Neural Content Moderation
By Isabella Garcia-Camargo, Martin Amethier, Guy Wuollet

Motivation
- Content Moderation is crucial for healthy online social spaces
  - Different communities have different standards
  - Content volume and moderator mental health requires automated moderation

Data
We have two datasets. The first is a Kaggle dataset from the Toxic Comment Classification Challenge and contains Wikipedia comments. The second is Reddit data and comes from Pushshift and from Hybrid Approaches to Detect Comments Violating Macro Norms on Reddit. The Reddit combines 2 million moderated comments and 2 million unmoderated comments from top 100 subreddits by popularity during 2016-2017.

Looking Forward
- Data Augmentation is a great next step. We built transformation functions to swap adjectives and replace nouns, verbs, and adjectives with synonyms. Unfortunately, Snorkel's SSL certificate failed and we were unable to troubleshoot.
- As different communities have different standards, it would be useful to train classifiers on specific subreddit data. This could be done using transfer learning from a more general model, or as a framework for creating domain specific models.

Models
- We ran a total of seven models: LSTM, GRU, Double-LSTM, Double-GRU, Triple-LSTM, Bidirectional-LSTM, and a TF-IDF NN.
- We found that the GRU outperformed the LSTM models. This is likely due to the shortness of our sequences, which were on average 300 tokens long with a median of 150 tokens.
- Bidirectionality also improved our results, signaling the importance of backwards relationships between tokens.

References