Sentiment-Based Auto-Complete for Customer Reviews

Marcus Alenius
malenius@andrew.cmu.edu

Emily Jiang
emilyjia@andrew.cmu.edu

Max Kulbida
mkulbida@andrew.cmu.edu

Henry Siegel
henrysie@andrew.cmu.edu

Abstract

We explore fine-tuning, using GPT-2 as a base model, as a method to generate customer reviews, when given a star rating, a tone, and an initial text stub. By using the Yelp Review dataset for training, we hope to train the model to output text that is consistent with initial text stub and the provided sentiments. The fine-tuned model showed notable improvements, including more relevant and consistent reviews. There was a quantifiable 20% reduction in cross-entropy loss from the one-hot encodings of the intended sentiments, when compared to the baseline, measured using a state-of-the-art BERT based classifier. However, there are still difficulties with the fine-tuned model, such as its struggle with "turning around" contradictory or confusing prompts as well as occasional large inaccuracies. Future efforts could explore larger models like GPT-4, or additional training techniques such as retrieval-augmented generation, which could enable greater factual consistency.

1 Introduction

In our project, we investigate the effects fine-tuning has on the quality of generated customer reviews, in hopes of generating a model which serves as an "auto-complete" for reviews which can extend the beginning of a review given the intended star rating for the establishment being reviewed, as well as the tone for the review. This is inspired from the new context-aware auto-complete features of for example, Gmail and Microsoft Word, which suggest in-context continuations of what you are typing, but our model also accounts for the star rating you are giving.

We attempt this by fine-tuning the smallest version of GPT-2, which has 124M parameters, on the Yelp Dataset, containing Yelp reviews as well as the star rating associated with each review. The hope is that this model and its performance can serve as evidence of the feasibility of such a tool in the future regarding customer reviews and general auto-completion for specific settings, whether that be other types of reviews or other tasks.

2 Summary of data

2.1 Dataset

We use the Yelp Reviews dataset:

`https://huggingface.co/datasets/Yelp/yelp_review_full`

The dataset contains customer reviews of businesses from Yelp. Each data point consists of an English-language textual review as well as the given star rating on a scale from 1 to 5.

Preprint. Under review.

The Yelp Reviews dataset was constructed by Xiang Zhang and first used in text classification benchmarking Zhang et al. (2015). The paper explores the usage of character-level convolutional neural networks for text classification. By construction of this dataset along with others, it was shown that character-level convolutional networks could achieve competitive performance compared to traditional state-of-the-art models such as bag of words and n-grams.

The dataset consists of 650,000 training samples and 50,000 testing samples. They were constructed by randomly taking 130,000 training samples and 10,000 testing samples for each rating from 1 to 5.

For the purpose of fine-tuning GPT-2 for the task of auto-completing customer reviews with a given sentiment, the Yelp Reviews dataset is well-suited. First, it is a large collection of real customer reviews and thus reflects authentic writing styles. We would expect this to align with how users expect an auto-completion tool to generate text. Second, its diverse nature – ranging across a wide variety of businesses and topics – allows the model to generalize across industries. Third, the star rating labels, serving as a proxy for sentiment, allow the model to learn how to generate text with a specific sentiment or tone.

2.2 Biases

There is an inherit self-selection bias in review datasets. Certain types of customers are more likely to leave reviews than others. Highly satisfied and highly dissatisfied customers are more likely to leave a review and spend time making it thoughtful and comprehensive. Mildly satisfied and mildly dissatisfied customers likely will not leave a review at all or if they do they may write a shorter, less detailed one.

While the dataset contains an equal number of samples for each rating, there are no other guarantees in the dataset. For example, one type of business may be overrepresented and another type may be underrepresented. One can imagine that restaurant reviews are more common on Yelp than auto shop reviews, and thus the former would be overrepresented and the latter underrepresented in the dataset.

Furthermore, there could be an imbalance between the distribution of business types over star ratings. For instance, a disproportionate number of 5-star reviews might come from luxury restaurants, while 1-star reviews might skew toward budget services. This could result in context-dependent biases, where the model associates sentiment with specific business types.

3 Methods

3.1 Existing baselines

Fine-tuning pretrained language models, such as GPT-2 and BERT, has shown to produce state-of-the-art results. Both GPT-2 and BERT have been used for language generation tasks, but Wang and Cho (2019) found that GPT-2 produces text of higher quality.

Previous work has fine-tuned GPT-2 to excel in domain-specific text-generation. Lee and Hsiang (2020) fine-tuned GPT-2 for generating patent claims. Using a dataset of 555,890 patent claims, the 355M parameter GPT-2 was fine-tuned. It was found that the model started generating patent-like text after as few as 36 fine-tuning steps.

van Stegeren and Myśliwiec (2021) fine-tuned GPT-2 on video game quests to generate dialogue for non-playable characters in video games and evaluated the quality of the generated text through human studies. Presented with both human-generated and model-generated quests, participants were asked to rate the text on a Likert scale in terms of language quality, coherence, and creativity. While it was found that the model performed statistically significantly worse on language quality and coherence, the performance on creativity did not show significant differences.

Most notably, Adelani et al. (2020) used GPT-2 to generate a large number of "fake" reviews based on a reference review with the desired sentiment. Their model was fine-tuned on Amazon and Yelp reviews. They observed that the pre-trained GPT-2 sometimes produced texts that did not read like reviews, however after fine-tuning, the generated texts were review-like. Additionally, they learned a sentiment neuron using a single-layer multiplicative LSTM. By replacing the output values of the sentiment neuron with +1 or -1, they were able to explicitly force the output to be conditioned by a specified sentiment.

Table 1: Fine-tuning hyperparameters

Parameter	Value
Batch size	8
Epochs	3
Learning rate	5e-5
Warmup steps	500
Weight decay	0.01

Table 2: Training and validation loss

Epoch	Training loss	Validation loss
1	2.5504	2.7707
2	2.5272	2.7574
3	2.4363	2.7619
4	2.2832	2.7797
5	2.1821	2.8088
6	2.1299	2.8359

Their review generation pipeline consists of (1) the fine-tuned GPT-2 model that generates the review and (2) a BERT-based text classifier that filters out reviews with undesired sentiments. They conducted human evaluations which showed that the real and fake reviews were indistinguishable in terms of fluency.

Our work is different from current approaches as we are tuning the model for auto-completion. That is, given the start of a review and a given sentiment the model should complete the review in a coherent and reasonable way . This is in contrast to Adelani et al. (2020), who used a human-written reference review to generate new reviews. Our approach is completely reference-free, and instead only uses the start of the review, and the specified star rating and tone.

This approach allows the model to be used in new applications, where current similar models do not work well. One can see how our model could be integrated into an online review application to assist customers with writing reviews.

3.2 Fine-tuning GPT-2

We first downloaded the smallest GPT-2 model, with 124M parameters (Radford et al., 2019), and prompted the model to output Yelp reviews in the following way:

Continue the following Yelp review which gives a rating of x stars and has a y tone: z

Here, x is the sentiment and would be a number in the range of 0 through 4 (representing 1 through 5 stars), y is the tone, and defaults to "serious" if left blank (but can be any word), and z would be the beginning of a possible Yelp review (usually a short phrase). The model then generates a review with these three pieces of information.

Next, we fine-tuned the model, hoping to improve the relevancy, coherence, and accuracy of reviews. To perform the fine-tuning we loaded the full Yelp Review dataset. Then, due to memory and time concerns, we selected a subset of the training and testing datasets to train and test on (sizes 5000 and 500 respectively).

We use the Huggingface Trainer class to fine-tune the model. See Table 1 for the used hyperparameters. We stopped training after 3 epochs as we started to see overfittng in further training. See Table 2 for training and validation loss values.

3.3 Assessing results

To qualitatively asses the results of our model we prompted both the pre-trained and fine-tuned model as specified above. We used the same combinations of sentiments, tones, and starting prompts for

both the pre-trained and fine-tuned model. We analyzed how closely it aligned with the prompt variables. See Section 4.1 for this analysis.

To provide a quantitative assessment of our model we prompted it with the same format as above. For the initial part of the review, z, we used the first 10 words from reviews in our test dataset. The tone was set as "serious". We then passed the generated reviews into a BERT-based sentiment classifier, fine-tuned to determine the star rating (1 through 5) given a product review. Namely, we use `bert-base-multilingual-uncased-sentiment` for this. Tested on 5,000 held-out product reviews, this classifier achieved 67% exact-match accuracy and 95% off-by-one accuracy on English product reviews (NLP Town, 2023).

Over 500 samples of this format, with intial review T_i and intended sentiment s_i, we had both the pre-trained GPT-2 model and our fine-tuned model auto-complete the reviews, resulting in a completed review R_i. Then, each R_i was fed into the classifier, outputting a probability distribution P_i on $\Omega = \{1, 2, 3, 4, 5\}$, indicating how "confident" it was that R_i had each sentiment – or for our purposes, that it had that Yelp review rating. Then, we calculate the binary cross-entropy loss between P_i and the one-hot encoding of s_i.

4 Results

4.1 Qualitative analysis

Before fine-tuning For the model before fine tuning, we observed that while the initial sentence would reflect the star rating, soon after, the output would contradict itself in opinion of what is being rated. Also, on prompts where it is left open ended what is being reviewed, oftentimes the output will mention reviewing something that isn't reviewed on Yelp, such as other Yelp reviews, or books. Below are some examples of outputs from GPT-2, with the italicized being the z passed in, and the star rating being 0 and the tone being "serious".

> *I am so disappointed* in this food, and the food is quite bad!! I've been here in the past two weeks, and haven't stopped and order food...which means it is all great, all the time. The only thing I have changed my mind is to buy from an old fashioned "Sour Diesel" and order from an unpretentious restaurant with great service and good food.

> *I am so disappointed* in this piece!!! I would like to see another review be on this book. I can already see what this review would be, and would like to be that very person! Thank you so much!!!

Though, having the initial prompt start with "I am in awe of" and supplying a star rating of 4 seemed to consistently lead to positive reviews. However, the positive nature of these outputs seems to be independent of the star rating mentioned in the initial prompt, as supplying a star rating of 0 still results in a positive review such as what follows.

> *I am in awe of* the beauty and sophistication of this place. Not only has the food great but the staff was professional or even very helpful with tips. The bartender is always good! They also have a pretty wide range of hot and cold drinks as well as a wide choice of cocktails. If you are looking for a tasty and refreshing treat, this could be your time. If you want high quality drinks and excellent service, this can be your favorite place. They are great if you want some dessert.

Furthermore, beginning the review with a neutral input and varying the star rating does not seem to have an affect on the output's tone, as it will typically veer positive or negative only depending on what the initial review is set to. Both of the following outputs come from the same starting text of the review, a star rating of 1, and a serious tone.

> *This restaurant is* amazing. It is a dive or a diner's place, it is nice to sit and chill and do not have to work too hard to make up for your days of getting busy with school...It will bring you a little taste of you to a more exciting level. I wish I had stayed with this. I'll never go back.

> *This restaurant is* completely filled with people, and that's really scary. They will wait up until 5:55 at midnight for something new to be served. They are totally open until 10 p.m. when they serve cocktails or toasts for customers in the dining hall. I had a small group of 6 in one room, which was crowded with tables, and people sitting around my table couldn't stop thinking what exactly it was. I am actually a big fan of these people and their food

Trying longer initial reviews didn't seem to improve the coherence of the outputted reviews:

> *Rose tea is great value for the price, especially for a college student. There's usually a 10 min wait if you order at the counter so I suggest ordering ahead of time. The fried chicken over rice is* one of my favorites. It cooks very well and the sauce is very good as you roll it in the rice. My only question is did it last longer than 2 hours if you had to wait or if it came out of the same container. Overall a good deal and I'll definitely...

It is unclear how the second-to-last sentence relates to the previous sentences about rice and fried chicken.

After fine-tuning The model after fine-tuning generated more coherent reviews as each sentence related to the overall message of the review. For example, after passing in the initial review "I am so disappointed in", 1 stars, and a serious tone, the model generated the following reviews:

> *I am so disappointed in this place.* I feel compelled to write a review with a thesis that the place is either understaffed or has been a complete disaster for years. To me, their chicken salad at the front counter looks like it's from a frozen chicken dumpling bowl, topped off with a frozen chicken salad, and tossed with a very greasy frozen chicken dip. I had no idea the fresh salad was so greasy, but they managed to leave a taste to go with it.
>
> *I am so disappointed in this place.* Although they did have good food and decent service, the food was below average. The cole slaw was soggy, the gyro was a piece of shank wrap (like some sort of Asian shank wrap with cole slaw still in it) and the cole slaw was mediocre at best. On top of that, I ordered the salad and the pizza which were subpar. To be fair I ordered a salad from
>
> *I am so disappointed in this place.* The only reason I stay here is so I can dine on the latest and greatest Chinese food on tap. The restaurant is really over crowded. The wait staff are all very apologetic. The only reason I don't stay here is because I wanted to find something fresh, trendy, trendy and delicious that wouldn't cost me money and no one would care about it. I actually found it in my salad box (for about 10 dollars
>
> *I am so disappointed in this place.* They are cash only. It's not even known if the drive thru is available or if they serve pizza via air. It's cash only. We sat at 1 PM and they said it will take another 30 minutes for it to refill. And that's what happened. The waiter said "we asked for you to put your order out because it IS jammed after we make reservations" and proceeded to "oh" wait a minute...

From inspection, we see that all the output reviews are highly specific to a culinary setting, and stay on topic. They address items like specific menu items and staffing that a typical Yelp reviewer might write about.

However, the model is still somewhat unreliable at generating a review matching a certain sentiment given seemingly "contradictory" input text. For example when asked to generate a review with 1 star and the starting text "I am in awe of ", the model generated the reviews:

> *I am in awe of* this place. The decor is fantastic; however, it is totally ridiculous for what it is. It is reminiscent of a real brothel with high ceilings and massive ceilings, and is dark, dingy, and very very, very old (in fact, the place itself is so old that the curtains would go flying, possibly in your face the second time). I had three meals here for dinner the other day, and I would call it mediocre
>
> *I am in awe of* this wonderful salon. My last visit was about a year ago and for 400 I felt I had earned an appointment. I called on an Avis with a recommendation and was instantly hooked when they greeted me on the phone, how attractive you are, and if there was any confusion they would assist with any questions. They quickly brought me in and I was ready to roll with my information. Since then they have expanded, they have more space
>
> *I am in awe of* Primanti Bros Market District. It is amazing! The employees are so knowledgeable, you can tell they are listening and they are making sure their products are getting quality testing. On top of that they also have a really friendly customer service team to help you in any situation. I have had many of the items that I have ordered at Primanti Bros market and they are always fresh and labeled. The variety of their frozen meats, cured meats,

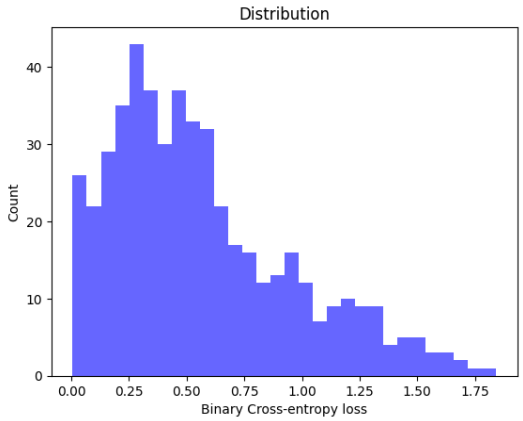

Figure 1: Pre-trained GPT-2 results. Average: 0.565

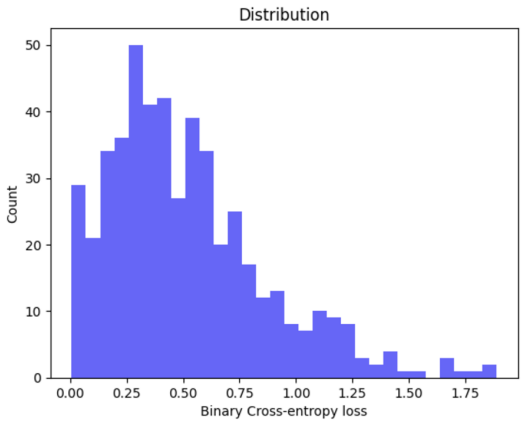

Figure 2: Fine-tuned model results. Average: 0.467

> *I am in awe of* the wonderful vegan restaurants in Pittsburgh; this review is NOT about them. My family loves veg restaurants; we've also eaten at many great restaurants, from Pittsburgh to NYC, but this is the most wonderful restaurant we've ever eaten at. Even though it's a super-small space, I can count on every guest (even the most casual veggie lover) to ask questions about what their favorite dish of all time is and when it's coming out. We've talked about

Out of the four reviews, only the first one is a negative review. Perhaps starting with the phrase "I am in awe of" is confusing the model as the phrase is often associated with other positive sentiments. It is interesting that fine-tuning did not significantly improve this issue.

4.2 Quantitative analysis

Over all 500 samples, which were identical across the pre-trained GPT-2 model and our fine-tuned model, we computed and plotted the probability distributions as described in Section 3.3. See Figure 4.2 for the histogram of the pre-trained model and Figure 4.2 for the histogram of the fine-tuned model.

As shown in the figures, our fine-tuned model results in a 20% reduction in cross-entropy loss across the same test examples. This shows that although it was only trained over 3 epochs, the model was able to learn how to output reviews consistent with the provided sentiment. An alternative way to quantify the performance would be to count the number of times the classifier's output probability distribution is heaviest on s_i, but this would force reviews to conform to one of five buckets and is not

an accurate cost model for how effective the auto-complete is, as sentiment is in reality a continuous spectrum.

5 Discussion and analysis

There is a noticeable improvement in the quality of the output after fine-tuning the model on Yelp reviews for our target task of auto-completing reviews. The outputs more consistently focus on the type of location being reviewed as opposed to accidentally veering off into talking about other matters that could theoretically be reviewed such as books, and additionally, have more salient compliments/complaints related to the type of location being reviewed, and come up with these compliments/complaints given how positive the star rating is. However, auto-completing reviews requires the model be able to infer qualities of the location being reviewed based on the initial review, which is a demanding task for GPT-2. While our fine-tuned model manages to stay on topic, it risks introducing commentary on the location that is not true.

Another issue was GPT-2's confusion for inputs involving initial prompts typically associated with positive emotions and negative sentiments. When given a positive prompt, that seemed to overpower the sentiment input, as GPT-2 would usually give positive reviews even after fine-tuning. To fix this problem, it necessitates having a robust understanding of the English language, as the model must know how to use a positive sounding phrase in a way that doesn't detract from the negative meaning. For example, using "I am in awe of" in an ironic way, would have worked but this is probably too much to ask from GPT-2. We suspect any quantity of fine-tuning won't help, and a more sophisticated model like GPT-4 is needed in order to better understand the English language.

For future work, we could consider fine-tuning a more complex langue model, such as a larger GPT-2 model or GPT-4. We could also invest more effort into in-context learning and creating a good prompt to improve the quality of our output. We could also consider incorporating retrieval-augmented generation. For example, a RAG system could potentially mitigate some of the confusion over contradictory prompts by retrieving examples of reviews that contain similar contradictory or nuanced sentiments. Additionally, our ability to fine-tune was limited by our lack of computational resources, so we were forced to use a small subset of the Yelp dataset.

References

David Ifeoluwa Adelani, Haotian Mai, Fuming Fang, Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. 2020. Generating Sentiment-Preserving Fake Online Reviews Using Neural Language Models and Their Human- and Machine-Based Detection. In *Advanced Information Networking and Applications*, Leonard Barolli, Flora Amato, Francesco Moscato, Tomoya Enokido, and Makoto Takizawa (Eds.). Springer International Publishing, Cham, 1341–1354.

Jieh-Sheng Lee and Jieh Hsiang. 2020. Patent claim generation by fine-tuning OpenAI GPT-2. *World Patent Information* 62 (2020), 101983. `https://doi.org/10.1016/j.wpi.2020.101983`

NLP Town. 2023. bert-base-multilingual-uncased-sentiment (Revision edd66ab). `https://doi.org/10.57967/hf/1515`

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

Judith van Stegeren and Jakub Myśliwiec. 2021. Fine-tuning GPT-2 on annotated RPG quests for NPC dialogue generation. In *Proceedings of the 16th International Conference on the Foundations of Digital Games* (Montreal, QC, Canada) *(FDG '21)*. Association for Computing Machinery, New York, NY, USA, Article 2, 8 pages. `https://doi.org/10.1145/3472538.3472595`

Alex Wang and Kyunghyun Cho. 2019. BERT has a Mouth, and It Must Speak: BERT as a Markov Random Field Language Model. arXiv:1902.04094 [cs.CL] `https://arxiv.org/abs/1902.04094`

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems* 28 (2015).

A Appendix / supplemental material

In the code submission assignment on Gradescope, there is the Jupyter Notebook we worked in, a README, and a JSON file containing 50 texts generated by our model, including the original review from the test split of the Yelp dataset, the true label (star rating), and the loss calculated by the fine-tuned BERT classifier.