Como segunda aproximación se ha elegido utilizar la arquitectura ResNet, adaptada para la clasificación de memes mediante técnicas de transfer learning. Esto se ha logrado utilizando un modelo pre-entrenado (ResNet-50) y al que se le ha sustituido el bloque de capas densas por un nuevo bloque conformado por capas densas específicamente dedicadas a la tarea de clasificación memes. Los pesos de la red original se han congelado de forma que solo se entrene el nuevo conjunto de pesos especializados. Este método es efectivo debido a que combina el conocimiento visual aprendido por el modelo con un entrenamiento focalizado en las características específicas de los memes sexistas, permitiendo un entrenamiento eficiente enfocado a la tarea.

Pseudo-CLIP

En el tercer enfoque se ha diseñado un modelo multimodal que integra información textual y visual. Este enfoque surge de la necesidad de capturar la riqueza semántica que emerge cuando se combinan diferentes tipos de datos, cómo son la transcripción del meme y el meme en sí junto con inspiración de modelos como CLIP. Utilizando la representación textual de DistilBert y la potencia de ResNet50 se busca extraer y fusionar características relevantes de ambos dominios

La arquitectura específica consiste en una primera etapa donde ambos modelos, DistilBert para texto y ResNet50 para imágenes, procesan sus respectivas entradas. Las representaciones obtenidas son luego procesadas por capas adicionales de reducción y activación, con el objetivo de normalizar y preparar estos datos para la fusión efectiva.



La fusión de características se realiza concatenando las representaciones de texto e imagen, seguido de múltiples capas lineales y de activación. Esta parte de la arquitectura permite que el modelo aprenda de la interacción entre las modalidades, potencialmente capturando patrones que no serían evidentes al tratar los datos de forma aislada.

Finalmente, la salida del bloque fusionado se pasa por un clasificador compuesto por varias capas lineales culminando en una predicción de clase.