ALBERT

- Versão mais leve
- Também possui diferentes versões: Base, Large etc.
- Menos parâmetros, utilizando técnicas de redução de parâmetros
- Possui performance superior a vários outros modelos baseados em Transformers

RoBERTa

- Robustly Optimized BERT Pretraining Approach
- Implementando com Pytorch
- Sem etapa de previsão de próxima sentença
- Treinado em diferentes tipos de textos (Notícias, novelas etc.)

ELECTRA

- Efficiently Learning an Encoder that Classifies Token Replacement Accurately
- Utiliza replace token detection technique (RTD) ao invés de mascaras
- RTD: tokens são substituidos ao invés de mascarados
- Diferentes versões: Small, Base, Large

XLNet

- Generalized Autoregressive Pretraining for Language Understanding
- Baseado em “Large Bidirectional transformer”: XL
- Utiliza técnica de permutação onde os tokens são previstos de forma aleatória

DistilBert

- Versão menor e mais rápida do Bert
- Desenvolvido pelo Hugging Face
- Baseado em knowledge distillation
  - Compressão de modelo, onde um modelo menor (estudante) é treinado a partir de um modelo maior (teacher)