Transformers:

The paper “Attention is all you need” introduced transformers to the world in 2017. They were designed as a way of improving on previous recurrent neural networks such as LSTMs, which only took data in sequentially and were relatively slow to train. Like the name suggests, the neural network is primarily built of ‘attention vectors’. Essentially, for each word in a sentence, calculate the importance of it in relation to each other word in the sentence. The result is an attention vector, where each element in the vector is how each word is related to every other word in the sentence, thus capturing context.

Transformer networks contain two similar but distinct building blocks: encoder and decoder blocks. Take the task of translating from English to German. The encoder block is responsible for learning what language itself is, while the decoder block then uses the output of the encoder block to map German language to English.

BERT:

In 2018, BERT was created by a Google research team. BERT was designed for language modelling, the task of predicting the next word in a sentence given all previous words. As such, an understanding of language itself is required, and BERT is therefore simply multiple encoder blocks stacked one on top of the other. BERT is pre-trained, which means it has already been trained to understand language, and as such it can be leveraged to solve other problems for which it was not initially designed to. For our purpose this is sequence classification, detecting whether a sentence is considered sarcastic or not.

We use a distilled version of BERT, DistilBERT. This is a student model that has learned the behaviour of the full model, denoted as the teacher. As explored in \_\_ paper, DistilBERT retains 97% of BERT’s performance, with almost 40% fewer parameters and is 60% faster than BERT. As such, this allows use to train our fine tuned model on much more data in shorter time.

Baseline:

While BERT based models provide state-of-the-art performance (or at least have done so prior to recent advancements such as XLNET), we decided to benchmark performance across a range of simple classification baselines, such as Logistic Regression, Naïve Bayes and Decision Trees which all are suited to classification tasks but not for language modelling. We also train rudimentary DistilBERT models across a single epoch and with limited training data to evaluate the model’s robustness.