Data Selection Proposal

MAIS 202 Deliverable 1

Sarvasv Arora, Kaiwen Xu

February 8, 2021

Project Idea

We plan to make a discord/reddit bot that replies to specified comment/post with a Deep Learning generated sarcastic — but fun — comment.

"Everybody a sarcasm gangsta until the bot kicks in #_-# "

Dataset selection

Reddit is a fun place, full of witty people who oftentimes post comments that are *r/nextf*ckinglevel* and sarcastic.
We found the following datasets that contains a list of such parent-comment pairs, and aim to use these in our project's hunger for data.

- **Sarcastic Comments - REDDIT**

  → This dataset consists of 1.3 Million sarcastic and non-sarcastic comments, along with the parent comment, and the score of the comment.

  → The sarcastic comments are determined by the "\ s" tag, which is often used by Redditors to explicitly state that their comment is indeed sarcastic. This makes the dataset more definitive, and a good choice for our project.

- **Reddit parent comment pairs**

  → This is a dataset of parent-comment pairs on Reddit, and is of size 1.53 GB. It also has the score of the comment.

  → We didn't explore this dataset much yet, but this is more like a backup dataset, in case we need some more training data.

Methodology and goals of the project

Since we have a pretty well defined dataset, we will first separate out the sarcastic comments out of it. Then, we will further clean the dataset by selecting only those comments that have a certain score higher than some threshold. This is a hyperparameter, and we will decide upon this later when we do more in-dept analysis.

Next, the model itself would be a bidirectional-LSTM (most probably, an attential model) which would take in the parent comment as the input (encoder) and we'll train it to produce a sarcastic comment (decoder). In this process we will use some pre-trained word embeddings, since training our own is out of the scope of this project.

Our evaluation metric would include calculating the BLeU Score on the test set data.

Application

For the application/demo, we aim to build a reddit or a discord bot that sends the specified comment/text to a custom API that does inference on the backend.