NLP Lab 8 - Generating reviews given class and rating

Neil de la Fuente, Joan Samper, Daniel Vidal

May 2024

Abstract

In this project we use a review dataset that includes reviews of both houses and automobiles, we use it to train a model that aims to generate reviews tailored for a given rating and class. This could be used for automatic review generation for companies to be used in platforms such as tripadvisor, google maps...

1 Introduction

In this report, we practiced the application of T5 for review generation to automatize this task. Leveraging its learnt properties from language modeling, we will explore how T5 can generate coherent and contextually relevant reviews that encapsulates the sentiment and type taking into account the overall score of the reviews. For this task the model receive the overall score and the type of product in order to fine-tune all the model parameters to generate plausible reviews.

2 Pipeline

This assignment was divided in three main parts.

1. First we analyzed, processed and understood the data for then visualizing some properties.
2. Then, we built a model to generate reviews based on rating and class.
3. Finally, we evaluated the model qualitative and quantitatively.

During this assignment we had a great time applying NLP techniques, data analytics, preprocessing, and much more for a cool and very useful end.

3 Data Understanding and Visualization

The dataset contains **378k reviews**. A few of them where Null reviews so they were removed (less than **300 Null reviews**). There were **49k duplicate reviews** (12%) and **174k duplicate summaries** (46%), since most of the summaries contains a few words like "five stars" or "well done" the summaries have high probability to be duplicated. The number of reviews that were the same as the summary is 4778.

In the following you will see the distribution of the reviews and summary lengths (Nº of characters, not tokens), and the distribution of the overall scores:

- Mean: 4.43
- Median: 5.00
- Mode: 5.00
- Standard Deviation: 1.07

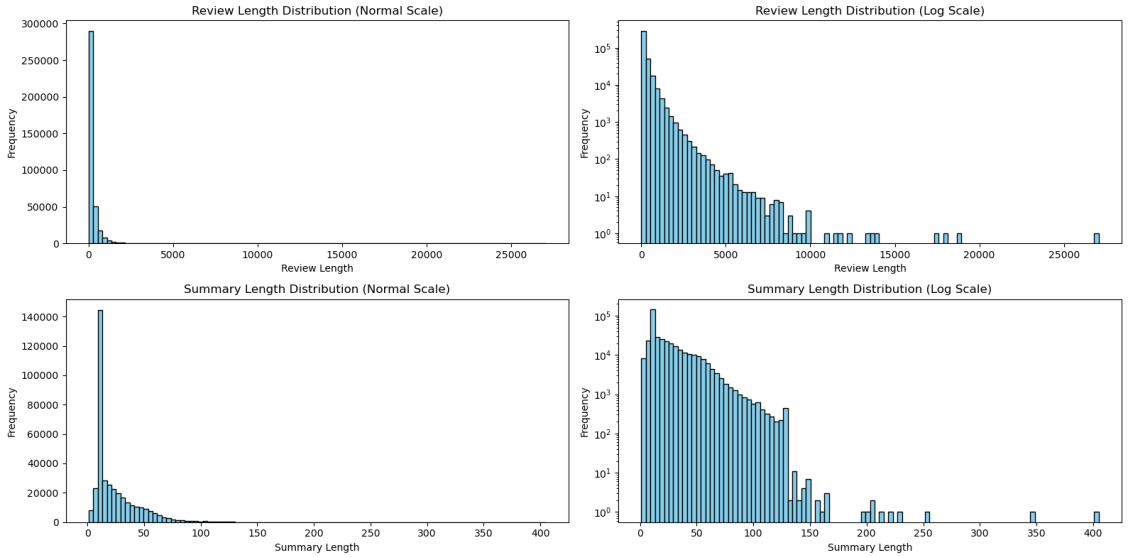

Figure 1: Reviews and Summaries length distributions

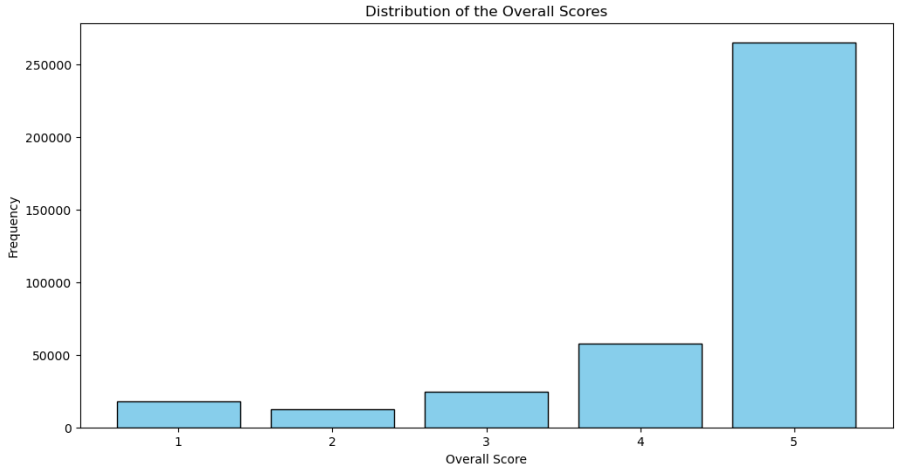

Figure 2: Overall Distribution

4 Preprocessing

We tokenized all the text from the reviews using the tokenizer from T5 that already provides the tokens with the corresponding ID's for the model and we apply a one hot encoding to the overall score and category columns

5 Model Training

The model training was conducted using a fine-tuned T5 model, which is known for its effectiveness in text generation tasks. We utilized the Hugging Face Transformers library to leverage the pre-trained 'google/mt5-small' model due to its balance between performance and computational efficiency.

5.1 Training Configuration

The training was set up with the following parameters:

- **Optimizer:** AdamW with a learning rate of 2e-5.
- **Batch Size:** 16 for both training and validation phases.
- **Epochs:** We trained the model for 5 epochs, monitoring the validation loss for early stopping to prevent overfitting.
- **Loss Function:** Cross-entropy, as it is standard for classification and generation tasks.

5.2 Training Process

The training process involved feeding tokenized text into the model, where the input included encoded overall scores and categories, and the output was expected to be a coherent review. The model learned to generate text that aligns with the given score and category, adapting the language model's weights accordingly.

5.3 Challenges

During training, we faced challenges related to:

- Balancing the dataset to prevent bias toward more frequent categories.
- Managing GPU memory constraints, which were mitigated by adjusting the batch size and using gradient accumulation.

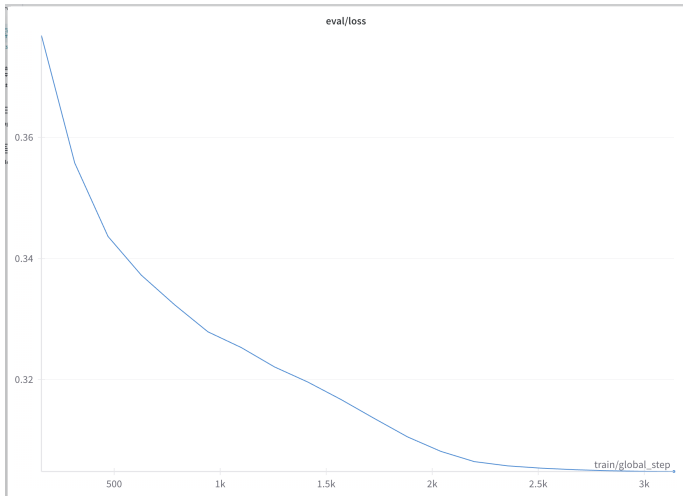

Figure 3: Validation Loss

6 Model Evaluation

Post-training, the model was evaluated on a held-out test set to assess its ability to generate reviews that are coherent, relevant, and correctly reflect the specified rating and category.

6.1 Quantitative Evaluation

We used the following metric to quantitatively evaluate the model:

- **ROUGE Scores:** To measure the similarity between the generated reviews and a set of reference reviews. We used Rouge-1, Rouge-2 and Rouge-L.

6.2 Qualitative Evaluation

Qualitative analysis involved manually reviewing the generated text to assess:

- **Relevance:** *Does the review pertain to the correct category and rating?* In general, the reviews generated by the model were aligned well with the specified categories and ratings. This indicates that the model effectively understands and applies the input parameters to produce relevant content, it may seem an easy task but still the model is able to generate varied results for each of the inputs, which was the goal.
- **Coherence:** *Are the reviews logically structured and grammatically correct?*

  Even though the model is not Miguel de Unamuno and sometimes has errors due to the lack of extensive training, it is generally grammatically correct and keeps coherence and structure as well.
- **Creativity:** *Does the model produce diverse and contextually rich reviews?* The model showed the ability to produce diverse and contextually rich reviews. It was not limited to repetitive or generic phrases but instead offered varied and interesting comments, enhancing the usefulness and engagement of the reviews. This was mainly thanks to dropout.

Examples:

For a given score of 4 and category Home the review returned was: Works well. The ground truth is: Works well as advertised.

For a given score of 1 and category Home the review returned was: not as expected. The ground truht is: Replaced existing heater/fan which was the same model as this one. The first one was satisfactory, so I bought this Broan 655 again to save having to retrofit another model in the ceiling. The noise is enough to make on wear ear plugs. Dreadful sounding, especially when the heater is turned. Wasted money on this purchase. Will not use this brand again. Would not recommend this particular model.

7 Conclusions

This project demonstrated the effectiveness of the T5 model in generating contextually relevant and coherent reviews based on specified product categories and user ratings. The fine-tuning of the T5 model on our dataset resulted in a capable system that can potentially be used to automate review generation for various applications, such as content creation for digital marketing or assisting in generating user feedback.

Key takeaways include:

- The T5 model, even in its smaller configuration, is robust enough to handle the nuances of text generation across different contexts and sentiment levels.
- Proper preprocessing, such as one-hot encoding of categorical variables and tokenization, significantly contributes to the model's performance.
- Challenges like data imbalance and computational constraints can be managed effectively with careful dataset preparation and training configuration.

In conclusion, the project fulfills its intended objective of automating review generation and also provides a foundation for further research and development in the area of natural language processing and artificial intelligence.