# Project Proposal

## Background

In the last few years, ChatGPT and other similar generative language models have changed how the world sees AI. But generative models like GPT (Generative Pre-trained Transformer) that create human-like text are not the only possible application. Discriminative models like BERT (Bidirectional Encoder Representations from Transformers) can be used to understand and interpret text, for specific tasks such as sentiment analysis and reading comprehension. There are many pre-trained BERT models available from HuggingFace, many of which have been trained for a particular task.

## This project

For my last capstone project of my time at Springboard, I propose to use a basic pre-trained BERT model and fine-tune it for sentiment analysis from multiple data sources found on Kaggle. The labeled data sets I plan to use include reviews from IMDB, Amazon, and Yelp. Each dataset contains reviews from these websites, categorized by high or low star ratings and labeled as either positive or negative accordingly. The IMDB dataset contains 50 thousand reviews, the Yelp dataset about 600 thousand, and the Amazon dataset nearly 4 million.

The steps of this project will include preprocessing the labeled data into a form usable by a BERT model, including tokenizing the data with a tokenizer from the HuggingFace Transformers library. Next, I'll set up a training pipeline. Then I'll be able to finetune the model, and finally to evaluate it and document my results.

## Possibilities for future extension

This should be a relatively straightforward task, but will give me experience in using libraries such as HuggingFace, as well as experience with Large Language Models and various aspects of Natural Language Processing. Although the course's timeline puts an upper bound on the complexity of this particular project, I should be in a good place once the course ends to extend the scope, perhaps by training my own BERT model, or using larger, less clean datasets. . One application I would be interested in in the future is a cross-lingual sentiment analysis. This project should be a stepping stone to many potential future projects in the field of Natural Language Processing.