Marco Lipari, Andrew Tomajian, James Wnek, Roy Elia

1. **Problem statement:**
   Our project aims to develop a machine learning model that can estimate the year a Reddit comment was written based. The ultimate goal is to create a web application where users can input a comment and see which time period the model predicts it belongs to.
2. **Data Preprocessing:**
   This is a link to the dataset used:
   https://academictorrents.com/details/ba051999301b109eab37d16f027b3f49ade2de13
   The dataset contains, by month from 2005-06 to 2024-12, json files with all the reddit comments for that month. Because that is too much data for this project's purposes we wrote a bash script which takes the first 20000 comments for each month and writes to a csv file the following for each comment: name of subreddit, subreddit id, body, date created. Each month has its own file.

3. Machine learning model:
   Before predicting an actual time period, we decided to test out the viability of the model by trying to classify comments into one of two time periods: either 2013 or 2024. The comments were first tokenized using the distilbert-base-uncased model and were turned into an array of contextualized embeddings for each token (indicating the context of each word). The main model used was Bert for sequence classification with a training/test split of 90-10.
4. So far our model was able to predict if a comment came from 2013 or 2024 with 72% accuracy, which is more than random guessing. This is especially impressive given that it is not a very obvious task even for us as humans; a potential metric for performance we could try is to compare it to the accuracy of a human classifying the comment. Although getting the model to take in as input any comment from 2005-2024 and output a prediction using regression will definitely be difficult, the preliminary results are encouraging.
5. The first step will be to tweak the hyperparameters of this model and do some research about optimal train-test splits and even potentially other models and compare their performance. Then we will attempt to use a regression head to expand our model to predicting the whole range of data. The biggest challenge will be managing the sheer amount of data that we have, so some research on the optimal amount to choose for our patience level will also need to be done.