Whiskey GPTaster

Stanford CS224N Custom Project

Akram Sbaih
Department of Computer Science
Stanford University
akram@stanford.edu

Mentor **John Hewitt**
Department of Computer Science
Stanford University
johnhew@stanford.edu

Abstract

Recent advances in language modeling and text generation started raising concerns on the potential for machines to generate fake news and reviews. These models now generate text shown to be indistinguishable from human text [1]. Some recent approachs [2] train classifiers to discriminate fake and real text. These usually require access to the generator or its training data which aren't usually available in the wild. We're also faced with the constraint of having few examples of identifiable good fakes. In this work, I study the performance of classifiers with combinations of [the generator, its training data, its generations]. To this end, I generated a set of fake whiskey reviews using GPT2 to evaluate these approachs.

1 Approach

The goal is to make a classifier that takes in a Whiskey review alongside its metadata (price, rating, and name) and decides whether it's human generated or machine generated. In this project, I define a human-generated review as one that has been scraped from a crowdsourced dataset, while a machine-generated review is one made by GPT2 finetuned on that human dataset. These definitions reflect only a small combination of the potential samples in the wild, but still serve as a good evaluation since GPT2 is one of the widely-used architectures in the market. To this end, there are multiple approaches that I would like to compare.

- Finetune GPT2 to be a classifier that takes these input as a sequence and outputs a token deciding whether the review is real or fake, starting from a publicly available GPT2 checkpoint.
- Finetune GPT2 as the previous approach but start from the checkpoint that was used for generation similar to the approach proposed in [2].
- Compare the preplexities on the real and generated sequences for both the public and generator-finetuned checkpoints of GPT2 and train a classifier on those preplexities if necessary.

We should expect that all three approaches perform well. However, each of them offers a different set of constraints. For example, we probably will not have access to the finetuned generator in the wild. We also might be faced with a very limited number of fake reviews from the same generator that we can train a classifier on. Therefore, comparing the results of these approaches offers a valuable understanding of what we can and cannot do to combat bad use of language models for generating fake reviews.
To evaluate these approaches, I set the following milestones

1. Prepare a dataset of human-generated reviews with their metadata.
2. Finetune Huggingface GPT2 on that dataset to generate novel fake reviews making a new fake-reviews dataset.
3. Implement each of the approaches and assess their accuracy.

2 Experiments

Data
www.whiskyadvocate.com is a website where people write their reviews and ratings for various whiskeys. I made a script that scraped all the reviews on the website totaling 5649 unique ones. Each review contains the name of the whiskey, its price in dollars, its rating out of 100, and the text review of its taste. Some examples are shown in Table 1.
I split the data into train, validation, and test sets being 90%, 6%, and 4% which are 5084, 339, and 226 samples, respectively. I intended for the training split to be bigger than the convention because of the limited number of total samples. The test split is also small because main evaluation done on it will be qualitative which is labor-intensive.
The reviews are reformatted as sequences to be fed into GPT2. An example of the formatting I used is `"<price>105<rating>89<whiskey>Springbank 11 year old, 58%<review>Finished in a rum cask. Gently sweet [...] exclusive.)"` where `[...]` indicates omitted text.

Evaluation method

1. Preplexity. This is the average uncertainity the language model has over the sequence it's evaluated on (the lower the better). Since GPT2 is a language model, I used this metric to track how well it's adapting to the new task of modeling whiskey reviews during finetuning. It's also a measure of how well the the final model performs.
2. Qualitative assessment. Since I don't have access to enough labor-work to turn human assessment of the generations into a quantative measure like in [1], I resorted to my personal assessment. This involves looking at the generated samples for common themes in whiskey reviews. This includes things like having a description of common flavors, distinction between nose and palate, having an after-tast, and mentioning the name of the whiskey in the review.
3. Accuracy. This is useful for the classifiers proposed in the approach section.

Experimental details
After preprocessing the dataset, I ran the Huggingface GPT2 Trainer on the training and validation splits for 5 epochs starting with their publicly available pre-trained GPT2 checkpoint. I chose a batch size of 2 per device beecause of the limited available memory. Each sample in the batch is the sequence concatenation of multiple samples from the training set so that we use up all the available GPT2 block size and train faster.
I trained on an Azure machine with two K-80's making a total batch size of 4 with 4 dataloading workers. I used the default learning rate of $5e-5$ and the default Cross Entropy loss function. Training took 31 minutes and ended up with a preplexity of 15.49 on the validation set. You can see the loss curve in Figure 1.

Results
The baseline for the generation task is the human generated text since we're not comparing the model to other models because that's not the goal of this project. I use my human evaluation to look for common patterns in the generated text as mentioned in the Evaluation Method section. The generations seem satisfying except for an observation that the model tends to list too many flavors sometimes. Table 1 shows a few randomly selected generations from the test set.

3 Future work

At this point, I finished the first step of generating the fake reviews dataset and evaluating it. Next, I need to do the following:

1. Implement and train classifiers proposed in the Approach section.
2. Gather statistics on the performance of these classifiers and their usability in the wild and what constraints they require.
3. Engage more into the ethics of the fake reviews problem and the philosophy of its constraints in the wild.

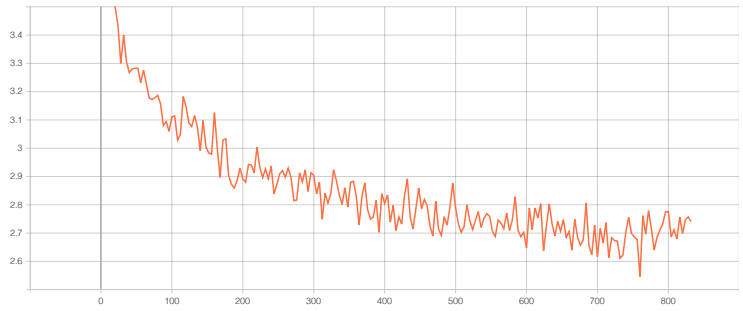

Figure 1: GPT2 Training Loss for 5 epochs.

Real	**Fake**
<price>495<rating>89<whiskey>Bainbridge Yama American Single Grain Barley Mizunara Japanese Oak Cask, 45%<review>This high-end whiskey exhibits restrained oak, elegance, and delicateness, with wonderful poached pear, cereal, crème caramel, floral, and lemon chiffon cake aromas that yield to a bright beam of tart, mouthwatering citrus—clementine, lemon, and yuzu—tingling with allspice. Bright, light, and lively, but not lacking in complexity, finishing with marshmallow, toasted almond, and marzipan. Very pretty! American single grain whiskey aged in Japanese Mizunara oak casks.	<price>495<rating>89<whiskey>Bainbridge Yama American Single Grain Barley Mizunara Japanese Oak Cask, 45%<review>Distilled in 2003 and bottled in September 2016. Rum mizunara has been a popular whiskey for a very long time, one worth considering. The nose offers light toffee, sweet orange syrup, lime juice, light honey, sandalwood, and light oak, adding some teasing dried spice notes. The palate is spicy, with caramel and milk chocolate, red currant, and dark fruit. It finishes with coffee, orange rind, and cocoa powder, finishing dry and spicy.
<price>225<rating>94<whiskey>Amrut Greedy Angels, 50%<review>A whopping three-quarters of the spirit put in these casks was taken by greedy angels. It has a big waft of crystallized pineapple, tropical fruits, and spiky spice on the nose. On the palate, red licorice, syrupy jellied fruits, some mandarin, cherry lozenge, and tinned strawberries, and the same menthol rancio you'd kill for in a 30 year old scotch. This is Amrut's oldest-ever whisky; it's as rare as hen's teeth. . . and just 8 years old. Awesome.	<price>225<rating>94<whiskey>Amrut Greedy Angels, 50%<review>After pushing back into the woods this year, an aggressive and well-integrated expression is finally showing its chest. A blend of Irish oak, red berry, and grape skins, there's no rush. The nose is fruity and sweet, with fragrant nuts and ripe rhubarb, ginger, nougat, tarry rope, and cocoa. It's a mix of different Irish oak expressions. But it's got more depth than the others, so give it a whirl. (6,650 bottles)

Table 1: Real on the left and Fake on the right from the test set.

References

[1] David Ifeoluwa Adelani, Haotian Mai, Fuming Fang, Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. Generating sentiment-preserving fake online reviews using neural language models and their human- and machine-based detection, 2019.

[2] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news, 2020.