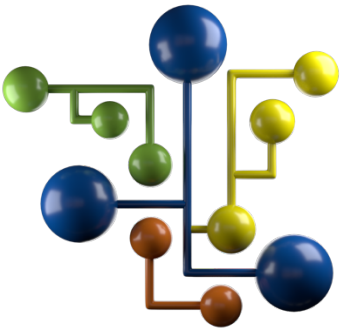


Data Science Academy

www.datascienceacademy.com.br

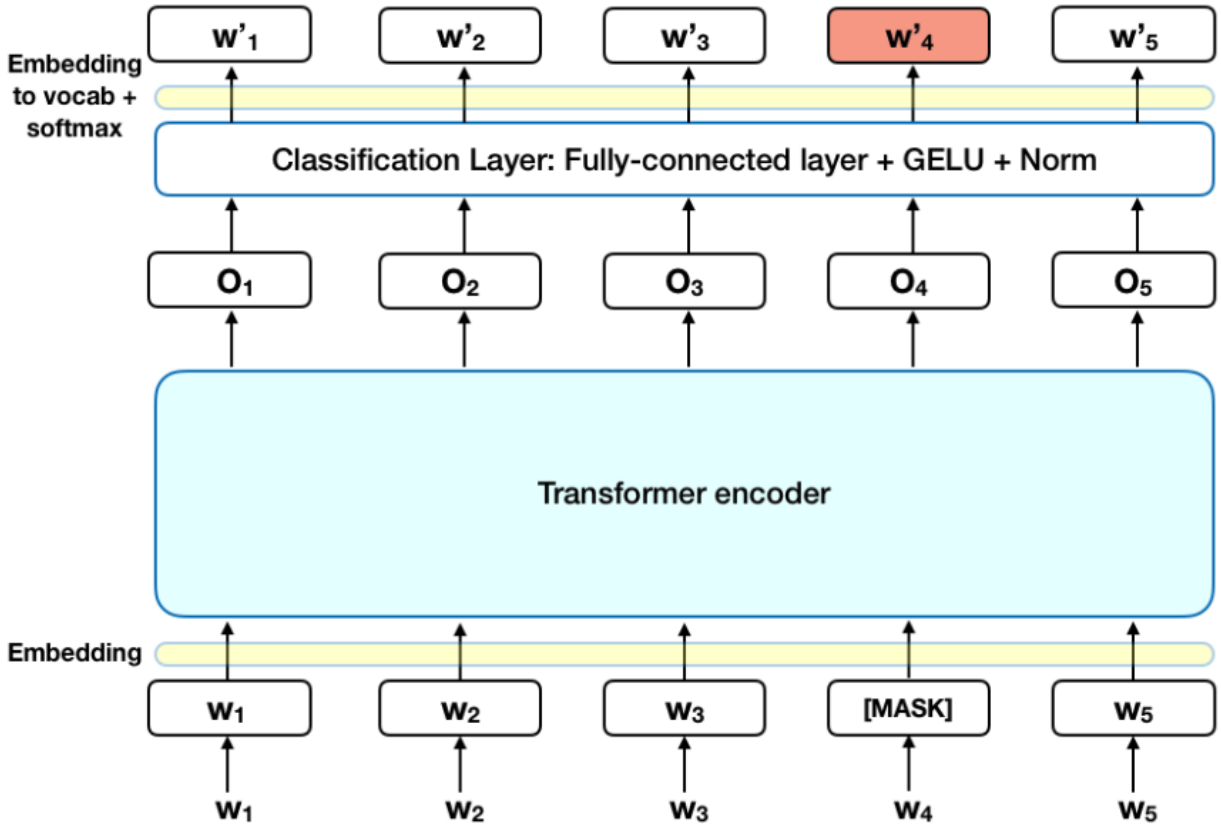
# Processamento de Linguagem Natural

## Masked LM (MLM)

Antes de alimentar as sequências de palavras no BERT, 15% das palavras em cada sequência são substituídas por um token [MASK]. O modelo tenta prever o valor original das palavras “mascaradas”, com base no contexto fornecido pelas outras palavras não mascaradas na sequência. Em termos técnicos, a previsão das palavras de saída requer:

- Adicionar uma camada de classificação na parte superior da saída do codificador.
- Multiplicar os vetores de saída pela matriz embedding, transformando-os na dimensão do vocabulário.
- Calcular a probabilidade de cada palavra no vocabulário com softmax.

Todos os passos estão descritos na imagem abaixo.




A função de perda do BERT leva em consideração apenas a previsão dos valores mascarados e ignora a previsão das palavras não mascaradas. Como consequência, o modelo converge mais lento que os modelos direcionais, uma característica que é compensada por sua maior conscientização sobre o contexto.

Referências:

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
https://arxiv.org/abs/1810.04805

Transformer: A Novel Neural Network Architecture for Language Understanding
https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html