Spam Detection With BERT

Kevin Zhuo & Daniel Zhang

1 Introduction

In our project, our goal is to utilize a BERT-based model in order to identify and filter out spam messages from emails, social media, or other types of communication. We believe that BERT has potential to be used for bidirectional encodings and allow for greater ability to distinguish between spam and non-spam. In our project, we will first tokenize the sentences using the BERT embeddings, converting them into numerical inputs which can be fed into the BERT model. Eventually, we plan to used a fine-tuned BERT model in a supervised learning approach on the training dataset. The hope is that eventually we can evaluate on the model on the held-out test set and achieve good results using the BERT model.