Bite-Sized Recommendations: Evaluating Abstractive and Extractive Summarization Approaches for Yelp Reviews

Jeremy Lan, Michelle Lee
University of California, Berkeley
jeremy.lan@berkeley.edu, michellelee2020@berkeley.edu

Abstract – **In this project, we evaluated diverse extractive and abstractive summarization methods using the Yelp restaurant review dataset. The goal of the project was to accurately summarize the top 10 reviews for each restaurant, to provide Yelp users with a condensed version of important business information. Extractive techniques like bag-of-words, TextRank, and spaCy were employed alongside abstractive approaches using a pre-trained BART/large-cnn model. We gauged sentiment using FLAIR sentiment analysis and quantified summarization quality through ROUGE scores. We found that although the BART abstractive model outperformed the extractive summarization models in ROUGE and FLAIR sentiment metrics, hyperparameter tuning on BART did not lead to significant improvements. We also noted the limitations of ROUGE as an evaluation metric and explored the benefits of FLAIR sentiment analysis for better capturing the quality of a review. This project contributes insights into optimizing summarization techniques and their practical application for meaningful content extraction from extensive textual data.**

I. INTRODUCTION

In today's world of information overload, abstractive and extractive NLP summarization techniques have been utilized extensively to provide individuals with key information. Yelp's business review platform presents as a unique data source as one of the most prominent platforms for user-generated reviews. Many restaurants and local businesses often have thousands of reviews about their services, and millions of users leverage these reviews to make both small and large decisions. However, the sheer volume of detailed reviews on Yelp makes it impractical for users to sift through them to form a general recommendation, and Yelp's other "generalization" metrics (5-star ratings, etc.) may not provide enough detail for users to make an informed decision.

This project aims to tackle this problem through the application of extractive and abstractive summarization techniques on the Yelp reviews dataset. Extractive summarization techniques directly identify high-value sentences or phrases from the original text, while abstractive summarization involves generating concise summaries that capture the essence of the original text while introducing paraphrasing and language generation capabilities.

Another challenge that we face when approaching this problem concerns model evaluation. The Yelp dataset contains raw user review data but does not include any reference summaries to utilize as training data. Thus we are unable to implement traditional summary evaluation metrics such as ROUGE or BLEU on a large scale for the entire dataset. Although we do manually generate a small subset of summaries to serve as labelled training data, we also introduced a sentiment analysis metric that allowed us to perform a wider analysis of model performance across the entire dataset.

Through this project, we seek to develop a robust and effective system for generating abstractive summaries from Yelp reviews. From the hundreds and thousands of reviews that each business receives, we hope to provide a concise summary of the top reviews so that an end user can receive useful information about a business as efficiently as possible.

II. BACKGROUND

As the amount of data on the Internet has grown exponentially over the past few decades, many strides have been made in the field of natural language processing (NLP) summarization. Both abstractive and extractive summarization methods are popular in current applications - abstractive summarization models form novel phrases and sentences to offer a more coherent summary, while extractive summarization models simply return important sentences from the input text (Awasthi, 2021). Depending on the model complexity and compute bandwidth required, different summarization models may be better well-suited for different appplications. In this section, we will evaluate the various abstractive and extractive summarization methods used in our study, and provide some context on the Yelp reviews dataset.

A. **Baseline "Summarization" - Bag of Words**

To serve as a baseline for our more complex summarization models, we leverage the "bag-of-words" approach, which returns keywords from the input document ranked by appearance frequency. The keywords returned from this approach may not make coherent sense as a sentence and thus act as a low-range baseline for which our summarization models can be compared to.

B. **Extractive Summarization - spaCy, TextRank**

We utilized spaCy and TextRank models for our extractive summarization task. TextRank uses a graph-based algorithm that treats text components as nodes and their interconnections as edges. Through an iterative process, it gauges the importance of each node, thus identifying the most crucial sentences within the text. On the other hand, spaCy, an open-source library, is centered around linguistic analysis and structured information extraction, encompassing features

like part-of-speech tagging and named entity recognition. Its primary focus is on delivering thorough linguistic annotations and robust features (Honnibal, 2017).

C. **Abstractive Summarization - BART**

We leveraged BART (Bidirectional and Auto-Regressive Transformer) for our abstractive summarization model. BART is pre-trained on the reconstruction of corrupted documents (denoising autoencoding) which allows BART to develop more robust representations of text and to handle more complex NLP tasks. Compared to other transformer architectures like BERT, BART's autoregressive decoder architecture allows it to be fine-tuned for sequence generation tasks such as abstractive question/answer and summarization (Lewis, 2019). On the CNN/Daily Mail dataset, which is one of the most widely used summarization datasets, BART outperformed BERT and other pointer-generator network models in both ROUGE metrics and model perplexity.

PEGASUS and T5 transformer-based models also appeared as candidate options for our abstractive summarization task given their similar prevalence in the field. However, in limited testing BART was shown to have superior performance to T5 and PEGASUS in zero-shot learning contexts, while performance was comparable across all three models in few-shot learning (Goodwin, 2020). BART was thus selected as our abstractive summarization model of choice.

Within BART, we investigated adjusting the *beam_size* and the *no_repeat_ngrams* parameter during the hyperparameter tuning process. While a larger *beam_size* would be expected to yield higher quality outputs as more output possibilities are put under consideration, this comes at the tradeoff of increased computational resources and decoding time. Leveraging a higher *no_repeat_ngrams* value can result in less redundancy and more diversity in the summarization output.

Although abstractive summarization models consistently show improved performance compared to extractive summarization models substantially more compute resources are often required for abstractive models (Mahajani, 2019). We take these compute limitations into account in our experimentation methods and performance evaluation approach.

D. **Yelp Review Dataset**

The Yelp Business and Review data used in this project was provided by Yelp in .json format, containing 7 million reviews for 150 thousand businesses (Clark, 2023). To streamline this data for our model, we only included restaurants that were still listed as open and that had accumulated 50 or more reviews since 2018. Thus, any lower-volume restaurants where review data might be more sparse and less reliable were excluded, while we only retained reviews post-2018 as their more recent information would be more valuable to a customer in the present day. The dataset includes the number of "Funny", "Useful", and "Cool" upvotes that each review received from other users.

To include the context from multiple impactful reviews, the top 10 most-upvoted reviews were selected and concatenated to serve as our "input" data. The top-voted review for each restaurant had an average length of 203 words, while the average length of the top 10 reviews concatenated had an average length of 1602 words. In total, 10318 restaurants met the above criteria.

E. **FLAIR: Sentiment Evaluation**

Due to the lack of training data for our dataset (no pre-written summaries of Yelp reviews), other evaluation metrics were needed to gauge model performance. Accurate sentiment analysis is an extremely important aspect of any summarization task, as a good summary (extractive or abstractive) should have a similar positive or negative sentiment to the original input text. We thus leverage the FLAIR NLP framework, which provides a harmonized interface for contextual string embeddings (Akbik, 2019). This allows researchers to combine various embeddings (BERT, ELMo, GloVe, etc.) with ease without falling victim to the computational complexity of any one embedding.

FLAIR's sentiment analysis module is based on a character-level LSTM network and is tuned to output not only a positive/negative sentiment, but also a numerical score (-1 to 1) regarding the intensity of the sentiment. Since this sentiment analysis can be run on both the input text and the generated summaries, we are thus able to obtain some level of comparison of performance between summarization models across the entire dataset without the need for training data.

During the design process it was also considered to leverage the "star" rating provided in a user review as an evaluation metric - a model could be trained to assign each generated summary a "star" rating which could be compared to the original input. However, this option was discarded due to heavy class imbalance in the Yelp dataset's star ratings (Figure 1), and use of an external sentiment classifier was preferred.

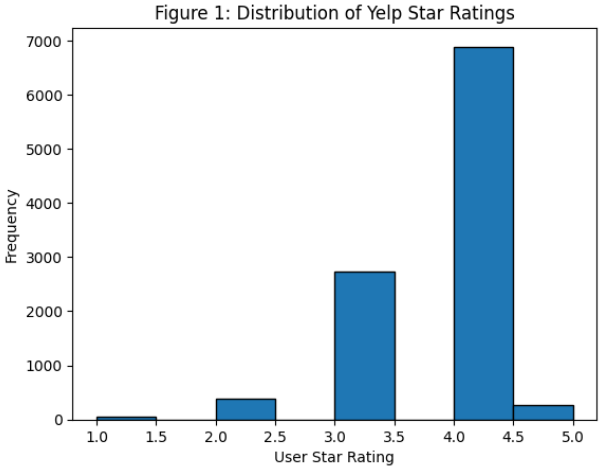

Fig. 1. Distribution of Yelp Star Ratings

III. METHODS

The Yelp data was first prepared as stated in the background section, the final concatenated output resulted in 10318 rows,

with the top 10 reviews from each restaurant concatenated into one string.

Since no training data summaries were available for our dataset, we leveraged ChatGPT 3.5 (OpenAI, 2023) to assist in manual summary generation for 100 restaurants ($\sim$1% of our dataset). ChatGPT was instructed to generate an abstractive summary of the input text with approximately 10% of the input length (an input of 2000 words would be condensed to $\sim$200 words). This subset of reviews would serve as our groundtruth to assist in model evaluation and hyperparameter tuning.

We were able to leverage the full dataset to run Bag of Words baseline analysis, which returned the top 5 words that occured in the concatenated review string. TextRank and spaCy extractive summarization models were run on default settings, with the models instructed to only return the top 3 input sentences.

The majority of our model tuning efforts went towards optimizing our BART abstractive summarization model. Having identified the beam size and the *no_repeat_ngrams* value as hyperparameters with a likely impact on model performance, we ran a GridSearch testing the beam size of (1, 3, 5), and *no_repeat_ngrams* lengths of (3, 5, 7).

Due to compute limitations with running a complex BART model and the lack of training data, the GridSearch was only performed on the same 100 restaurants where ChatGPT had generated training summaries, thus allowing us to evaluate the hyperparameters using ROUGE. The *max_length* parameter for BART was also set to 10% of the input review length to attempt to mirror the length of the training reviews.

For all summaries generated (Bag of Words, spaCy, TextRank, BART), FLAIR sentiment analysis was applied which provided every summary with a positive/negative sentiment as well as an intensity score (-1 to 1). FLAIR was also applied to the input concatenated reviews to provide a groundtruth for comparison. To evaluate model performance, we computed precision, recall, accuracy, and F1 scores using the positive/negative labels from FLAIR sentiment, as well the MSE between the groundtruth intensity scores and those of the generated summaries. For the subset of rows where a training summary was generated, ROUGE-1, ROUGE-2, and ROUGE-l scores were computed.

An example is provided below presenting the generated summaries for a specific restaurant. The input text (10 concatenated reviews) and the ChatGPT reference summary are shown in Appendix A and B.

> **Bag-of-Words:** "['coffee', 'place', 'great', 'would', 'delicious']" (word-count: 5)
>
> **spaCy:** "My mom and I have been going around tucson trying a bunch of different coffee places during this quarantine and this place is SOOOOO good!! Their coffee tastes fresh, the service is fast, and it's generally such a great place. The girl that was working and I had a good laugh together about how we always drink iced coffee." (word count: 59)
>
> **BART with beam_size=1, no_repeat_ngrams=3:** "Tucson Coffee Roasters is a great place to drink coffee. The coffee is balanced and the syrups are made in-house. The atmosphere is great and the coffee is delicious. The service is great. The food is good." (word count: 37)
>
> **TextRank:** "Their coffee tastes fresh, the service is fast, and it's generally such a great place. They will do great with that location. They do the coffee with a cop." (word count: 29)

IV. RESULTS AND DISCUSSION

A. Mean ROUGE Scores Comparison: BART GridSearch

We used the pre-trained BART-large-cnn model and generated summaries for the same set of one hundred rows for differing parameters of *beam_size* and *no_repeat_ngrams*. Then we compared those model-generated summaries to summaries generated by ChatGPT as shown in Appendix B.given the concatenated inputs (Table 1, Figure 1)

TABLE I
ROUGE Scores for BART GridSearch

beam_size	no_repeat_ngrams	rouge-1	rouge-2	rouge-l
1	3	**0.2413**	**0.0778**	0.1641
1	5	0.2353	0.0738	0.1657
1	7	0.2342	0.0737	0.1669
3	3	0.2374	0.0602	**0.1673**
3	5	0.2342	0.0586	0.1661
3	7	0.2340	0.0585	0.1662
5	3	0.2357	0.0602	0.1614
5	5	0.2369	0.0617	0.1628
5	7	0.2371	0.0621	0.1636

We observed that the distributions of ROUGE scores were fairly similar across the *beam_size* and *no_repeat_ngrams*. The lack of improvement in model performance with increased *beam_size* may be due to the redundant nature of the concatenated reviews in the input document. Since many reviews from Yelp users may include similar topics (multiple people liking the same food or service, etc.), there is likely substantial redundancy. The further exploration that the model performs with a larger *beam_size* thus may not yield as diverse a set of outputs, thus offsetting any expected improvements. Similarly, the lack of change in model performance with adjustments in *no_repeat_ngrams* may be due to the inherent nature of the dataset, with each review already having a limited occurrence of repetitive n-grams. As we are implementing an abstractive summarization model with a substantial document length reduction, BART is not likely to repeat a 5-gram or 7-gram with standard settings. Thus setting an upper bound with *no_repeat_ngrams* may not have a substantial effect, while setting *no_repeat_ngrams* below 3 runs the risk of the model not being able to discuss common 2-grams like food items or restaurant names.

Low variability in ROUGE scores in the GridSearch may also be attributed to the small sample size of 100 samples - in future work expanding this analysis across a full-size validation set would yield more robust insights, but would require a fully labeled dataset.

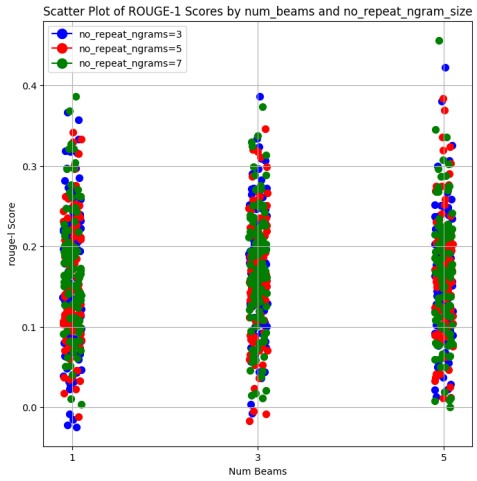

Fig. 2. Scatterplot of Rouge-L scores by num_beams and no_repeat_ngrams

B. Mean ROUGE Scores Comparison: BART vs. Extractive Summarization

Despite the lack of significant variability between our BART models in the hyperparameter tuning process, we did see a marked improvement in ROUGE performance when comparing our BART models to our extractive summarization models (spaCy and TextRank), as shown in Table 2.

TABLE II
ROUGE Scores for BART vs. spaCy and TextRank

Model	rouge-1	rouge-2	rouge-l
BART (1,3)	**0.2413**	**0.0778**	**0.1641**
TextRank	0.0782	0.0068	0.0583
spaCy	0.2220	0.0452	0.1302

Because our 100 reference summaries from ChatGPT are written abstractively, it makes sense that extractive summarization models would perform poorly, as the reference summaries do not contain many string segments from the original text. If our reference summaries had purely been written extractively it would likely result in higher ROUGE scores for spaCy and TextRank compared to BART. This discrepancy highlights some of the limitations of ROUGE as a metric as well as the importance of understanding the context and style of the reference summaries from which these models are evaluated (Ng, 2015).

C. FLAIR Sentiment Evaluation

Given the limitations of ROUGE evaluation and the small quantity of reference summaries for our dataset, the FLAIR sentiment analysis shows promise as we were able to perform this analysis across the entire Yelp review dataset.

FLAIR was used to assign each of the input text with a "POSITIVE" or "NEGATIVE" sentiment as well as a sentiment intensity from -1 (very negative) to 1 (very positive). Using these data as groundtruth, the sentiment label and magnitude was also calculated for each of the Bag-of-Words, spaCy, and TextRank summaries for the entire dataset. BART abstractive summarization was only run on 100 summaries, however we were also able to calculate FLAIR sentiment labels and magnitudes to serve as a comparison against the reference summaries (BART run with *beam_size* of 1 and *no_repeat_ngrams* of 3).

From comparing the sentiment labels from the input text with each of the summarization methods, we are able to present precision, recall, accuracy, and F1 scores. We are also able to compare the sentiment magnitudes between the input and the summary texts to generate a mean-squared error (MSE) quantifying the difference between the groundtruth and the model-generations, as an additional metric of prediction accuracy (Table 3).

TABLE III
FLAIR Sentiment Evaluation Metrics

Model	Precision	Recall	Accuracy	F1 Score	MSE
BoW	0.65	0.69	0.69	0.66	1.1
TextRank	0.72	0.68	0.68	0.7	1.17
spaCy	0.7	0.65	0.65	0.67	1.28
BART (1,3)	0.83	0.79	0.79	**0.81**	**0.76**

The results from table 3 must be put in context with the sentiment baseline of 74% (7659 of the 10318 input rows were classified with positive sentiment). Consequently, it can be seen that only BART abstractive summarization produced evaluation metrics exceeding the baseline, with both strong precision and recall. Due to the disjointed nature of the text string returned by the Bag of Words model making it difficult to accurately capture sentiment, it understandably performed below the baseline. When discussing the poor performance of the extractive summarization methods, one must understand the format of the source text. Since the input text is a concatenation of multiple reviews (which can be both positive and negative), and the extractive summarization methods only return a limited number of sentences that are deemed most important, the mixed sentiment in the input may not always be captured in the output thus leading to misclassifications (if reviews contributing negative sentiment are not returned in the output). On the other hand, BART abstractive summarization is more well-suited to capturing mixed sentiment by reconciling conflicting sentiments from multiple sources in the input text, when compared to spaCy and TextRank. Likewise, the improved accuracy for BART compared to the other models also resulted in a lower MSE.

Given that the distribution of FLAIR sentiment values is only from -1 to 1, the MSE values recorded are quite high. This is partly due to the extremely tight distribution of scores around the poles of -1 and 1 (see figure 3 for TextRank example) - thus, any review that is misclassified would end up on the far opposite side of the scoring spectrum, leading to a high MSE.

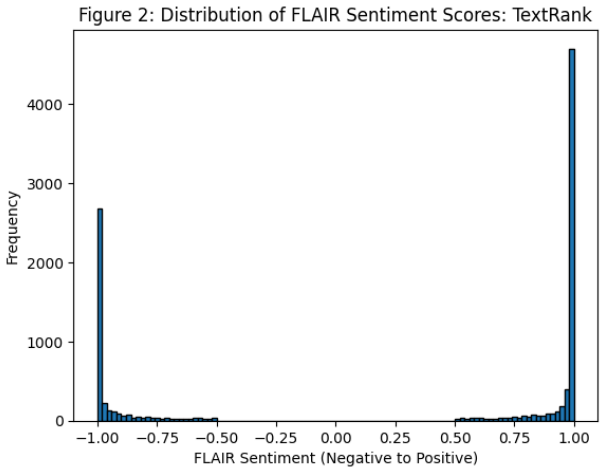

Fig. 3. Distribution of TextRank FLAIR Sentiment

It is important to consider that BART was only applied towards 100 rows of the input text due to computational limitations. Replicating this improvement on a larger portion of the entire dataset would improve the robustness of these results, especially if the sentiment evaluation metric is to be valued equally with more traditional evaluation metrics such as ROUGE.

D. **Limitations and Future Work**

Our main limitations in this project concerned the lack of training data as well as the lack of sufficient compute resources to run BART on a larger scale. We believe the FLAIR sentiment analysis is able to provide additional quantitative support, but the lack of reference summaries do affect the robustness of our findings. Ultimately, summary evaluation is quite subjective and even reference summaries can vary widely in length, quality, and writing style. If Yelp or another organization would plan to implement such a model, a valuable metric could instead be the feedback from customers on the usefulness of these summaries in recommending products or businesses.

Compute limitations also affected our ability to run BART on a larger scale. Each BART summary run took on average 1 minute to complete on the systems available to us (Google CoLab with default GPU), thus making it infeasible to apply BART to the entire dataset of over 10,000 rows. This limitation also affected our ability to tune our BART model with a wider GridSearch of hyperparameters and limited us to only tuning on 100 rows. The obvious candidate for future work would be the implementation of BART across the entire dataset with additional compute power - combining that with a reference summary for each row would improve robustness of results. If reference summaries were available, we would have also been able to leverage additional evaluation metrics like METEOR and CIDEr, which further account for word order, synonyms, and agreement with multiple references.

Another option to bypass the lack of reference summaries would be to tune our BART model on a smaller labeled large-language-model dataset (SQuAD Q&A Dataset would be a feasible candidate). Evaluating a hyperparameter GridSearch on another labeled dataset would also allow us to leverage previous work performed on these datasets as a starting point for a tuned model. Our Yelp dataset would still lack the reference summaries to evaluate performance with ROUGE; however it would be interesting to compare the FLAIR sentiment metrics we obtained with those from a fully tuned BART model.

V. CONCLUSION

In this project, we performed a comprehensive exploration of extractive and abstractive summarization techniques applied to Yelp restaurant reviews. Our goal was to generate insightful summaries from these reviews to help users in their decision-making process.

We leveraged voting data to select top 10 reviews per restaurant and used the concatenated string as the text input. Summaries generated from extractive methodologies, including bag-of-words, TextRank, and spaCy, and abstractive strategies involving a pre-trained BART/large-cnn model were evaluated through FLAIR sentiment analysis, providing an integral facet of assessment, and ROUGE scores, facilitating a quantitative measure of summarization quality.

In our experiment, we found the BART model, particularly with *beam_size* set to 1 and *no_repeat_ngrams* set to 3, to consistently outperform other methods across various metrics. The model exhibited superior ROUGE-1 and ROUGE-2 scores, along with robust FLAIR sentiment analysis results, accentuating its proficiency in capturing the essence and sentiment of the original text. However, hyperparameter tuning was overall inconclusive with small variations in ROUGE performance - additional compute power would have enabled a wider GridSearch across a larger validation set.

Moreover, our study underscores the importance of considering the intricacies of data context and characteristics when evaluating summarization models. The limitations of existing metrics, such as ROUGE, became evident in the context of abstractively written reference summaries, where traditional extractive models faltered due to misalignment between source and generated content.

In conclusion, our project brings insights into the optimization of summarization techniques and practical strategies for deriving meaningful insights from textual data. For future work, we want to explore other pre-trained abstractive summarization models and fine-tune on a larger set of data to see if we can further improve performance. Ultimately, Yelp review summarization presents as an intriguing business-case with the potential for significant improvements to customer experience.

REFERENCES

[1] I. Awasthi, K. Gupta, P. S. Bhogal, S. S. Anand, P. K. Soni, *Natural Language Processing (NLP) based Text Summarization - A Survey*, 2021, doi: 10.1109/ICICT50816.2021.9358703.

[2] S. Clark, *Yelp Review Dataset*, Retrieved from https://www.yelp.com/dataset, 2023.

[3] M. Honnibal, I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing", , 2017, to appear.

[4] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension", , 2019, 1910.13461.

[5] T. R. Goodwin, M. E. Savery, D. Demner-Fushman, "Flight of the PEGASUS? Comparing Transformers on Few-Shot and Zero-Shot Multi-document Abstractive Summarization", *in Proceedings of COLING. International Conference on Computational Linguistics*, pp. 5640–5646, 2020.

[6] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, R. Vollgraf, "FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP", *in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pp. 54–59, Association for Computational Linguistics, Minneapolis, Minnesota, Jun. 2019, doi:10.18653/v1/N19-4010, URL: https://aclanthology.org/N19-4010.

[7] A. Mahajani, V. Pandya, I. Maria, D. Sharma, "A Comprehensive Survey on Extractive and Abstractive Techniques for Text Summarization", *in* Y.-C. Hu, S. Tiwari, K. K. Mishra, M. C. Trivedi, eds., *Ambient Communications and Computer Systems*, pp. 339–351, Springer Singapore, Singapore, 2019.

[8] OpenAI, "ChatGPT (Aug 3 Version)[Large language model].", .

[9] J. Ng, V. Abrecht, "Better Summarization Evaluation with Word Embeddings for ROUGE", *CoRR*, vol. abs/1508.06034, 2015, URL: http://arxiv.org/abs/1508.06034, 1508.06034.

[1] [2] [3] [4] [4] [5] [6] [7] [8] [9]

VI. Appendix

A. **Example Input Text - 10 Concatenated Reviews**

Uh, yummy? Delicious? Amazing service?!? What else can I say! The coffee is balanced AND they make their syrups in-house. I got the caramel macchiato, not too sweet, not too milky, just how I love my lattes! My goal is to try their whole menu cause I'm sure it'll all be delicious. Customer service was also on point, friendly and helpful. He even gave us some extra stamps! Don't forget to bring your stamp cards! They are also pup friendly, gotta wear your masks and make sure to social distance. Tables are spread out 6ft apart which is a plus and they also have outdoor tables as well. Delicious cold brew! I love checking out local coffee shops, especially during these hard times. The cold brew is delicious! I will definitely be returning! This place is great! The staff was so friendly, it caught my eye when they first opened and I saw the store owner standing outside with a "Free Coffee" sign. I decided I would stop to try it on my way to school. It was great! The atmosphere was very comfortable as well. They do the coffee with a cop. They had their coffee for sale in cute little bags for the holidays. I found the coffee smooth and not to bitter. I drink cold coffee all year young so iced is always for me. The girl that was working and I had a good laugh together about how we always drink iced coffee. They have punch cards which is a nice touch. I could definitely see myself coming in here to hang out and drink coffee. I wish they had a few more tables and maybe some more food like pastries etc. They will do great with that location. I have lived in Tucson all my life. Almost 43-gritting through my teeth and summer-years. I hadn't heard of this place!!

Two of our closest local coffee shops recently closed. I wanted a local place that roasted beans to have at home. I searched on IG and came across this gem!

Awesome coffee! Great breakfast tacos!! Excellent spot to chat and read. Amazing date spot. My husband and I are hooked! So glad I found This was a really cool location to visit for my first stop this morning for a cup of coffee. I was able to walk to this location from my hotel stay, was greeted when I walked in the door as well as had a chance to view a little bit of the menu and the surroundings on the inside. The associate working the counter she was helpful she asked me if it was my first time and I informed to her yes. So she proceeded to inform me of what they offer and also a few suggestions on what I possibly might like. So I ordered a small cinnamon crunch coffee. Was very delicious and I was able to enjoy my coffee while sitting outside on their seating patio. We just moved to an apartment close by and heard that this locations was pretty good so we wanted to give it a shot. Living within walking distance we decided to take a walk to the Tucson Coffee Roasters this morning. Upon entering we were greeted by a polite barista who made us feel welcomed. The atmosphere was great, it would be a place for anyone who wants to come to study or even sit and to relax. I didn't know what to try out but was recommended by the barista to try it out. So I decided to try the whisky aged coffee in a size medium. The portion size was not bad and for the price it was reasonable. I suggest getting it with sugar since it would be bland. The aroma on the other hand is splendid, it remind me of when I was in Chicago trying it out for the first time. I recommend giving this place a try if you're on this side of town! I got a prickly pear iced tea here with zero prickly pear in it. I asked if it would be sweet and they said no, but there was cane sugar at the cream and sugar station. I just wish they would have also told me it was going to be plain tea.

That said, the inside is pretty nice. They have high ceilings, a couch, interesting art work, and a take-a-book-leave-a-book policy.

If my wife had liked what she ordered, I might have raised the rating. Maybe the straight up coffee is good, but they probably need to take the decaf and tea options off the menu. I hate to give places such a low rating, but unfortunately Tucson Coffee Roasters did not deliver on my particular order. I got an iced decaf latte with oat milk. I was so excited to see they had oat milk, as I'm trying to reduce my dairy intake. I love oat milk since it's usually super smooth and creamy. My drink was anything but, and had a watered down kind of grainy taste. I'm wondering if that's because it was decaf? Not sure but many other coffee shops have made delicious decaf lattes. I couldn't finish my drink and that's a pretty bad sign. That being said, the interior is super cute and modern and would be a great place to study or get some work done. I'm a frequent visitor to this pleasant and high quality coffee House. They're roast coffees are very very fine and the coffee house itself has much indoor seating and a small amount of outdoor seating. Covid regulations seem to be pretty well observed. My mom and I have been going around tucson trying a bunch of different coffee places during this quarantine and this place is SOOOOO good!! Their coffee tastes fresh, the service is fast, and it's generally such a great place. If you haven't tried it, I would go today!!! (word count: 994)

B. **Example ChatGPT Summary: (90% Length Reduction**

Tucson Coffee Roasters offers delicious coffee with amazing service and a comfortable atmosphere. Their caramel macchiato and cold brew are highlights, and they have pup-friendly outdoor tables. The staff is friendly and offers helpful suggestions. However, some find the iced tea lacking in flavor, and the decaf latte with oat milk was disappointing. Overall, it's a great local spot with high-quality roast coffees, indoor and outdoor seating, and good adherence to Covid regulations. Customers praise their breakfast tacos and recommend trying their menu offerings. It's a gem for coffee lovers and a great place to relax, study, or chat. (word count: 99)