SI630 Project Proposal: Witty Reddit Bot

Anonymous ACL submission

1 Introduction

Reddit is the fifth largest social media platform in the United States, and is a microblogging site with over a billion registered accounts. All of Reddit is broken up into sub-reddits, each related to a specific topic. Users can create posts, and infinitely deep comment threads on each sub-reddit. The primary goal of this project is to create a bot that can post a top level comment on an arbitrary Reddit post. A bot is essentially software that has access to the Reddit API, and can interact with the application automatically. These interactions will be sourced from a trained model, that will dictate what the bot comments on a Reddit post. The intent is to have these comments be witty and humorous, as is usually the case with most Reddit posts.

2 Task Definition

It must be noted that this project is not aimed at being a conversational AI, or to engage in social or political discourse on the internet. It will merely generate quips and witty remarks, which may end up relying on nonsensicality in order to be funny.

To break down the goal further, there are two major components to this project.

2.1 Model

This component is the core part, which spits out the comment body to be posted. It will take as input the text of an arbitrary Reddit post from a sub-reddit, and generate text that is relevant to the post (and hopefully funny). For example, a top comment on a post about a user's date getting too drunk and ruining his house is "My guy you brought in the equivalent of a stray cat".

2.2 Bot

This component will utilize the Reddit API to fetch posts (preferably new ones, as the visibility for comments is higher) and then feed them into the model. Once the model produces an output of the potential comment body, the bot will then make a top level comment on the post with the generated text. The comment IDs will be saved for future lookup to analyze how well the comment performed (in terms of upvotes, described further in the evaluation section)

3 Data

There exists a significantly large data dump of Reddit posts and comments compiled by Baumgartner et al. (2020) called Pushshift which makes available an API as well as raw downloads of content aggregated by month. The data is largely in JSON format, and an example comment has the following structure with some fields left out for brevity:

```
{"author":"just-a-stoner",
"body":"ahhhh i remember the days of
water-bottles, pen tubes, tape
and sockets....",
"controversiality":0,
"created_utc":1506816003,
"id":"dnqik3y",
"parent_id":"t3_73ifkm",
"retrieved_on":1509189608,
"score":12,
"subreddit":"trees",
```

To start off with, the model may be trained on a subset of this data (perhaps only posts and comments from within a text-only sub-reddit). However, there is no shortage of data as the repository contains "51,778,198 submissions and 5,601,331,385 comments posted on 2,888,885 subreddits" (Baumgartner et al., 2020)

4 Related Work

Writing an NLP based social media bot is not the most novel idea, and has been attempted before in a variety of contexts.

In a much more serious endeavor than is intended here, Sager et al. (2021) attempt to tackle misinformation on Reddit. Their model attempts to identify misinformation around the topics of essential oils in Reddit's dermatology related forums, and intervenes with text that is meant to convey the right information. However, their focus is on identifying the misinformation and not necessarily on the text-generation as that is made up of templated responses with general facts about essential oils.

Another effort, which has more focus on text generation is by Çetinkaya et al. (2020). They build a model using state-of-the-art architectures, to both classify a tweet as well as generate a reply that either agrees or disagrees with the original tweet. The model is trained and tested on the gun control debate in the United States, and the bots are trained to first classify if a tweet is for or against gun control. Following the classification, a bot from either side generates a tweet reply that either agrees with its side, or provides a meaningful response to a tweet from the opposing side. They also propose a new salience metric, to evaluate the relevance of the generated tweet to the original.

Finally, there is a thesis by Willemsen (2017) in which the author trains a sequence-to-sequence model to reply to Reddit comments. They use the same dataset (Baumgartner et al., 2020) intended to be used in this project, and employ different kinds of models. They evaluate models with character-level representations, word-level representations, and hybrid-level representations to get the best of both worlds and present a comparative study. They also use a retrieval based model that uses a weighted score to pick the best reply from a collection of replies as a gold standard to compare against. They also use multiple evaluation techniques like computing scores for BLEU, ROUGE, and CIDEr, while also critiquing overlap based evaluation criteria as being ineffective for these kinds of generative tasks as the potential responses span a large space.

All of these ideas tie into the goals of this project, and using these works as references will aid in its progress.

5 Evaluation

Since this task is primarily based on text-generation, none of the classic evaluation metrics like accuracy or F1-scores are applicable. However, there are alternative evaluation metrics which are applicable for this project, two of which are detailed here.

5.1 Perplexity

The first is the perplexity score of the model for a test set. The idea of perplexity is that the model should be able to assign relatively high probabilities to sentences that generally fit the bill of "witty" and "humorous" while assigning lower probabilities to sentences that do not.

This can easily be computed on a collection of hand-selected comments from the dataset, which can serve as the test data.

5.2 Upvotes

Once the model is trained, and the bot is deployed, a live feedback loop is established where the users of Reddit are potentially upvoting or downvoting the comments made by the bot. An average score of these upvotes can potentially be an interesting metric to gauge the performance of the model.

6 Work Plan

While we work our way to seeing material relevant to this project in lecture, I plan to use the buffer time to set up the peripherals of the project. This allows more flexibility for training and fine-tuning of the model during the latter half of the semester, when deadlines are closer.

Writing the bot does not require any NLP expertise, and I expect to have an extensible prototype available in about two weeks from the date of this submission.

In the coming weeks, I also plan to set up a testing framework, with hand annotated "good" and "bad" examples for comparing perplexity scores, as well as a way to monitor the number of upvotes a comment has received.

Finally, this will open up enough time later in the semester to work on the main chunk of the project which is the model training, for which I currently do not possess complete clarity in terms of methodology.

References

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *International Conference on Web and Social Media*.

Yusuf Mücahit Çetinkaya, İsmail Hakkı Toroslu, and Hasan Davulcu. 2020. Developing a twitter bot that can join a discussion using state-of-the-art architectures. *Social network analysis and mining*, 10:1–21.

Monique A Sager, Aditya M Kashyap, Mila Tamminga, Sadhana Ravoori, Christopher Callison-Burch, and Jules B Lipoff. 2021. Identifying and responding to health misinformation on reddit dermatology forums with artificially intelligent bots using natural language processing: Design and evaluation study. *JMIR Dermatology*, 4(2):e20975.

Bram Willemsen. 2017. *I Am A Sequence-to-Sequence Model trained on Reddit Data, Ask Me Anything! Generating Replies to Reddit Comments with Attentive Encoder-Decoder Networks*. Ph.D. thesis, Tilburg University.