Whisky GPTaster

Stanford CS224N Custom Project

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

I still have no mentor for this project.

1 Research paper summary

Title	Generating Sentiment-Preserving Fake Online Reviews Using Neural Language Models and Their Human- and Machine-based Detection
Authors	David Ifeoluwa Adelani, Haotian Mai, Fuming Fang, Huy H. Nguyen, Junichi Yamagishi, Isao Echizen
Venue	Advanced Information Networking and Applications – AINA
Year	2020
URL	`https://arxiv.org/abs/1907.09177`

Background.

The authors point out the huge implications of reviewes on businesses in our online world given their status as a reference for potential new clients. They cite research showing how the business could gain big profits or suffer losses as the result of a mass review attack on their products. To be able to detect and stop such attacks, the authors pursue a better understanding of powerful methods to generate fake reviews.

In their discussion of related works, they talk about techniques like manual generation and basic language model (LSTM-based) utilization. The former is too expensive while the latter requires post-processing to make its results centered around the business attacked (to stay on topic). They also mention the recent introduction of more sophisticated language models like GPT-2 and BERT and their advantages of pre-training and fine-tuning leading to more consistent generation of longer sequences.

The authors use these points to motivate their goal of using GPT-2 and BERT to generate fake reviewes of a given sentiment for a given business with minimal human internvention. Concretely, they want the model to take an example review for a business, and generate a pool of fake reviews describing the same business with the same sentiment. They also want to evaluate existing machine and human techniques on distinguishing these fake reviews from real ones.

Summary of contributions.

The authors implemented their novel algorithm as shown in Figure 1. It consists of a generation step that takes a sample existing review as a seed, and employs GPT-2 to generate a new fake one. After that, BERT is used in the validation step to find the sentiment of the fake generation and compare it to the sentiment of the original review. This way, BERT filters out all bad generations that have mismatching sentiment as desired.

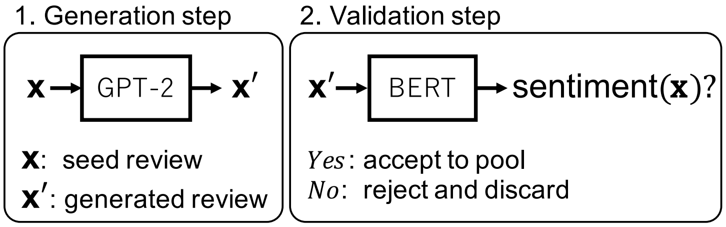

Figure 1: Fake review generation procedure

They use the pretrained version of GPT-2 (with 117M parameters) and BERT and fine-tune them on positive and negative real reviews from Amazon and Yelp (they don't disclose the size of these datasets). No pre-processing is done except concatenating all the reviews into one huge text file separated by new lines and shuffled on sentiment. After fine-tuning, GPT-2's generation started looking more like reviews and BERT achieved 96% accuracy on positive/negative sentiment classification on the test set.

Furthermore, human and machine evaluation were mostly incapable of distinguishing fakes from reals. The subjective evaluation by 80 humans showed that they pointed out fake reviews at random while machines (Grover, GLTR, and OpenAI GPT-2 detector) showed similar behavior.

Limitations and discussion.

Reviewes tend to be similar in their content leading to a faster training process. However, the authors' method had to train GPT-2 for 485K epochs (2 weeks!) to achieve the mentioned results. They don't talk about the training it took the baseline LSTM to achieve its somewhat similar results. This makes me concerned that training is what improved the results, not the new architecture.

Moreover, the use of BERT as a sentiment classifying filter and training it separately sounds like a convoluted way instead of just using BERT as a data labeler and training GPT-2 on this synthetic labeled data. This is different from an adversarial training setup where the adversary's gradients actually help the generator learn, which they don't have.

They also don't yet propose, or even discuss, potential ways to successfully distinguish their fake generations from real reviews, which is against the goal of their research effort stated at the beginning.

Why this paper?

I chose this paper after deciding to use the whisky dataset to generate whisky reviews since it tackles a similar goal. I didn't find much literature as close as this one is because the task of generating fake reviews is a trivial one generally speaking. It becomes interesting when it's coupled with new techniques and variables. In the case of this paper, they track the variable of sentiment and they use the technique of a sentiment classifier on top of the vanilla generation task, which is relatively similar to what I want to explore.

It at least showed me that GPT is capable of doing this task. It also reiterated to me that a classifier (BERT in this case) is faster to train than a generator (2 hours vs 2 weeks), which might be useful if I consider using an aversarial training procedure in my project.

Wider research context.

This paper comes in the context of pretraining and fine-tuning that we talked about in class. Its task of generating fake reviews designed for financial gain comes in the greater context of fake news generation which is a hot political issue. This work was able to generate sentimental text that's indistinguishable for previous synthetic text detection research works (cited in the paper) and therefore calls for more research on those ends.

2 Project description

Goal.

I would like to investigate on whether GPT could do the following tasks:

1. Predict the rating and/or price of a whisky by reading its human review in a format similar to the birthdate question-answering we did for assignment 5.
2. Do the reverse and generate fake human-like reviews for whiskies based on their rating and/or price in a similar input/output formatting to assignment 5.
3. After evaluating the generations, seek higher diversity in the reviews generated by employing an adversarial training scheme.

In other words, I want to evaluate the fidality and diversity of GPT generations after fine-tuning it with a basic formatting like in assignment 5. After that, I want to evaluate how they improve under adversarial training with a random seed especially given that there aren't many available reviews on the dataset and they're sparse with respect to rating/price.

Task.

I show the proposed tasks in the Goal section and will share an example for each of the tasks here:

```
1. x:  The aromas from the base rye moonshin[...]?68%?$20?XXXXXXXXXX
   y:  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX[...]?68%?$20?XXXXXXXXXX
2. x:  XXXXXXXXXX?68%?$20?XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX[...]
   y:  XXXXXXXXXXXXXXXXXX?The aromas from the base rye moonshin[...]
3. x:  ZZZZZZZZZZ?68%?$20?XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX[...]
   y:  XXXXXXXXXXXXXXXXXX?The aromas from the base rye moonshin[...]
```

where X and ? are special characters and Z is random noise.

Data.

I found this dataset on **this Kaggle entry** and later found out it was used in **this similar application**. Both of these have used portions of the data available on the original whisky review website **Whisky Advocate**. I'm planning to use **this github** simple scraping script(or my own) used to generate the partial dataset to generate a bigger one which shouldn't take more than a minute to run.

I'm expecting to have 4k+ human reviews. Each has its corresponding rating, price, and whisky name. No processing is needed other than the processing described in the Task section.

Methods.

The method for the first two tasks is very similar to assignment 5 and therefore won't elaborate. The method for the adversarial training task (number 3) is very different from the original paper. I will use another GPT as the critic and will have both the generator and critic training at the same time with gradients flowing from the critic to the generator.

Baselines.

Since the most original part of this project is task 3, I'm planning to use the generations of task 2 as the baseline.

Evaluation.

The difference between task 3 and task 2 is in the diversity of outputs while still being conditioned on the price and rating. And task 1 offers a classifier (regresser) that evaluates how much a review predicts price and rating. We can use task 1 as a feature extractor and feed the generations of both task 2, 3 and the human reviews to it. After that, we can use the Fréchet distance (which compares two distributions) to see how much the generations of task 3 are closer to the human reviews than are the generations of task 2. This kind of evaluation is widely used in GAN's as the Fréchet inception distance.