# Comparison study of BERT, RoBERTa and DistilBERT

The BERT algorithm from Google and the transformation based NLP methods have been very popular recently. After BERT is published by Jacob Devlin et. el there was a big trend of exploring similar models and coming up with variations which are targeted at a certain category of NLP. XLNet and RoBERTa are flavors of BERT which improve performance of the original model while the DistilBERT improves inference drawing speed.

## BERT

This is a bi-directional transformer for pre-training on a lot of text data to learn language representation that can be then used to achieve specific natural language processing tasks using machine learning. While the BERT is better on various parameters, the main improvement is on performance which can be attributed to the bi-directional transformation, pre-training tasks of Masked Language Model and Next Sentence Prediction.

### *Using BERT in the analysis*

BERT is open sourced at https://github.com/google-research/bert and pre-trained for 104 languages with implementations in TensorFlow and Pytorch. It can be fine-tuned for several types of tasks, such as text similarity, text classification, text labeling such as parts of speech, question and answer, named entity recognition etc. However, pre-training BERT can be computationally expensive unless you use a GPU.

BERT authors have also released a single multi-lingual model trained on an entire Wikipedia dump of 100 languages. Multilingual BERT has a lower performance compared to English.

According to the blog, below are the statistics of training the BERT model.

| | TPU Pod | TPU Chips | TPU Cores*1 | PFLOPS*2 | GPU*3 |
|---|---|---|---|---|---|
| $BERT_{BASE}$ | 4 x 4 days | 16 x 4 days | 32 x 4 days | 0.7 x 4 days | 2 x 50-70 days |
| $BERT_{LARGE}$ | 16 x 4 days | 64 x 4 days | 128 x 4 days | 2.9 x 4 days | 8 x 50-70 days |

Dataset above 100K shows robust performance. BERT outperforms many algorithms including Pre-Open AI SOTA, BiLSTM and OpenAI GPT.

## RoBERTa

This algorithm was introduced by Facebook and very well optimizes original BERT. This is a retraining of BERT using the training methodology by 10 times.

Works Cited

"[1810.04805] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *arXiv*, 11 October 2018, https://arxiv.org/abs/1810.04805. Accessed 29 October 2022.

"Home." *YouTube*, https://arxiv.org/ftp/arxiv/papers/2104/2104.02041.pdf. Accessed 29 October 2022.

Khan, Suleiman. "BERT, RoBERTa, DistilBERT, XLNet — which one to use?" *Towards Data Science*, 4 September 2019, https://towardsdatascience.com/bert-roberta-distilbert-xlnet-which-one-to-use-3d5ab82ba5f8. Accessed 29 October 2022.

Khan, Suleiman. "BERT Technology introduced in 3-minutes | by Suleiman Khan, Ph.D." *Towards Data Science*, https://towardsdatascience.com/bert-technology-introduced-in-3-minutes-2c2f9968268c. Accessed 29 October 2022.