Impact of Text-to-Image and Image-to-Text Transformation on Yelp Review Sentiment

Michael Perry
Carnegie Mellon University
mperry2@andrew.cmu.edu

Cem Adatepe
Carnegie Mellon University
cadatepe@andrew.cmu.edu

Maxwell Soh
Carnegie Mellon University
msoh@andrew.cmu.edu

Abstract

This paper investigates the impact of text-to-image-to-text transformations on sentiment analysis using Yelp reviews. We fine-tune a DistilBERT model, achieving an accuracy of 95% on the original text reviews for binary sentiment classification. Subsequently, we generate images that encapsulate the emotional content of these reviews using DALL-E 3, followed by text generation from the images using GPT-4o. Our findings reveal that while the accuracy of sentiment classification decreases when utilizing images as intermediaries, dropping to 75%, the overall sentiment remains discernible. This leaking of sentiment can be attributed to signal loss over the course of transformation and the systematic positive bias introduced by enterprise models, affecting the integrity of sentiment representation when utilizing these models. This study contributes to understanding the challenges and potential of multimodal approaches in sentiment analysis, highlighting areas for future research to enhance model reliability and performance.

1 Introduction

Understanding sentiment in text data is crucial for many applications, including review aggregation and brand monitoring. In this work, we explore the use of image-based transformations to analyze how the sentiment of Yelp reviews is impacted throughout text and image generation. By transforming reviews into images and generating a corresponding review-like description, we aim to study how well sentiment is preserved through this multimodal approach. An in-depth investigation into this process could yield relevant results in the fields of synthetic data generation, commentary on the use of transformers in processing data, and insights into the bias of multimodal models.

In general, passing data through any model will introduce noise, making the preservation of sentiment less expected. However, certain models may have the ability to amplify the sentiment, which leads to the opposite effect. Thus, this study seeks to address the following questions: How does sentiment change during these transformations? What implications do these changes have for applications relying on accurate sentiment analysis? By examining these questions, we aim to contribute valuable insights into the effectiveness of multimodal approaches in maintaining sentiment integrity across different forms of data.

2 Summary of Data

The Yelp Polarity Dataset consists of user reviews labeled as positive or negative, which is ideal for sentiment classification tasks. The dataset is balanced, with an equal distribution of positive and negative reviews. These text reviews are used to train a transformer for binary sentiment classification. For our experiment, we use original text data and also transform the reviews into images using Dall-E 3. These images aim to capture the emotional content and sentiment of the reviews. We then

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

process the images and generate text reviews using GPT-4o, allowing us to explore the influence of image-to-text transformations on sentiment analysis.

One theme of this paper is the inherent positive bias of enterprise models. GPT-4o and Dall-E 3 are typically optimized to avoid producing outputs that could harm user experience or reflect poorly on associated businesses. Thus, in the event of ambiguous inputs, like an unemphatic review that could lean positive or negative purely based on language, these models will tend to default to positive sentiment outputs. Due to this design choice, each further transformation will likely add more positive bias to our generated reviews. This bias becomes clear in our results and influenced our prompt-engineering when generating the transformed data.

3 Methods

3.1 Sentiment Classification Fine-Tuning

3.1.1 Model Selection

Our goal was to select a model that could achieve reasonably high performance on at least the original dataset. Pre-trained transformer architectures in recent years have shown remarkable ability to process language and model complex sentences, making them a prime candidate for sentiment classification fine-tuning. Due to open source and compute constraints it was only feasible to use off-the-shelf HuggingFace models in the range of 10 million to 1 billion parameters. We ultimately settled on DistilBERT and tielectra-small, parameter-efficient pre-trained transformers with around 66 million and 14 million parameters, respectively.

3.1.2 Fine-Tuning

We began the fine-tuning process by first preparing the dataset. We took a subset of the entire dataset, creating a split with train, validation, and test set sizes of 8,000, 2,000, and 1,000, respectively. We trained the models using gradient descent with AdamW optimizer and binary cross-entropy loss. Weight decay was effective in counteracting overfitting, though due to the limited size of the subset of data used (as a result of compute constraints), hyperparameter optimization was relatively straightforward.

3.2 Image Generation

We used Dall-E 3 to generate images that capture the sentiment of the Yelp reviews on a subset of the dataset separate from those used in the train, validation, and test splits for fine-tuning. We tried open-source models like Stable Diffusion 2.1 and Dall-E 2, but these did not generate images of high-enough quality. While Stable Diffusion 3.5 Medium would have provided us with high-enough quality images but was too big to run.

3.3 Text Generation from Images

From the generated images, we created review-like descriptions of the image. We turned to captioning models like BLIP, CLIP, and uform-gen2-qwen to provide captions of these images. Preliminary results showed that BLIP and CLIP were ineffective at generating captions that captured any sentiment at all, oftentimes only generating several word descriptions of the elements located in the image. uform-gen2-qwen was more expressive and left more sentiment in its description but notably suffered from a strong positive bias that we were not able to completely eliminate even with a negative biased prompt. We ultimately used GPT-4o to generate the text that would be used to evaluate our fine-tuned transformer, since we found it to be the most consistent, expressive, and objective of any model that was used.

3.3.1 Prompting

With uform-gen2-qwen, we used the following unbiased prompt "Give a detailed description of this scene with reference to lighting, coloring, and scenery." for image generation from a review. To attempt to mitigate the inherent positive bias, we also tested the following negative-biased prompt: "This scene likely depicts a very unpleasant situation. Based on this, give a detailed description of

(A)
Original Review:
This is not a good spot if you like clean modern hotel rooms. I upgraded to an executive suite and the couch is so nasty as well the ceiling has the popcorn style. If I didn't already change one room today I would actually ask to be changed again. This is more of a large motel.

(B)
Image of Review:

(C)
Unbiased P1 Response: Positive

Unbiased P2 Response:
The image features a spacious living room with a large, colorful sign that reads "BEVERAGE SULT" on the wall. The room is furnished with a brown sofa, two armchairs, and a coffee table. The room is illuminated by two lamps, one on each side of the room. The floor is covered in a dark gray carpet, and there are curtains on both sides of the room. A green plant is placed near the window, and a large yellow sign is hanging on the wall.

Biased P1 Response: Negative

Biased P2 Response:
The scene depicts a room with a large, yellow sign that reads "BEVERLY SULF" on the wall. The sign is positioned above a large, brown sofa and a brown armchair. The sofa is adorned with a white pillow and a brown armchair is positioned next to it. The room is illuminated by two lamps, one on each side of the room. The floor is covered in a dark gray carpet. The walls are painted in a light brown color. The room has a window with beige curtains, and there is a green plant in a pot near the window. The overall atmosphere of the room is gloomy and unpleasant.

Figure 1: Example of workflow results. (A) An original negative review. (B) Dall-E 3 generated image of the review using our text-to-image prompt. (C) UForm-Gen responses to our 4 image-to-text prompts based on the generated image. The P1 and P2 prompts are described in sections 3.4 and 3.3.1 respectively. The unbiased results describe the situation positively while the biased prompts led to a description with negative sentiment.

it with reference to lighting, coloring, and scenery". These prompts are referred to as P2 in Figure 1. Although negatively biasing the prompt led Uform-Gen to generate negative descriptive of some images based on negative reviews (such as the one in Figure 1), most were still described positively or neutrally. GPT-4o did not suffer from the same positive bias issue nearly as much, and we had success with the following unbiased prompt: "I am going to give you an image generated from a Yelp review. Write me a paragraph review trying to capture the sentiment based on aspects like coloring, scenery, tone, etc. Here is the image: ". We chose this prompt so that the generated review would be as descriptive as possible.

3.4 Direct Image Classification

Though uform-gen2-qwen could not give unbiased descriptions of the images for later classification via the fine-tuned transformer, we found that it could be used to directly classify an image as either positive or negative. We evaluated it with an unbiased and biased prompt in accordance with the early finding of consistent positive bias. The prompts we used were "Is the sentiment of this image positive or negative?" and "This image comes from a set of mostly negative images, is its sentiment positive or negative?". These prompts are referred to as P1 in Figure 1.

3.5 Evaluation Metrics

The performance of all models is evaluated using accuracy and we pay particular attention to false positives, as they appear to be the primary type of misclassification in our experiments.

4 Results

4.1 Text Sentiment Classification

The fine-tuned DistilBERT model achieves an accuracy of 95% on the original Yelp review data with scores seen in the leftmost matrix in Figure 2. This strong performance confirms the model's ability to effectively classify sentiment in the original text data. This performance is not particularly noteworthy, but it will serve as a baseline on which to compare future classification. The fine-tuned

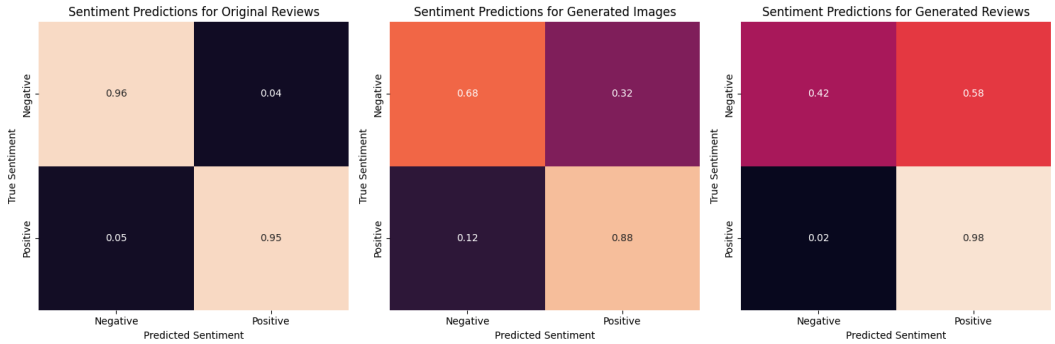

Figure 2: The confusion matrices for sentiment prediction based on Left: the original reviews, Center: the DALL-E 3 generated images of the reviews, and Right: the GPT-4o generated reviews of the generated images. The image generation and review generation steps both led to significant increases in the false positive rate (top right quadrant of each matrix).

tielectra-small achieves a meager 65% on the original Yelp review data and so we did not use it in further analysis on transformed reviews.

4.2 Image Sentiment Classification

We used Uform-Gen to classify the DALL-E 3 generated images of the original reviews with the biased version of P1. Even though this prompt is negatively biased, we observe a significant increase in the false positive rate compared to the classification of the original reviews (Figure 2), which was the primary cause of overall accuracy decreasing to 79%.

4.3 Generated Review Sentiment Classification

When we generate reviews from the DALLE-3 images using GPT-4, the sentiment classification accuracy drops to 75%. The majority of these misclassifications are false positives, suggesting that while the generated reviews retain the general sentiment, some subtle nuances may be lost or misinterpreted during the transformation process.

5 Discussion and Analysis

The results indicate that the process of transforming reviews into images and then back into text does lead to a noticeable drop in classification accuracy, though not as extreme as we may have initially expected. We believe this can most likely be traced back to two sources: positive biasing and signal loss.

Positive biasing during image generation and image description seemed to be rampant. Our fine-tuned classifier on the original dataset shows no sign of biasing more towards false positives or negatives, yet using uform-gen2-qwen to directly classify the generated images gives rise to a 32% false positive rate compared to 12% false negative rate. This disparity between false positives and false negatives is further widened after using our fine-tuned classifier on the reviews generated from the images (58% false positive vs 2% false negative). This seems to indicate that each step of the data transformation (text to image back to text) injects more positive bias into the underlying sentiment.

Signal loss and differences in model performance is also another important consideration here since even the false negative rate of uform-gen2-qwen on generated images increased by 10% from the baseline false negative rate of our fine-tuned classifier on original reviews. This suggests that the transformation, particularly the text to image component, in addition to positive biasing may also fail to capture some subtleties of the original review.

Many limitations of this research project arose due to the use of pretrained enterprise models and computation limits. As stated in the Summary of Data section, the use of enterprise models likely contributed significantly to the number of false positives. Our attempts to address this biasing by

introducing counteractive bias did not change results drastically. This phenomenon of positive biasing is worth further investigation, as isolating this component in a model or pipeline can improve data integrity and make model outputs more reliable. While we wanted to investigate how nonenterprise models performed in similar transformation pipelines, we faced limitations as smaller models performed poorly and larger models were unable to run on Google Colab. Similarly, cost and compute constraints limited the size of the datasets we were able to use and generate via these enterprise models and Google Colab's GPUs.

6 Conclusion

In this paper, we have explored the use of image-to-text and text-to-image transformations for sentiment analysis. Our results show that the accuracy of sentiment classification decreases noticeably mainly as a result of positive biasing introduced by both the image-to-text and text-to-image components of our model as well as signal loss that would be expected from any transformation process.

Acknowledgments

We would like to thank our TA mentor Emily for her valuable guidance throughout this project.

References

[1] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

[2] Ramesh, A., et al. (2021). Zero-Shot Text-to-Image Generation. arXiv:2102.12092.

[3] Brown, T. B., et al. (2020). Language Models are Few-Shot Learners. arXiv:2005.14165.

[4] OpenAI. GPT-4o: Advanced Conversational AI Model. Available in free and paid modes, designed for detailed and citation-rich responses. Accessed December 2024.

[5] RunComfy. UformGen2 Qwen Node: A Multimodal AI Model Integrating Image Analysis and NLP for Contextual Responses. Available at ComfyUI. Accessed February 2024.

[6] Clark, K., Luong, M.-T., Le, Q. V., Manning, C. D. (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. arXiv:2003.10555.

[7] Stability AI. Stable Diffusion 3.5 Medium: A Multimodal Diffusion Transformer for Text-to-Image Generation with Improved Performance in Image Quality and Resource Efficiency. Hugging Face repository, accessed October 2024.