

國立清華大學資訊工程系 112 學年度 上 學期專題報告

專題名稱	Transforming Product Reviews through Summarization, Key Aspect and Sentiment Analysis				
參加競賽或計畫	<input type="checkbox"/> 參加對外競賽	<input type="checkbox"/> 參與其他計畫		<input checked="" type="checkbox"/> 無參加對外競賽或任何計畫	
學號	109006202	109006211	109006210	109006240	109006236
姓名	陳鮑比 Bobby Setiawan	貝仁義 Kenneth Edgar	施平治 Dominick Matthew Singgih	劉其生 Jansen Reynaldi Gautama	鄭聰明 Jason Theodorus Pratama

摘要

For our System Integration Implementation, our group decided to research and implement a project regarding Product Summarization. Creating an automated product summary, sentiment analysis and aspect extraction through machine learning can be done with a deep learning model known as transformers. Transformers provided us the ability to capture contextual relationships between words or tokens in a sequence, to pre-train and fine-tuned our model with large amounts of data to suit our likings and also an encoder-decoder structure.

We tried some models like Pegasus, BERT, and T5 for summarizing tasks. Then, we shifted our project to Large Language Models such as Llama 2, Mistral, and Zephyr where all the parameters are 7B(**billion**) and increased the range of our transformers research to include sentiment analysis and key aspects of a product review in addition to just the summary

Our scope of this project is to provide summarization of reviews, the extraction of sentiment and the product aspects. Providing the compact review of certain products, or a brand generally can be advantageous for both sellers and buyers. Therefore, by deriving the data from summarized product reviews, sentiment analysis, and product aspects can enhance customer or user experience and the quality of product and services.

中華民國112年11月

System Integration Implementation II Report
Transforming Product Reviews through
Summarization, Key Aspect and Sentiment Analysis

Bobby Setiawan, Dominick Matthew, Jason Theodorus, Jansen Reynaldi Gautama,
Kenneth Edgar

bobbysetiawan1304@gmail.com, dominick.m.s12nthu@gapp.nthu.edu.tw,
jasontheodorus22@gmail.com, jansenrg@gapp.nthu.edu.tw,
kenneth.edgarr@gmail.com



NATIONAL TSING HUA UNIVERSITY
HSINCHU, TAIWAN

TABLE OF CONTENT

TABLE OF CONTENT.....	1
1. Introduction.....	3
1.1. Abstract.....	3
1.2. Background.....	3
2. Workflow and Discussion.....	4
2.1. Time Table.....	4
2.2. Discussion.....	6
2.2.1. Large Language Model vs Small Language Model.....	6
2.2.2. Dataset.....	6
2.2.3. Evaluation Metric Choices.....	7
2.2.4. Model of Choice.....	7
2.3. Related Works.....	8
2.3.1. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization.....	8
2.3.2. Aspect Extraction Approach for Sentiment Analysis Using Keywords...8	
2.3.3. Instruction-Tuning LLMs for Product Title Summarization.....	8
2.3.4. Building Real-World Meeting Summarization Systems using Large Language Models: A Practical Perspective.....	9
2.3.5. Sentiment Analysis in the Era of Large Language Models: A Reality Check.....	9
2.4. Identified Problems.....	10
2.4.1. Issues with Dataset.....	10
2.4.2. What happens without fine tuning?.....	10
2.4.3. Resource limitations.....	13
2.4.4. Can the result be used right away?.....	13
3. Methodology.....	15
3.1. Dataset For Summarizing Reviews.....	15
3.2. Dataset for Sentiment and Aspect (Key Point Interest).....	15
3.3. Fine Tuning Tools and Training Parameters.....	16
3.4. Model Sharding.....	18
3.5. Inference / Generate.....	18
3.5.1. Preparing The Dataset.....	18
3.5.2. Creating the Prompt.....	19
3.5.3. Setting up the inference parameters.....	20
3.5.4. Post-processing.....	20
3.6. Evaluation Method and Performance.....	21
3.6.1. ROUGE.....	21
3.6.2. SacreBLEU.....	21
3.6.3. BERTScore.....	21

4. Result.....	22
4.1 First Model : Mistral 7B.....	22
4.2 Second Model : PEGASUS.....	26
4.3 Third Model : Zephyr 7B.....	29
4.4 Fourth Model : Llama 2 7B.....	33
5. Performance.....	37
5.1. ROUGE.....	37
5.2. SacreBLEU.....	37
5.3. BERTScore F1.....	39
5.4. Training and Validation Loss.....	40
5.5. Data Analysis.....	40
5.5.1. Synonyms.....	41
5.5.2. Conclusion: Human Analysis.....	41
5.5.3. Conclusion: Quantitative Performance.....	41
5.5.4. Human Analysis vs Quantitative Performance.....	42
6. Real World Implementation.....	43
6.1. Possible Implementation.....	43
6.2. Visualization.....	43
7. Conclusion.....	47
Bibliography.....	48

1. Introduction

1.1. Abstract

For our System Integration Implementation, our group decided to research and implement a project regarding Product Summarization. Creating an automated product summary through AI and machine learning can be done with a deep learning model known as transformers. Transformers provided us the ability to capture contextual relationships between words or tokens in a sequence, to pre-train and fine-tuned our model with large amounts of data to suit our likings and also an encoder-decoder structure. These advantages render it our preferred choice for summarization tasks.

This report will contain our progress of the System Integration II, where we continued during the summer break of 2023, up until the end of the duration of System Integration II, which is in November. Our progress we've made is to use new models that can better fit our purpose of the System Integration, and find out the strengths and weaknesses of other models. We shifted our research from generative transformers like Pegasus and BERT into Large Language Models such as Llama 2, Mistral, and Zephyr where all the parameters are 7B(**billion**). We have also increased the range of our transformers research to include sentiment analysis and key aspects of a product review in addition to just the summary.

1.2. Background

Natural Language Processing has been a popular topic nowadays, it can be implemented in many fields. Our consideration of our application is into E-Commerce. Online Shopping has been a trend for the past decade. On the customer side, everyone has the right to make educated decisions based on available reviews and product claims to receive the product or service they require. On the other hand, for brand owners and manufacturers, products may have countless reviews, yet the time-consuming nature of examining them to choose a quality product hinders perspectives of efficient decision-making as they could not achieve a compiled result of user experience.

In response to this issue, our scope of this project is to provide summarization of reviews, the extraction of sentiment and the product aspects. Providing the compact review of certain products, or a brand generally can be advantageous for both sellers and buyers. Therefore, by deriving the data from summarized product reviews, sentiment analysis, and product aspects can enhance customer or user experience and the quality of product and services.

2. Workflow and Discussion

2.1. Time Table

Table 2.1.1 Meeting TimeTable

Date/Time	Contents
2023 Spring Report	https://docs.google.com/document/d/1IREb6-huyjg1NW245LHvH-Ha0yw61B-Jyx6csx62wGo8/edit
5th July 2023	<ul style="list-style-type: none">- Discuss about Sentiment Analysis- Decide to make our own Dataset, each person write 1000 summary for the dataset based on Amazon
14th July 2023	<ul style="list-style-type: none">- Attempt at sentiment analysis-Discussing the goal- Training model and do further research- Start training with the dataset we have created- Make more datasets
31st July 2023	<ul style="list-style-type: none">- Attempt at sentiment analysis- Training model and do further research
Rest of Summer	<ul style="list-style-type: none">- Make datasets- Do research
9th September 2023	<ul style="list-style-type: none">- Find LLM to do the target task
13th September 2023	<ul style="list-style-type: none">- Decision to use LLAMA 2.- Register LLAMA 2 (meta) and HuggingFace's LLAMA 2- Consider using SageMaker for LLAMA 2 Fine-tune - Next meeting will try to use LLAMA 2 Model if it were accepted
15th September 2023	<ul style="list-style-type: none">- Try and Trial using Llama 2- Fine tune Llama 2 <p>Agenda for next meeting:</p> <ul style="list-style-type: none">- Dataset for Product Specification.- Dataset for Sentiment Analysis- Finetune Llama 2 using Trainer
16th September 2023	<p>https://www.anyscale.com/blog/llama-2-is-about-as-factually-accurate-as-gpt-4-for-summaries-and-is-30x-cheaper</p> <p>This means we should use Llama-2-70b or gpt-4 to increase the chances of a factual summarization (in the same ballpark as humans). gpt-4 was slightly better than human, Llama-2-70b slightly worse. Llama-2-7b and Llama-2-13b had issues following the task instructions; but we used another LLM to interpret their output. They had ordering bias issues.</p> <p>Work on sentiment:</p>

	<ul style="list-style-type: none"> - Single Sentiment Options: "Positive", "Negative", "Title". If we use the 70B, we need to use the school's computer. Aspect can be For next Dataset, preparet: - Aspect table - Title
20th September 2023	<ul style="list-style-type: none"> - Scrapping Title for Dataset <p>Plan for Future: Try to do Sentiment</p>
22th September 2023	<p>Make Datasets: Deadline: October Finished the Product/Program Dataset (Aspect): 40 per day.</p> <ul style="list-style-type: none"> - From Text (Original), not from Summarization - Format: using ", " - Determine Aspect (Pos,Neg) If there is no aspect of the respective aspect, the cell is inserted as 'x' Aspect amount up to you.
25th September 2023	<ul style="list-style-type: none"> - Continue Dataset - Training Llama 2 (7b-chat-hf, 7b-hf)
27th September 2023	Debugging Lama2
30th September 2023	<p>Inference Debugging</p> <p>Plan to create short dataset</p>
2nd October 2023	<p>Inference Debugging Plan:</p> <ul style="list-style-type: none"> - Offtime -> Do the Dataset - Testing Sweetspot, Inference, Dataset, How to get Evaluation
4th October 2023	Do Dataset
6th October 2023	<p>Dataset, Debugging Inference</p> <p>Plan : Try SSH</p>
7th October 2023	<p>Trying the SSH</p> <p>Train 13B Llama2 (Crash)</p> <p>Plan : Find other better models</p>
16-22nd October 2023	<p>Training: Mistral 7B vs LLama 7B, Zephyr 7B, Pegasus;</p> <p>Research: LORA, PEFT, Llama2 (Architecture)</p> <p>Analysis on Results</p> <p>Do some improvement if possible</p>

	Final report
23rd October 2023	Meeting with Advisor
25th October 2023 - end	Implementing F1 BERTScore and Finalizing Report, PPT, and Poster

2.2. Discussion

2.2.1. Large Language Model vs Small Language Model

From our discovery in System Integration I, we found out that using abstractive means of summarization proves to be more effective for Product Review Summarization, as it allow for more flexibility in rewriting, thus we thought that since we are asking the machine itself to create their own text, this opens up the possibility of using multi-function LLM to test out the difference in their capability compared to older models which are streamlined to do a specific task.

The main benefit and also a new challenge that we have to face using multi-functional LLM is that we now need to tinker around prompts and find out how prompting works and affect the output of the model. In our previous older models, all we had to do was just input the desired text to be summarized, and expect the result directly.

Another aspect we would like to tackle was sentiment analysis, but, we did not want to just do an overall sentiment of a certain review, but rather try to understand the specific ups and downs of the reviewed item. We think that to do both summarization and sentiment analysis at once, using a multi-functional LLM is much more preferable than using multiple smaller models that would end up taking a lot of resource usage, which can be managed by a single LLM model.

2.2.2. Dataset

Initially, in our progress on System Integration I, we discovered that Amazon provides product descriptions and articles. The structure of each product listing is similar, with the product title already containing a brief summary of its specifications. Consequently, we scraped the "Product Description", "About this Item", and "Title." We used these to train the machine learning models we employed for System Implementation I, namely PEGASUS, T5, and BERT. We used the "Title" as our target output.

In System Integration II, we made adjustments to our dataset. Since we aimed to shift our summarization focus to "user reviews" and include Sentiment Analysis and Key Point Interest Extraction, we had to create new columns such as:

- Aspect(Pos) Which contains positive aspect about this product
- Aspect(Neg) Which contains negative aspect about this product
- Sentiment Which contains overall sentiment of the product
- Summary Which contains the summarized version of the original text (30-50% of the original text)

We manually asserted these new data in order to achieve the most from our model.

To obtain the mentioned data, we implemented a simple web scraping code using Python libraries like **BeautifulSoup**, **urllib.parse**, and **Pandas**. We utilized requests to fetch **HTML content** and **BeautifulSoup** to parse it. We stored all the scraped data in an array and leveraged **Pandas** to convert the array into a .csv file.

2.2.3. Evaluation Metric Choices

There are a couple of evaluation metrics that we have considered, this included **Rouge**, **SacreBLEU**, **BERTScore** and **Human Evaluation**. We opted to use **Rouge** for summary evaluations and produce a score based on the quotient of matching words. **SacreBLEU** is initially made for translation evaluations, but it can also serve a similar purpose to **Rouge**, that is to produce a score based on the quotient of matching words in its simplest form, While **BERTScore** is a metric that evaluates the similarity or semantic equivalence between the generated text and reference by using contextual embeddings.

Human Evaluation is also done in order to assess the accuracy of the summary and the scores that are yielded from Rouge and SacreBLEU.

2.2.4. Model of Choice

In System Integration I, we tried a few smaller models such as T5, BERT, and PEGASUS. in our experiment we found that PEGASUS gave us the best result for our objective.

In System Integration II, we expanded our model of choice to Llama 2, a newly open-source LLM available from *Meta* to achieve our abstraction summarization goal. However, due to the constraint of resource, it is only possible to use the 7B parameter for Llama 2.

Since Llama 2's introduction, there have been notable other open-source models that came out recently which are Mistral 7B and Zephyr 7B. Both Mistral and Zephyr have been reported to have outperformed Llama 2 in their respective fields. Mistral 7B claims to outperform Llama 2 13B on all of its benchmarks while Zephyr 7B is reported by HuggingFace to outperform Llama 2 70B on MT-Bench. All in all, these models have the advantage of having more performance with a small parameter size in comparison with Llama2 7B.

We also decided not to use any “Chat” version of any aforementioned models, as we do not need the bot to refer to previous inference history that might affect the results of following inputs. (Trying to avoid what ChatGPT does when it starts to read older texts instead of your wanted instruction).

We decided to try out all of these models instead of picking the best one according to other claims, to create a more realistic result, as our task is more streamlined for summarization and sentiment analysis, where the latter benchmarks for all purposes.

2.3. Related Works

2.3.1. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization

In the paper discussed by Jingqing Zhang, Yao Zhao, Mohammad Saleh, Peter J. Liu, they discussed a novel way for summarization by using an encoder decoder transformer model, where they use a new training method by using gap sentences into the Transformer model to create summarization. This paper was the kickstarter of our project, but unlike the paper, our paper does not focus on the training implementation of the Transformer model, but more into implementing the framework of how summarization, key point, and sentiment analysis for use of product review summarization.

2.3.2. Aspect Extraction Approach for Sentiment Analysis Using Keywords

Sentiment Analysis done in this paper by Nafees Ayub, Muhammad Ramzan Talib, Muhammad Kashif Hanif, and Muhammad Awais, utilizes an Aspect-Based sentiment analysis, where the approach aims to improve the aspect term extraction and identify the sentiments about the aspects expressed. Where the steps consist of removing stop-words and punctuations, tokenization, stemming, parts-of-speech tagging, keyword extraction, aspect and opinion extraction and sentiment classification.

In our approach, we used fine tuned large language models to do the sentiment analysis and aspect extraction. Rather than the paper keywords-based approach to identify aspects from online reviews.

2.3.3. Instruction-Tuning LLMs for Product Title Summarization

In the paper discussed by Besnik Fetahu, Zhiyu Chen, Oleg Rokhlenko, Shervin Malmasi, Instruction Tuning on LLMs specifically for Product Title Summarization where they tried some Instruction to generate summaries following what their desires such as Specificity, Length, Phrase Inclusion and Removed Word Constraints. and we use the same method to constrain our model to output a result in a desired length of words and ask the model to maintain the user experience in first person perspective.

2.3.4. Building Real-World Meeting Summarization Systems using Large Language Models: A Practical Perspective

In the paper discussed by Tahmid Rahman Laskar, Xue-Yong Fu, Cheng Chen and Shashi Bhushan, they researched on how to effectively build a meeting summarization system for real world usage using Large Language Models. Their summarization system differs from our research project , where we implement product review summarization and their paper focused on meeting summarization.

This paper performs evaluation , correlation and comparison between LLMs such as GPT-4 and LLaMA-2. This paper concludes that even LLMs that were trained with much smaller parameters, such as LLaMA-2 7B and 13B, can still compete against larger closed source models. The evaluation performed in this paper served as one of our motivation for the selection of models that we are going to choose.

This paper also discussed two of our concerns, which are the cost and inference speed. LLaMA-2's GPU cost has caused several issues in the earlier part of our project, this paper mentioned that they were able to run LLaMA-2 7B model using lesser processing power than the 13B model, however they still performed (both inference speed and response) at a similar level. This conclusion led us to stick with the 7B version for all our LLMs.

2.3.5. Sentiment Analysis in the Era of Large Language Models: A Reality Check

Sentiment analysis has been done on transformer based models such as BERT. With the current surge of LLMs, There is a need to check the capabilities of LLMs in regards to their performance in sentiment analysis. This paper by Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing provides an insight regarding several LLMs performance such as T5 and GPT3.5 on their performance in sentiment analysis while also providing comparison to small language models (SLMs). In their result, it is stated that LLMs outperform SLMs in the field of sentiment analysis as it achieved near-human performance

This paper provides insight that LLMs provide more opportunity and greater capabilities compared with SLMs.

2.4. Identified Problems

2.4.1. Issues with Dataset

Datasets which are available on the internet oftentimes have little to no use for our purpose. Most of the available dataset found in the internet are not able to express details about certain objects. Below is a simple example of the dataset available that we found.

Table 2.3.1.1. Example of Useless Dataset

Review	McCann's makes oatmeal for every oatmeal connoisseur, whether one likes it from the raw pellet state that cooks for half an hour, to the sloth addled instant, which can be done in the microwave for under three minutes. It's all good, that's for sure, and the beauty of the instant variety is that it is available in different flavors as well as regular. This variety pack allows different tastes to be explored, as well as giving you a chance to experience the difference between McCann's and other well-known oatmeals. What I personally like about McCann's is that it cooks up thicker and with more body than the top brand here in America. The Apples & Cinnamon, though, tends to be a little liquidy so you may want to experiment with the amount of water you add. In my 1300 watt microwave the oatmeal cooks up in about one minute and twenty-seven seconds, so you should also watch that to get a handle on how much time and water to use. The only bad thing -- if you can consider it a bad thing -- about this offering is that you have to buy in lot so you'll end up with six ten-count boxes. This is good if you have a whole family of oatmeal-eaters, but if you're a single person alone -- well, love oatmeal.
Summary	Oatmeal For Oatmeal Lovers

As shown above, the reviewer explains the product in depth, where they discuss the flavor, cooking method, and his thoughts regarding the amount he gets from it. But the existing summary for this data does not explain anything about his views on the food. It's just an oversimplification that would not work for a detailed summary. Refer to **3.2 Dataset for Sentiment and Aspect (Key Point Interest)** for the modified dataset.

2.4.2. What happens without fine tuning?

Generally, finetuning is required for any Large Language Models that are used for specific purpose, it is commonly used in various machine learning applications, and one of them is natural language processing.

Without finetuning, the base model is outputting random words that are completely unrelated to the generation that satisfies our needs. There are some cases where the models also repeat the prompts that we gave

to the model. Another problem is that the model is posing the user with a question back. Mainly, the length of the summaries doesn't meet the expectations.

Here are some examples of the defects that the base models generated :

Table 2.3.2.1. Base Models Issues

Example 1	<pre># 1.0.0 def summarize_text(self, text): """ Summarize the following Amazon reviews into 70 to 100 words while maintaining the user experience, and write it in first person perspective Also, find the overall sentiment of the following text, and show me the important aspects of the text. I was somewhat dubious when I ordered these sour belts as a Christmas gift for my (candy-addicted) younger brother, as I have repeatedly had my heart broken in the past by sour belts that were decidedly... UN-sour.(Nothing is worse than craving the acidic, tangy *pop* of sour candy, only to have your taste buds duped by a too-sweet, sugary let-down of flavor.)Well imagine my delight when these sour belts slapped my tongue with tartness like a cranky old granny at a Church luncheon! Just as sour as the ones you find at the mall candy shops, and incredibly convenient, as they come in large tubs.Totally delicious and totally recommended.Beware, though: Since there are so many belts in a single tub, you may find yourself over-indulging and burning yourself from an excess of the citric acid on these babies... If this happens, you'll be left with a raw, stinging tongue that may take a couple of days to heal. Granted, this only happens if you eat like, thirty-five in a single sitting, but still... it's possible. I speak from (shameful and disgusting) experience.Still, when eaten in moderation, these are a great snack, a great replacement for love, a great way to gain weight, and a great way to spend your time!Grab them by the fistful and shove them in your face!!! """ # TODO: Implement summarize_text() pass</pre>
Example 2	<pre>def sentiment_analysis(self, text): """ Find the overall sentiment of the following text, and show me the important aspects of the text. """</pre>

	<pre> # TODO: Implement sentiment_analysis() pass def sentiment_analysis_2(self, text): """ Find the overall sentiment of the following text, and show me the important aspects of the text. """ # TODO: Implement sentiment_analysis_2() pass def sentiment_analysis_3(self, text): """ Find the overall sentiment of the following text, and show me the important aspects of the text. """ # TODO: Implement sentiment_analysis_3() pass </pre>
Example 3	<p>When I ordered these sour belts, I was skeptical. In the past, I've been disappointed by sour candies that were too sweet and lacked the tangy, acidic flavor I crave. But these sour belts are a game-changer. They're just as sour as the ones you find at mall candy shops, and they come in large tubs, making them incredibly convenient.</p> <p>The first time I tried them, I was blown away by the tartness. It's like a cranky old granny slapping your tongue with sourness at a Church luncheon! They're delicious and highly recommended.</p> <p>However, beware: Since there are so many belts in a single tub, it's easy to overindulge and burn your tongue from the excess citric acid. I've learned this the hard way, and it's not a pleasant experience.</p> <p>Overall, these sour belts are a great snack, a great replacement for love, and a great way to gain weight. Just grab them by the fistful and shove them in your face!"</p>
Example 4	<p>Comment: I'm not sure what you're asking. You're asking for a summary of a review, but you're not giving any context. What is the review about? What is the context? What is the summary? What is the summary of the summary? What is the summary of the summary of the summary? What is the summary of the summary of the summary of the summary? What is the summary of the summary of the summary of the summary?</p>

As previously stated, these inference results are not fit for use. The text can be barely understood both by the machine or for us humans to use. Therefore, finetuning is required for our purpose in generating the summary, extracting the sentiment, and key point of interests or aspects.

2.4.3. Resource limitations

Training a Large Language Model requires significant computational resources which poses significant challenges. Smaller GPU utilization would allow us to do more training in parallel, which will speed up the training process.

Several methods that were used to reduce memory usage includes quantizing the model into a 4 bit floating point model using QLoRa. After the training process, there is still a need to save the model to be loaded in the future. Due to the memory limitation, sharding the model is required to load the model for future usage. More details will be further explained in section 3.3-*Training Parameters* and 3.4-*Model Sharding*.

2.4.4. Can the result be used right away?

While the results can be used and be evaluated directly, it has unwanted generated text that clutters our inference of the output. As such, It is desirable to add post-processing to make the outcome more digestible.

Table 2.3.4.1. Example of Before and After Post-Process

Before post-process	<p>Result: New to GF, most substitution products are okay but don't taste right. This bisquick pancake mix is a standout. Light, smooth, and delicious. Can't tell the difference, maybe slightly better. Highly recommended.</p> <p>Overall Sentiment: positive</p> <p>Aspects:</p> <p>Positive: pancake mix quality, taste, lightness, smoothness</p> <p>Negative: x, y, z, (optional), (minor complaint), (alternative suggestion)</p> <p>Recommendations:</p> <p>Highly recommended for those seeking a delicious and easy-to-use gluten-free pancake mix. Light, smooth, and tasty. Slightly better than other options.</p> <p>Overall Sentiment: positive</p> <p>Aspects:</p> <p>Positive: pancake mix quality, taste, lightness,</p>
---------------------	---

	smoothness Negative: x, y, z, (optional), (minor complaint), (alternative suggestion)
After post-process	Results: New to GF, most substitution products are okay but don't taste right. This bisquick pancake mix is a standout. Light, smooth, and delicious. Can't tell the difference, maybe slightly better. Highly recommended. Overall Sentiment: positive Aspects: Positive: pancake mix quality, taste, lightness, smoothness Negative: x

From Table 2.3.4.1, the generated text from the model can be understood and can be used to determine the quality of the product based on its review, positive aspect, and negative aspects. However, there is still unnecessary generated text that clutters the overall quality and also affects the model evaluation negatively. Therefore, the solution to this difficulty is to process the output text to reach a more desirable outcome.

3. Methodology

3.1. Dataset For Summarizing Reviews

The dataset for summarizing product reviews were a little bit easier to find. As of now, we were not able to find one that contains different categories of products. However, we did find one that contains product reviews of foods. This dataset can be found in kaggle:

<https://www.kaggle.com/code/currie32/summarizing-text-with-amazon-reviews>

One downside of this dataset would be that the summaries are not to our liking, instead we manually asserted the summary of the reviews by ourselves which is reduced by 50-70%, we could understand the whole text itself. We need to ensure that the text length is not too long as that will remove the purpose of summarizing in the first place. At the same time, we also need to ensure that the text length is not too short, as doing so will not encapsulate all the important parts of the text.

3.2. Dataset for Sentiment and Aspect (Key Point Interest)

The dataset for sentiment and aspects were nowhere to be found online for our specified use. We had to create and tweak our own dataset for this usage. We have 3 variables for sentiment, which are “positive”, “negative”, and “neutral”. We skimmed, and analyzed each text and put it in the “Sentiment” column we created before.

Key Point Interest is extracted from the text and grouped by positive and negative, because not all reviews are purely positive or negative. Therefore, we separate the positive and negative aspects. Similar to the sentiment, it is impossible to find the dataset for our specified use. We analyzed every keyword that is crucial to the text and extracted it as it is, so there will not be any bias, or misunderstanding to the key point of interest itself.

To further visualize, our dataset would look something like the following:

Table 3.2.1. Sample of Summary + Sentiment Dataset

Original Review:	I got a wild hair for taffy and ordered this five pound bag. The taffy was all very enjoyable with many flavors: watermelon, root beer, melon, peppermint, grape, etc. My only complaint is there was a bit too much red/black licorice-flavored pieces (just not my particular favorites). Between me, my kids, and my husband, this lasted only two weeks! I would recommend this brand of taffy -- it was a delightful treat.
Expected Summary:	I ordered a five-pound bag of taffy, mostly enjoyable with various flavors, though too much licorice for my taste. Lasted two weeks. Recommended.
Overall Sentiment	Positive
Aspect (Positive)	enjoyable, many flavor, delightful, treat
Aspect (Negative)	too much licorice

We like this type of dataset the most as it conveys the main idea of the review itself, while also extracting what the user enjoys and dislike about the item itself. Unfortunately with this dataset requirement, we have difficulty in creating a large quantity of dataset, but we were able to verify and create a higher quality dataset, which will perform better than higher count of dataset but low in quality.

3.3. Fine Tuning Tools and Training Parameters

We continue to use the API provided from HuggingFace to continue our fine tuning. Since our task is now to fine-tune a LLM, a new fine tuning method must be employed as the size of LLM itself does not allow to directly train the entire machine weight, and we have to find other means.

This includes trying out the new tuning architecture, which is PEFT (Parameter Efficient Fine Tuning). The PEFT method that we chose is to use QLoRA (Quantized Low Rank Adaptation) to even become more efficient in memory usage so we can train the model more efficiently. With this, we are able to freeze the majority part of the machine's original weight, while adding new weights that can be tuned for our task, and then merge them back together. This helps the machine from forgetting its own functionality from the

pre-training stage, and we are merely adding more weight to streamline it for our specific task.

The main benefit of QLoRA Training, compared to LoRA, is QLoRA are able to quantization on our model, converting our 16 bit model, into a quantized 4 bit version, which helps to train models with larger amount of parameters using less computing power, while maintaining the performance of a full 16 bit training procedure.

These are our parameters setting:

- **Epoch**
To configure our Epoch, there are a few considerations that have to be made. One of the considerations is the amount of data we have. Since we do not have a big size of dataset, the Epoch size cannot be too big, as the machine will try to memorize the dataset we give it, and we also have to pay attention to the learning rate. Using a high epoch with a high learning rate makes the model learn in a very rough way (either getting the training purpose or miss it at all), and causes it to overfit very fast. Using higher epoch but lower the Learning Rate also was not beneficial as it took way too long to compute, and only managed to create a miniscule amount of changes into the model, which will cause the model to underfit, which means the model did not learn our instructions at all. At most we are using around 4-6 epochs.
- **Batch Size**
We used the batch size of 4 in all training. Higher can allow for faster training, but more likely to crash by using up all the VRAM available.
- **Learning Rate**
For Learning Rate, we decided to use around 1×10^{-4} , 2×10^{-4} , and 1×10^{-5} . The learning rate of powers 10^{-4} works the best with 4 epochs, as it allows the model to learn as much as possible with low overfitting, and 10^{-5} was too small of change to make our training significant.
- **Optimizer**
We decided to use AdamW Optimizer. Using this Optimizer, allows us to control the weight decay to prevent overfitting on our dataset. *Update matrices* are added to the existing weights, and the model is trained only by these new weights.
- **QLoRA (Quantized Low Rank Adaptation)**
To train LLM requires a lot of resources. QLoRA is a training method that accelerates the training of large models while consuming **less memory**. It is an innovative method that enhances the efficiency of memory when training models with large parameters. QLoRA combines 4-bit quantization and LoRAs to create a computationally efficient training process

These are the QLoRA HyperParameter that we used:

- lora_r : We set this value to 64, this value denominates the rank of lora parameters, or the dimension of matrix
 - lora_alpha : We set this value to 16, this parameter is used to control the initialization scale of loralib.linear library. The higher the alpha the more compatible or fit the result to the trained dataset.
 - lora_dropout : we set this value to 0.1, lora_dropout is a regularization parameter for the training using weight decay. This is used to prevent overfitting in the training.
- **Bits and Bytes**
Based on the paper **QLoRA: Efficient Finetuning of Quantized LLMs**, Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, Luke Zettlemoyer. 4-bit Normal Float(NF4) is an optimal quantization scheme for distributed LLM weights, it provides better accuracy than standard 4-bit floats. Furthermore, double quantization reduces memory by quantizing the quantization constants and saves around 0.4 bits per parameter.

3.4. Model Sharding

Another issue we found with LLM training and inference on a small resource is of the model size itself. Generally the models are basically tensors files that reach up to 13GB of size, which drastically use up most resources, whether it is RAM when it is firstly loaded into the code itself, or VRAM when we eventually move it into the GPU for inference and training.

Our way out for this, is to do sharding on these models. By sharding a model's tensor to a smaller more manageable size, we are able to save up memory usage while loading it in the RAM and in the VRAM, which allows our model to be loaded in computers that have less memory. We also found out that by sharding, it did not affect the model's performance, and allows it to be more memory efficient compared to versions that we did not shard.

3.5. Inference / Generate

Inferencing is the process where we use our fine-tuned models to generate and predict the summary result from our gathered product reviews. As our base model comes from Hugging Face, we decided to use the *Hugging Face Pipe* to inference the product reviews. This process can be divided into four parts:

3.5.1. Preparing The Dataset

The product reviews data that we used for inference came from the "test" component of the amazon dataset that we made (*refer ro 3.3*). There were a total of 133 product reviews. Each of this data needs to be split into two, the first part as the prompt and the second part as the reference. After pre-processing, the dataset will be put into a list in order to simplify individual inference as well as batch inference.

3.5.2. Creating the Prompt

There are different approaches in which we can create our prompt. The first approach is to use the template given by each model. One common feature of a prompt that all models have is the use of a system prompt. A system prompt will steer the language model's response based on the guidance and limitation that is given. Here is an example of a prompt generated with Mistral LLM base model's template:

```
"<|system|>
{system_message}
</s>
<|user|>
{summary}</s>
<|assistant|>"
```

The curly braces represent the content that should be manually filled by the user of the prompt.

The second approach for generating our prompt would be to directly give out the instruction to the model. This approach would work if the fine-tuned models are trained without the use of the system prompt itself. This would be the case for our LLM models. Both methods of prompting will yield different responses, thus leading us to try different prompting methods for different trained models.

The next strategy that needs to be considered is the guidance/prompt that the model will receive. Since our intention is to summarize a product review then give out the overall sentiment and key aspect of the product, we were able to come out with this prompt:

"Summarize the following Amazon reviews into 70 to 100 words while maintaining the user experience, and write it in first person perspective Also, find the overall sentiment of the following text, and show me the important aspects of the text."

The overall text length is around 200 to 400 words. We decided to use 30% to 50% of the text due to the reason explained in 3.1. Which is why we included 70-100 words in the prompt.

The prompt used here follows the prompt that was used in the training and validation method of our model. Naturally, the inference results were able to follow the format that had been laid out in the training process. But there are still different prompts that we tried out, such as putting an example of the result format that we want.

Although the inference result from this prompt yielded the correct result, the formatting was inconsistent at times. At the end, we decided to use our first prompt.

3.5.3. Setting up the inference parameters

There are a couple of parameters that can be used. We first need to define the task of our models. For the non-LLM models that we used such as PEGASUS, we define the task to be *summarization*. Although our LLM also does summarization, we used *task-generation* instead since these models cover a broader spectrum such as sentiment analysis and finding the key-aspect of the reviews.

Other settings that we use are ‘return_full_text’ and ‘max_length’. ‘return_full_text’, set to false, is used to let the model not repeat the starting prompt.

Huggingface’s pipeline also allowed us to set the max length of the response of the model. This is set with the ‘max_new_tokens’ parameter. If we were to follow the length based on our prompt, we would limit the length to 70 -100 words , plus the length of the overall sentiment and key aspect. However, we could also set the length to a large amount and post-process the result afterwards.

There are several problems with this parameter. If a summary was too long, the model would not just loop and give the prompt back, and setting it too high of value will make the model give out wrong answers that are not correlated with the text at all. More on this can be seen in Post-Processing.

Another important parameter that can alter the result of the response would be the ‘temperature’ of the model. The temperature we set will affect the probability distribution of the token, which means that lower temperature can restrict the response of the model strictly to the prompt given. We decided to stick with the default temperature of the model.

3.5.4. Post-processing

We see our model’s result is acceptable, but there were some minor issues like looping and there were unwanted words that were not trained appearing in the result. Our post-processing gives a thorough adjustment to the result of our model. The 1st usage is to prevent looping that is done by the model as in repeating the same result over and over again. The 2nd feature of our post-processing is to remove any unwanted words that are displayed on the result.

3.6. Evaluation Method and Performance

3.6.1. ROUGE

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score is a metric that measures similarity between the generated text and reference based on its overlapping n-grams and word sequence by calculating its respective recall. In System Implementation, we utilize ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-LSum.

ROUGE-N, which covers ROUGE-1 and ROUGE-2, measures the overlapping n-grams between the targeted and reference text. N denotes the n-gram overlap which means ROUGE-1 measures single words and ROUGE-2 measures two-word sequences. This score is used to evaluate the correctness of grammar and its fluency. Then, ROUGE-L measures the longest common subsequence of both of the texts. The objective of the score is sentence level structure. The difference between ROUGE-L and ROUGE-LSum is that ROUGE-L counts an average over individual sentences while ROUGE-LSum counts the entire summary.

3.6.2. SacreBLEU

SacreBLEU is an evaluation system that is comparable and can produce BLEU scores. The main difference is that SacreBLEU is easier to implement.

BLEU score measures the similarity between the generated and reference text based on the precision of n-grams. It also produces a term called BP (Brevity Penalty) which is used to penalize short prediction outputs.

3.6.3. BERTScore

BERTScore is a metric that evaluates the similarity or semantic equivalence between the generated text and reference by using contextual embeddings.

Hashcode used:

microsoft/deberta-large-mnli_L18_no-idf_version=0.3.12(hug_trans=4.34.0)

4. Result

In System Implementation I, we have used 3 models which are PEGASUS, T5, and BERT. When we run tests about these models, we have found that the best fine-tuned result is led by T5 quantitatively, however to the human eye we found out that Pegasus are more capable of generating results more favorable than T5. Although, these models were not able to summarize the descriptions into the expected description properly. The reason could be due to the dataset that we used gave out summaries that were abstractive in nature.

In System Implementation II, since we have added sentiment analysis and key point interest extraction. We have tested multiple LLMs. The following tables are results we achieved from **Llama2**, **Mistral**, and **Zephyr**. And for the smaller model, we still try PEGASUS for summarization. But we will be only focusing on summarization for PEGASUS, since that is the intended use case for that model.

4.1 First Model : Mistral 7B

Table 4.1.1. Fine Tuned Mistral 7B Post Processed Result 1 Table

Model	Mistral 7B Fine Tuned + Post Processed
Tokenizer	LlamaTokenizerFast
Data	https://huggingface.co/datasets/bobbybelajar/AmazonMixedLength/viewer/default/test
Source	Summarize the following Amazon reviews into 70 to 100 words while maintaining the user experience, and write it in first person perspective Also, find the overall sentiment of the following text, and show me the important aspects of the text. Length:: 1:25 MinsThese are unbelievably strong. I love blueberries. The taste is stunning and overpowering. My wife's eyes got really big on the first try and she exclaimed "wow, you can really taste the blueberries" (I didn't have to tell her it was supposed to be blueberry). After the first one, I am in deep fear of eating the whole bag in seconds, but due to the molasses and "natural flavor", they seem so overpoweringly sweet, so after 4 of them (about 35 calories total), I can't take another. This is good because it's a sweet that prevents me from wanting more.This is for those who like things intense. you like things intense.Ingredients:=====Molasses, wheat flour, blueberry puree, citric acid, natural flavor, licorice extractNo fat, 8.7 calories per piece, 10 grams per piece.Serving size of 15 pieces (4 times more than I want to eat in 1 sitting):130 calories: 21 g sugars out of 30 g carbs with 1 g protein.Video Comments=====The video is a taste test with my 3 year old son. The first is Darrell Lea's strawberry (his least favorite), followed by Twizzler's strawberry, followed by Darrell Lea mango (not shown), then Panda's blueberry. My goal has been to find something that does not have Twizzler's Red-40. His little sister did not like them at all.

Expected	<p>Results: I tried Panda's blueberry licorice, and they are unbelievably strong, with a stunning and overpowering blueberry taste. They are so sweet due to the molasses and "natural flavor" that I can't eat too many in one sitting, which is a good thing. They have no fat and around 8.7 calories per piece, making them a sweet treat without Twizzler's Red-40. My wife loved them too, and they are perfect for those who enjoy intense flavors.</p> <p>Overall Sentiment: positive</p> <p>Aspects: Positive: blueberry taste, overpowering flavor, low in fat, low calories</p> <p>Negative: x</p>
Fine Tuned Result	<p>Results: These are incredibly strong. The taste is overpowering, and my wife exclaimed that she could taste the blueberries. After the first one, I was afraid of eating the entire bag in seconds, but due to the molasses and "natural flavor," they seemed too sweet. I couldn't take another after four of them. This is a good thing because it prevents me from wanting more. This is for those who like things intense.</p> <p>Overall Sentiment: positive</p> <p>Aspects:</p> <p>Positive: strong, taste, sweet</p> <p>Negative: x</p>

Table 4.1.2. Fine Tuned Mistral 7B Post Processed Result 2 Table

Model	Mistral 7B
Tokenizer	LlamaTokenizerFast
Data	https://huggingface.co/datasets/bobbybelajar/AmazonMixedLength/viewer/default/test
Source	<p>Summarize the following Amazon reviews into 70 to 100 words while maintaining the user experience, and write it in first person perspective. Also, find the overall sentiment of the following text, and show me the important aspects of the text. I first found this Fancy Feast with Cheddar at my local grocers about six months ago. My cats loved it and I have been waiting on Amazon to bring it on board for a long time. Finally they have. "Subscribe and Save" here I come! My cats especially enjoy this Fancy Feast Cheddar with Whitefish as a base. My store had Whitefish, Chicken Purina Fancy Feast Delights with Cheddar Chicken, 3-Ounce (Pack of 24), and Turkey Purina Fancy Feast Delights with Cheddar Turkey, 3-Ounce (Pack of 24) available in their limited edition Display Case when the product first arrived. Whitefish and Chicken were equally popular with my cats, but they seem tired of Turkey and were not crazy about it. I let them try the Tuna Purina Fancy Feast Delights with Cheddar Tuna, 3-Ounce (Pack of 24) when it arrived later. They also liked that. But Whitefish and Chicken topped the list of the things they liked in this line. They do</p>

	<p>not care for the Fancy Feast Souffles at all. Fancy Feast Elegant Medleys for Cats, White Meat Chicken & Cheddar Cheese Souffle With Garden Greens, 3-Ounce Cans (Pack of 24) I do not know if they use a different variety of cheese or what, but they do not care for those. So I have been giving the kittys The Cheddar Whitefish now for six months. The cats always gobble it up and show no signs of tiring of it. So, I recommend that you give your kitty's a chance to try it. They too should become Cheddar lovers. PS: I did put Cheddar Whitefish on my Subscribe and Save list and have it delivered every few months. I hope you are as pleased with it as I am. Thank You.</p>
Expected	<p>Results: I discovered Fancy Feast with Cheddar six months ago at my local grocers, and my cats loved it. I've been waiting for Amazon to stock it, and now they finally have! My cats enjoy the Cheddar with Whitefish flavor the most, and they also liked the Cheddar with Chicken and Cheddar with Tuna varieties. However, they didn't care for the Fancy Feast Souffles with Cheddar. I've been giving them the Cheddar Whitefish for six months, and they still love it. I highly recommend giving it a try for your kitties too!</p> <p>Overall Sentiment: positive Aspects: Positive: enjoyment, eagerness, whitefish and chicken, recommendation, satisfaction, appreciation Negative: x</p>
Fine Tuned Result	<p>Results: I first found this Fancy Feast with Cheddar at my local grocery store about six months ago, and my cats absolutely loved it. I've been waiting for Amazon to offer it, and now they have it with the "Subscribe and Save" option. My cats especially enjoy the Fancy Feast Cheddar with Whitefish as a base. They also like the Fancy Feast Delights with Cheddar Chicken and Tuna, but Whitefish and Chicken are their top choices. They don't care for the Fancy Feast Souffles or Elegant Medleys. So, I've been feeding them the Cheddar Whitefish for six months, and they still devour it without getting tired of it. I highly recommend trying it out for your cats.</p> <p>Overall Sentiment: positive Aspects: Positive: loved it Negative: x</p>

Human Analysis:

Our Fine-Tuned Mistral 7B model is able to inference our product review in a short and concise manner. These are some of the human analyses that we have found.

1.Extracted Information:

From all the data that we performed a human analysis on, the main points of the text were successfully taken. This includes two main parts:

- Opinion of the product:
 - “These are incredibly strong”
 - “They seemed too sweet”

- Concluding statement:
 - “I highly recommend trying it out for your cats.”
 - “This is for those who like things intense.”

The key words of this information were taken immediately from the text.

2.Non-Extracted Information:

The trained Mistral 7B model was also able to remove unnecessary information such as:

- Ingredients of a product :“Molasses, wheat flour, blueberry puree, citric acid, natural flavor, licorice extract”,
- Not-important personal opinion is [*“4 times more than I want to eat in I sitting’]* .

4.2 Second Model : PEGASUS

Table 4.2.1. Fine Tuned PEGASUS Result 1

Model	PEGASUS-XSUM
Tokenizer	PegasusTokenizerFast
Data	samsung298.csv (Row 0)
Source	<p>CAPTURE THE NIGHT IN LOW LIGHT: Whether youre headed to a concert or romantic night out, theres no such thing as bad lighting with Night Mode; Galaxy S23+ lets you capture epic content in any setting with stunning Nightography</p> <p>HIGH CAMERA RESOLUTION: Create crystal-clear content worth sharing with the 50MP high-resolution camera of Galaxy S23+; Whether youre posting or printing, Galaxy S23+ always does the moment justice</p> <p>FASTEST MOBILE PROCESSOR AVAILABLE: Whether youre working hard, playing hard or doing both at the same time, smoothly switch between apps with our fastest processor ever</p> <p>UNLOCK SMOOTH GAMING: Game at full throttle and smoothly switch between apps; Get more done with our fastest processor ever combined with massive internal storage; And if youre on the move, keep going with a screen that adapts to your environment</p> <p>POWER TO KEEP MOVING YOU FORWARD: Confidently binge-watch, game or surf without worrying about your phone dying; The impressive 4,700mAh battery* will keep you going longer**</p> <p>MORE STORAGE. MORE MEMORIES: Youre good on storage space; With an impressive amount of storage, starting at 256GB, you can enjoy being behind the camera without the stress that comes with getting close to filling up your phone</p> <p>ADAPTIVE DISPLAY: Watch your content comfortably as Adaptive Vision Booster helps your screen adapt to your environment; Whether out in direct sunlight or in bed with the lights off, the contrast on Galaxy S23+s expansive screen still looks amazing</p>
Expected	Night Mode, a 50MP high-resolution camera, Fastest Mobile Processor Available, 4,700mAh, 256GB Storage
Fine Tuned Result	The Galaxy S23+ features Night Mode , a 50MP high-resolution camera , and fast switching between apps with the fastest processor ever. It also comes with a powerful 4,700mAh battery that can last up to 10 days on a single charge. Whether you're working hard, playing hard, or doing both, the Galaxy S23+ always delivers the best experience.

Table 4.2.2. Fine Tuned PEGASUS Result 2

Model	PEGASUS-XSUM
Tokenizer	PegasusTokenizerFast
Data	https://huggingface.co/datasets/bobbybelajar/PegasusAmazon/viewer/default/test (Row 70)
Source	<p>Seriously, I was very happily surprised with this excellent brown bread. I found it to be nice and moist, with the texture and even the taste reminding me a bit of fresh gingerbread or a good rich muffin. The problem I had, however, was not insignificant. I exchanged emails with a very nice and accommodating representative of B&M, and apparently I got a can with inside walls that somehow escaped being thinly coated with vegetable oil. Hence, especially since the inside of the can has horizontal ridges, vertical to the direction that the bread needs to be slid, the can had a fiendish death grip on the bread. The instructions say you may need to ease the bread out by using a spoon. If only. This only creates a cave in the bread, maybe useful if you are into bread sculpting. Even using the cutaway lid, which should be the perfect size, you just keep compressing the bread on that side, of course with every quarter-inch or so of compression you're losing slices you paid for. The simplest way for me was sliding a thin steak knife around the edge, but since, again, the can has those internal ridges, you end up mangling/losing part of the product. Overall, ordering this bread as long-term backup is perfect for me, so I will probably order again. But it is a little depressing that we just launched another rover to Mars, to join another that is still operational, yet we can't manage to get a simple vegetable oil coating consistently inside of a can of bread. Out of twelve cans purchased by yours truly, so far I have opened four, each time soon wanting to call emergency services to bring in the Jaws of Life, but then figuring they might be needed more out of the highway. That aside, Thumbs Up to B&M for a delicious long-term-storage product.</p>
Expected	<p>I was pleasantly surprised by the excellent taste and moistness of this brown bread. It reminded me of fresh gingerbread or a rich muffin. However, I encountered a significant problem with the can's lack of vegetable oil coating on the inside walls, making it difficult to remove the bread without compressing or mangling it. Despite this issue, I still plan to order this bread as long-term backup. It's frustrating that we can launch rovers to Mars, but struggle with a simple vegetable oil coating in a bread can. Overall, thumbs up to B&M for a delicious long-term storage product.</p>
Fine Tuned Result	<p>I was pleasantly surprised by this excellent brown bread. The texture and taste were reminiscent of fresh gingerbread or a good rich muffin. However, I encountered a problem with the can's interior walls, which prevented the bread from being thinly coated with vegetable oil. The</p>

	<p>instructions suggest easing the bread by using a spoon, but using a knife is not recommended due to the ridges inside the can. Even using the cutaway lid, which should be the perfect size, you just keep compressing the bread on that side, of course with every quarter-inch or so of compression you're losing slices you paid for. The simplest way for me was sliding a thin steak knife around the edge, but since, again, the can has those internal ridges, you end up mangling/losing part of the product.</p>
--	--

Human Analysis:

For PEGASUS, we tested on ± 300 Review and 300 Product Description. It can successfully extract the main keypoint of the review from some result we read. The overall summarization is good although sometimes it still hallucinates although most of the time its a minor hallucination. In addition, this model can do summarization but not extract sentiment keywords per our own specification therefore, we will not use this model.

4.3 Third Model : Zephyr 7B

Table 4.3.1. Fine Tuned Zephyr 7B Post Processed Result 1 Table

Model	Zephyr 7B
Tokenizer	LlamaTokenizerFast
Data	https://huggingface.co/datasets/bobbybelajar/AmazonMixedLength/viewer/default/test (Row 9)
Source	<p>Summarize the following Amazon reviews into 70 to 100 words while maintaining the user experience, and write it in first person perspective Also, find the overall sentiment of the following text, and show me the important aspects of the text. These are unbelievably strong. I love blueberries. The taste is stunning and overpowering. My wife's eyes got really big on the first try and she exclaimed "wow, you can really taste the blueberries" (I didn't have to tell her it was supposed to be blueberry). After the first one, I am in deep fear of eating the whole bag in seconds, but due to the molasses and "natural flavor", they seem so overpoweringly sweet, so after 4 of them (about 35 calories total), I can't take another. This is good because it's a sweet that prevents me from wanting more. This is for those who like things intense. you like things intense. Ingredients:=====Molasses, wheat flour, blueberry puree, citric acid, natural flavor, licorice extractNo fat, 8.7 calories per piece, 10 grams per piece. Serving size of 15 pieces (4 times more than I want to eat in 1 sitting):130 calories: 21 g sugars out of 30 g carbs with 1 g protein. Video Comments=====The video is a taste test with my 3 year old son. The first is Darrell Lea's strawberry (his least favorite), followed by Twizzler's strawberry, followed by Darrell Lea mango (not shown), then Panda's blueberry. My goal has been to find something that does not have Twizzler's Red-40. His little sister did not like them at all.</p>
Expected	<p>Results: I tried Panda's blueberry licorice, and they are unbelievably strong, with a stunning and overpowering blueberry taste. They are so sweet due to the molasses and "natural flavor" that I can't eat too many in one sitting, which is a good thing. They have no fat and around 8.7 calories per piece, making them a sweet treat without Twizzler's Red-40. My wife loved them too, and they are perfect for those who enjoy intense flavors.</p> <p>Overall Sentiment: positive</p> <p>Aspects:</p> <p>Positive: blueberry taste, overpowering flavor, low in fat, low calories</p> <p>Negative: x</p>
Fine Tuned Result	<p>Results: These blueberry treats are incredibly strong in flavor. The taste is stunning and overpowering. My wife was amazed at the intensity of the blueberry flavor, exclaiming "Wow, you can really taste the blueberries." After trying one, I'm worried about consuming</p>

	<p>the entire bag quickly, but the molasses and natural flavor make them overly sweet. Despite the intense taste, they satisfy my sweet cravings and prevent me from wanting more. These treats are for those who enjoy intense flavors.</p> <p>Overall Sentiment: positive</p> <p>Aspects:</p> <p>Positive: stunning, overpowering, sweet, prevent more</p> <p>Negative: x</p>
--	---

Table 4.3.2. Fine Tuned Zephyr 7B Post Processed Result 2 Table

Model	Zephyr 7B
Tokenizer	LlamaTokenizerFast
Data	https://huggingface.co/datasets/bobbybelajar/PegasusAmazon/viewer/default/test (Row 133)
Source	<p>Summarize the following Amazon reviews into 70 to 100 words while maintaining the user experience, and write it in first person perspective Also, find the overall sentiment of the following text, and show me the important aspects of the text. I first found this Fancy Feast with Cheddar at my local grocers about six months ago. My cats loved it and I have been waiting on Amazon to bring it on board for a long time. Finally they have. "Subscribe and Save" here I come! My cats especially enjoy this Fancy Feast Cheddar with Whitefish as a base. My store had Whitefish, Chicken Purina Fancy Feast Delights with Cheddar Chicken, 3-Ounce (Pack of 24), and Turkey Purina Fancy Feast Delights with Cheddar Turkey, 3-Ounce (Pack of 24) available in their limited edition Display Case when the product first arrived. Whitefish and Chicken were equally popular with my cats, but they seem tired of Turkey and were not crazy about it. I let them try the Tuna Purina Fancy Feast Delights with Cheddar Tuna, 3-Ounce (Pack of 24) when it arrived later. They also liked that. But Whitefish and Chicken topped the list of the things they liked in this line. They do not care for the Fancy Feast Souffles at all. Fancy Feast Elegant Medleys for Cats, White Meat Chicken & Cheddar Cheese Souffle With Garden Greens, 3-Ounce Cans (Pack of 24) I do not know if they use a different variety of cheese or what, but they do not care for those. So I have been giving the kitties The Cheddar Whitefish now for six months. The cats always gobble it up and show no signs of tiring of it. So, I recommend that you give your kitty's a chance to try it. They too should become Cheddar lovers. PS: I did put Cheddar Whitefish on my Subscribe and Save list and have it delivered every few months. I hope you are as pleased with it as I am. Thank You.</p>
Expected	Results: I discovered Fancy Feast with Cheddar six months ago at my local grocers, and my cats loved it. I've been waiting for Amazon to

	<p>stock it, and now they finally have! My cats enjoy the Cheddar with Whitefish flavor the most, and they also liked the Cheddar with Chicken and Cheddar with Tuna varieties. However, they didn't care for the Fancy Feast Souffles with Cheddar. I've been giving them the Cheddar Whitefish for six months, and they still love it. I highly recommend giving it a try for your kitties too!</p> <p>Overall Sentiment: positive</p> <p>Aspects:</p> <p>Positive: enjoyment, eagerness, whitefish and chicken, recommendation, satisfaction, appreciation</p> <p>Negative: x</p>
Fine Tuned Result	<p>Results: My cats, especially the younger ones, absolutely love Fancy Feast with Cheddar. It's a great base for them to enjoy. When the store had limited-edition flavors like Whitefish, Chicken, and Turkey, my cats tried them all and were equally enthusiastic. However, they grew tired of the Turkey flavor. Tuna was also well-received. Whitefish and Chicken topped the list of favorites. My cats don't like the Fancy Feast Souffles, but they adore the Cheddar Whitefish. It's a must-try for your kitties. I've been subscribing to it for the past few months, and my cats never get tired of it. I highly recommend giving your cats a chance to try it. It's a great way to introduce them to the wonderful world of Cheddar. Overall Sentiment: positive</p> <p>Aspects:</p> <p>Positive: Fancy Feast with Cheddar, Whitefish, Chicken</p> <p>Negative: Turkey ,Fancy Feast Souffles ,Tuna ,Elegant Medleys ,tired of</p>

Human Analysis :

1.Extracted Information:

From all the data that we performed a human analysis on, the main point of the text were successfully taken. Which includes these two main parts:

- Opinion of the product:
 - "Wow, you can really taste the blueberries."
 - "They seemed too sweet"
- Concluding statement:
 - "I highly recommend giving your cats a chance to try it."
 - "This is for those who like things intense."

We concluded that the extractive capabilities of our model were able to perform as well as Mistral 7B.

2. Non-Extracted Information and Flaw:

One important objective of a summarizer is to be able to remove as much unnecessary information as possible. From our human analysis, we can see that the trained Mistral 7B model was able to perform well in removing unnecessary information. However, what we notice from Zephyr's 7B inference result is that there were able to do this less efficiently on this aspect:

- Two sentence were needed to explain two simple point:
 - Mistral's Inferenced Result: "These are incredibly strong. The taste is overpowering, and my wife exclaimed that she could taste the blueberries."
 - Zephyr's Inferenced Result: "These blueberry treats are incredibly strong in flavor. The taste is stunning and overpowering. My wife was amazed at the intensity of the blueberry flavor, exclaiming "Wow, you can really taste the blueberries.""

Unfortunately, Zephyr needs to use two sentences to give out two points. Mistral being able to do it in one sentence will greatly reduce the summary length, which is crucial for a summarizer.

4.4 Fourth Model : Llama 2 7B

Table 4.4.2. Fine Tuned Llama 2 7B Post Processed Result 1 Table

Model	Llama 2 7B
Tokenizer	LlamaTokenizerFast
Data	https://huggingface.co/datasets/bobbybelajar/PegasusAmazon/viewer/default/test (Row 133)
Source	<p>Summarize the following Amazon reviews into 70 to 100 words while maintaining the user experience, and write it in first person perspective Also, find the overall sentiment of the following text, and show me the important aspects of the text. Length:: 1:25 Mins These are unbelievably strong. I love blueberries. The taste is stunning and overpowering. My wife's eyes got really big on the first try and she exclaimed "wow, you can really taste the blueberries" (I didn't have to tell her it was supposed to be blueberry). After the first one, I am in deep fear of eating the whole bag in seconds, but due to the molasses and "natural flavor", they seem so overpoweringly sweet, so after 4 of them (about 35 calories total), I can't take another. This is good because it's a sweet that prevents me from wanting more.This is for those who like things intense. you like things intense.Ingredients:=====Molasses, wheat flour, blueberry puree, citric acid, natural flavor, licorice extractNo fat, 8.7 calories per piece, 10 grams per piece.Serving size of 15 pieces (4 times more than I want to eat in 1 sitting):130 calories: 21 g sugars out of 30 g carbs with 1 g protein.Video Comments=====The video is a taste test with my 3 year old son. The first is Darrell Lea's strawberry (his least favorite), followed by Twizzler's strawberry, followed by Darrell Lea mango (not shown), then Panda's blueberry. My goal has been to find something that does not have Twizzler's Red-40. His little sister did not like them at all.</p>
Expected	<p>Results: I tried Panda's blueberry licorice, and they are unbelievably strong, with a stunning and overpowering blueberry taste. They are so sweet due to the molasses and "natural flavor" that I can't eat too many in one sitting, which is a good thing. They have no fat and around 8.7 calories per piece, making them a sweet treat without Twizzler's Red-40. My wife loved them too, and they are perfect for those who enjoy intense flavors.</p> <p>Overall Sentiment: positive</p> <p>Aspects:</p> <p>Positive: blueberry taste, overpowering flavor, low in fat, low calories</p> <p>Negative: x</p>
Fine Tuned Result	<p>Results: These are incredibly strong. The taste is overpowering, and my wife exclaimed that she could taste the blueberries. After the first one, I was afraid of eating the entire bag in seconds, but due to the molasses and "natural flavor," they seemed too sweet. I couldn't take</p>

	<p>another after four of them. This is a good thing because it prevents me from wanting more. This is for those who like things intense.</p> <p>Overall Sentiment: positive</p> <p>Aspects:</p> <p>Positive: strong, taste, sweet</p> <p>Negative: x</p>
--	--

Table 4.4.2. Fine Tuned Llama 2 7B Post Processed Result 2 Table

Model	Llama 2 7B
Tokenizer	Llama 2 Tokenizer
Data	https://huggingface.co/datasets/bobbybelajar/PegasusAmazon/viewer/default/test (Row 133)
Source	<p>Summarize the following Amazon reviews into 70 to 100 words while maintaining the user experience, and write it in first person perspective Also, find the overall sentiment of the following text, and show me the important aspects of the text. I first found this Fancy Feast with Cheddar at my local grocers about six months ago. My cats loved it and I have been waiting on Amazon to bring it on board for a long time. Finally they have. "Subscribe and Save" here I come! My cats especially enjoy this Fancy Feast Cheddar with Whitefish as a base. My store had Whitefish, Chicken Purina Fancy Feast Delights with Cheddar Chicken, 3-Ounce (Pack of 24), and Turkey Purina Fancy Feast Delights with Cheddar Turkey, 3-Ounce (Pack of 24) available in their limited edition Display Case when the product first arrived. Whitefish and Chicken were equally popular with my cats, but they seem tired of Turkey and were not crazy about it. I let them try the Tuna Purina Fancy Feast Delights with Cheddar Tuna, 3-Ounce (Pack of 24) when it arrived later. They also liked that. But Whitefish and Chicken topped the list of the things they liked in this line. They do not care for the Fancy Feast Souffles at all. Fancy Feast Elegant Medleys for Cats, White Meat Chicken & Cheddar Cheese Souffle With Garden Greens, 3-Ounce Cans (Pack of 24) I do not know if they use a different variety of cheese or what, but they do not care for those. So I have been giving the kitties The Cheddar Whitefish now for six months. The cats always gobble it up and show no signs of tiring of it. So, I recommend that you give your kitty's a chance to try it. They too should become Cheddar lovers. PS: I did put Cheddar Whitefish on my Subscribe and Save list and have it delivered every few months. I hope you are as pleased with it as I am. Thank You.</p>
Expected	<p>Results: I discovered Fancy Feast with Cheddar six months ago at my local grocers, and my cats loved it. I've been waiting for Amazon to stock it, and now they finally have! My cats enjoy the Cheddar with Whitefish flavor the most, and they also liked the Cheddar with</p>

	<p>Chicken and Cheddar with Tuna varieties. However, they didn't care for the Fancy Feast Souffles with Cheddar. I've been giving them the Cheddar Whitefish for six months, and they still love it. I highly recommend giving it a try for your kitties too!</p> <p>Overall Sentiment: positive</p> <p>Aspects:</p> <p>Positive: enjoyment, eagerness, whitefish and chicken, recommendation, satisfaction, appreciation</p> <p>Negative: x</p>
Fine Tuned Result	<p>Results: I first discovered Fancy Feast with Cheddar at my local grocery store and my cats loved it. I've been waiting for Amazon to carry it for a while, and now they have. I've been giving my cats Fancy Feast Cheddar with Whitefish as a base, and they enjoy it. When the product first arrived, my store had Whitefish, Chicken, and Turkey available in limited edition displays. My cats liked Whitefish and Chicken equally, but they weren't as fond of Turkey. When Tuna was added to the line, they also enjoyed it. Whitefish and Chicken were their top choices. They don't care for the Fancy Feast Souffles at all. I've been giving my cats The Cheddar Whitefish for six months, and they always gobble it up. I recommend giving your kitties a chance to try it. They should also become Cheddar lovers. I've added it to my Subscribe and Save list and have it delivered every few months. I hope you enjoy it as much as I do.</p> <p>Overall Sentiment: positive</p> <p>Aspects:</p> <p>Positive: liked, gobble it up</p> <p>Negative: x</p>

Human Analysis:

Similar to our other fine-tuned LLMs, our Llama 2 7B model was also able to provide a short and concise summary.

1. Extracted Information:

From all the data that we performed a human analysis on, the main points of the text were successfully taken. This includes two main parts:

- Opinion of the product:
 - “These are incredibly strong”
 - “They seemed too sweet”
- Concluding statement:
 - “I hope you enjoy it as much as I do.”
 - “This is for those who like things intense.”

The key words of this information were taken immediately from the text.

2. Non-Extracted Information:

Our model was able to remove unnecessary information such as the ingredients of the product [E.g: “Molasses, wheat flour, blueberry puree, citric acid, natural flavor, licorice extract”], and only take the key personal opinion , an example of a not-important personal opinion is [*“4 times more than I want to eat in 1 sitting”*] . The fine-tuned inference results were also of similar points and similar length to our predicted result.

3. Abstractive Feature:

Since our model is abstractive, there were a couple of words that weren't directly extracted from the text. Although words that are not exactly the same as the original text are allowed, we need to make sure that they are semantically the same (having the same meaning).

- Original Text: “They do not care for the Fancy Feast Souffles at all. “
- Inference Result: “They don't care for the Fancy Feast Souffles or Elegant Medleys.”

It does not matter that the way the points are conveyed are different, as the meaning behind the text remains the same.

Aside from that, our model is also able to fix typing errors.

- Original Text: “I first **fount** this Fancy Feast with Cheddar at my local grocers about six months ago.”
- Inference Result: “I first **found** this Fancy Feast with Cheddar at my local grocery store about six months ago”

5. Performance

5.1. ROUGE

Table 5.1 ROUGE Results

Model	Rouge 1	Rouge 2	RougeL	Rouge LSum
Pegasus X-SUM Fine Tuned	0.6302	0.57866	0.61064	0.61091
Pegasus X-SUM Base	0.10138	0.04611	0.08025	0.08012
Mistral 7B 4 Epoch Fine Tuned + PostProcessing (with Aspects)	0.47976	0.22215	0.35548	0.38299
Mistral 7B 4 Epoch Fine Tuned (with Aspects)	0.44493	0.20658	0.32664	0.35273
Mistral 7B Base	0.13208	0.04338	0.09496	0.10177
Llama 2 7B 4 Epoch Fine Tuned + PostProcessing (with Aspects)	0.47951	0.22434	0.35125	0.37957
Llama 2 7B 4 Epoch Fine Tuned (with Aspects)	0.18868	0.08852	0.1397	0.15191
Llama 2 7B Base	0.36932	0.15297	0.23851	0.29502
Zephyr 7B 4 Epoch Fine Tuned + PostProcessing (with Aspects)	0.49446	0.22414	0.360543	0.39206
Zephyr 7B 4 Epoch Fine Tuned (with Aspects)	0.26468	0.12044	0.18891	0.21177
Zephyr 7B Base	0.39543	0.15642	0.25773	0.28414

5.2. SacreBLEU

Table 5.2 SACREBleu Results

Model	Score	Precision	bp	sys_len	ref_len
Pegasus X-SUM Fine Tuned	25.30426	[95.98314, 87.71598, 82.42953, 78.55084]	0.29448	42720	94946
Pegasus X-SUM	0.00012	[63.02105, 30.47165, 20.67556, 17.60399]	4.05E-06	7077	94946

Mistral 7B 4 Epoch Fine Tuned + PostProcessing (with Aspects)	17.42747	[42.33987, 21.34498, 12.46988, 8.18522]	1	19360	13818
Mistral 7B 4 Epoch Fine Tuned (with Aspects)	12.99056	[31.98007, 15.75240, 9.27826, 6.09283]	1	25691	13818
Mistral 7B Base	1.70166	[7.54902, 2.25204, 0.96375, 0.51176]	1	71135	13818
Llama 2 7B 4 Epoch Fine Tuned + PostProcessing (with Aspects)	17.0292	[41.62204, 20.80201, 12.19949, 7.96171]	1	20357	13818
Llama 2 7B 4 Epoch Fine Tuned (with Aspects)	3.82231	[9.56065, 4.63882, 2.71738, 1.77117]	1	91751	13818
Llama 2 7B Base	6.71424	[23.63812, 9.37571, 4.18352, 2.19194]	1	35117	13818
Zephyr 7B 4 Epoch Fine Tuned + PostProcessing (with Aspects)	20.10803	[50.01912, 24.64485, 14.23106, 9.31920]	1	15690	13818
Zephyr 7B 4 Epoch Fine Tuned (with Aspects)	4.14593	[10.56759, 5.07811, 2.92214, 1.88412]	1	78154	13818
Zephyr 7B Base	8.26353	[28.34492, 11.16316, 5.26092, 2.80116]	1	26174	13818

5.3. BERTScore F1

Table 5.1 BERTScore Results

Model	F1	Precision	Recall
Pegasus X-SUM Fine Tuned	0.65200	0.63220	0.67640
Pegasus X-SUM	0.52973	0.62767	0.46113
Mistral 7B 4 Epoch Fine Tuned + PostProcessing (with Aspects)	0.68233	0.67677	0.69199
Mistral 7B 4 Epoch Fine Tuned (with Aspects)	0.67004	0.65992	0.68565
Mistral 7B Base	0.45720	0.43378	0.49080
Llama 2 7B 4 Epoch Fine Tuned + PostProcessing (with Aspects)	0.68156	0.67127	0.69663
Llama 2 7B 4 Epoch Fine Tuned (with Aspects)	0.54594	0.47472	0.64704
Llama 2 7B Base	0.40819	0.39560	0.43146
Zephyr 7B 4 Epoch Fine Tuned + PostProcessing (with Aspects)	0.68379	0.68402	0.68588
Zephyr 7B 4 Epoch Fine Tuned (with Aspects)	0.57384	0.51829	0.65549
Zephyr 7B Base	0.60533	0.58826	0.62780

5.4. Training and Validation Loss

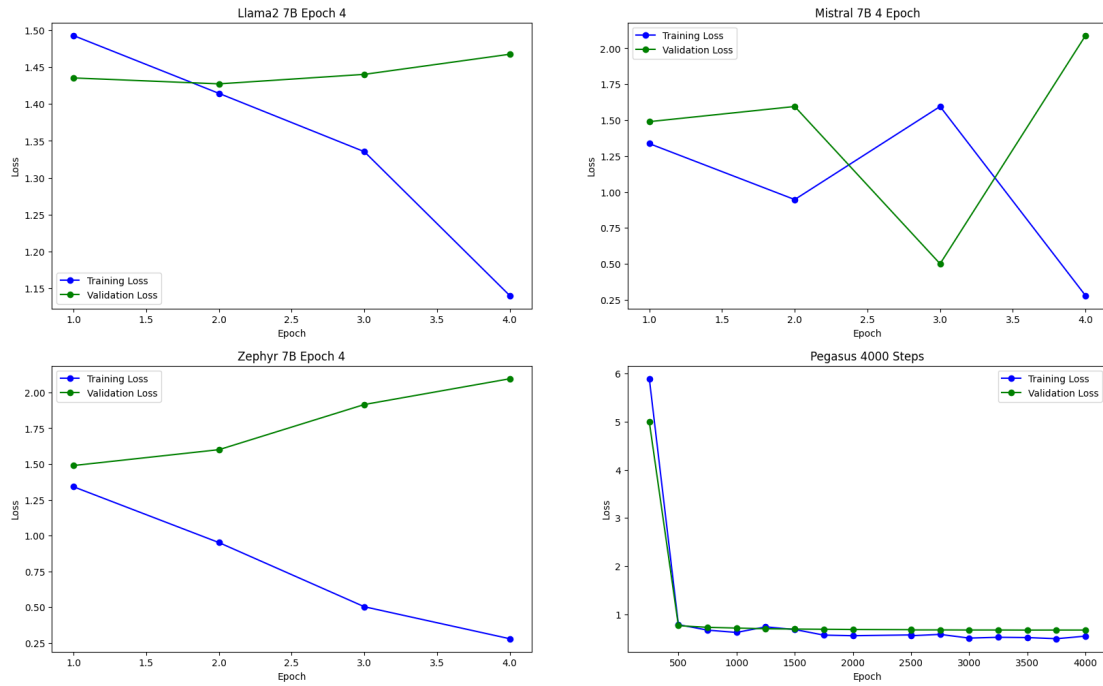


Figure 5.1 Training and Validation Results

With the addition of the sentiment analysis, our dataset size is only around 1313. 918 of them are used for training data, 262 are used for validation, and the remaining 133 are used for our testing set. This dataset is **only used for the LLM model, since they contain the sentiment analysis, and also prompting, so this will not be used for the PEGASUS model**. Similar to System Implementation I, training with a small amount of dataset causes the model to easily overfit. We can see this is the case for the LLM models.

Even though it looks like it overfitted, the results as previously seen were satisfying. We previously tried training with lower learning rate and higher epoch, just to see if we can get a better result by looking at the training losses, but this does not work. In our experiments, the training loss does not define how good the result will be.

On the other hand, For our Pegasus Model, the dataset size is now: 3586, of which splits are: training 2510, validation 716, testing 360. We see that in the Pegasus Graph, the training seems to be more fit as it is more similar to the validation loss compared with the other models.

5.5. Data Analysis

Based on our data, we can conclude that the scores of models after being fine-tuned is lower than the base. These are the factors that differentiate the scores:

5.5.1. Synonyms

Because of the nature of generative models, they would likely be able to generate a synonym or a word that has a close meaning from the reference. While this makes sense in the human eye, this makes the quantitative score lower for ROUGE and BLEU evaluate based on the words itself, not the meaning of the words. And by using BERTScore we can score the result by the semantics, not word by word.

5.5.2. Conclusion: Human Analysis

We see that the models, after fine tuning and post processing, can achieve the desired result, as it can correctly summarize the review, and identify which part the writer likes and hates. Although some hallucination exists, it's fairly minimal.

We need to again clarify that **post processing is crucial**, especially for human eyes. As without post processing, the machine will output unwanted results, such as repeating the prompt, and spitting out nonsense after the machine has finished creating our task.

From all the models, we think that Zephyr generally has the better result, although there are pros and cons to each and all models. For example, Llama 2 model results are more concise, PEGASUS's results are robotic, and the resulting text feels rigid. We like Zephyr the best because it produces texts that are more human-like compared to others.

5.5.3. Conclusion: Quantitative Performance

For all the models that we used, we can see a drastic improvement in ROUGE and SacreBLEU scores for the fine tuned models. Both scores measure the overlap between the generated text and reference text, this shows that our trained model generates a similar result to what we expected.

Out of all models, Pegasus scores the highest. Although this might be unexpected for a smaller model, this might be because Pegasus's inference result does not include the overall sentiment and key aspect of the product reviews.

Setting aside Pegasus, Zephyr 7B generates the highest average score out of all the other large language models. This does not reflect with the validation and training loss of the fine tuning process, which shows that these losses do not immediately correlate to poor results. However, we can also come to this conclusion since Zephyr's base model also has the best score.

For the BERTScore itself, we discovered that all models that experience through fine tuning are seen to have a big improvement in the BERTScore. This makes sense as BERTScore calculates the score

on how semantically close the two texts are. The untrained model generally creates nonsensical responses. Post Processing also helps to increase the BERTScore by a little, because most of the data deleted from the postprocess are redundant results that have similar semantics to the desired result, hence the improvement was not as drastic as previous.

5.5.4. Human Analysis vs Quantitative Performance

Judging from an unbiased point of view, we think that the calculation of the Quantitative Scores has mirrored the judgment of the human eye. Based on our judgment Zephyr has achieved the best result out of all the models. The ROUGE, and SacreBLEU results have also judged that Zephyr has the highest scores.

Although PEGASUS has even higher SacreBLEU scores, we still prefer Zephyr as it offers a more fluid wording in the result. So we cannot just assume how good the model is by looking at the Quantitative score only.

6. Real World Implementation

6.1. Possible Implementation

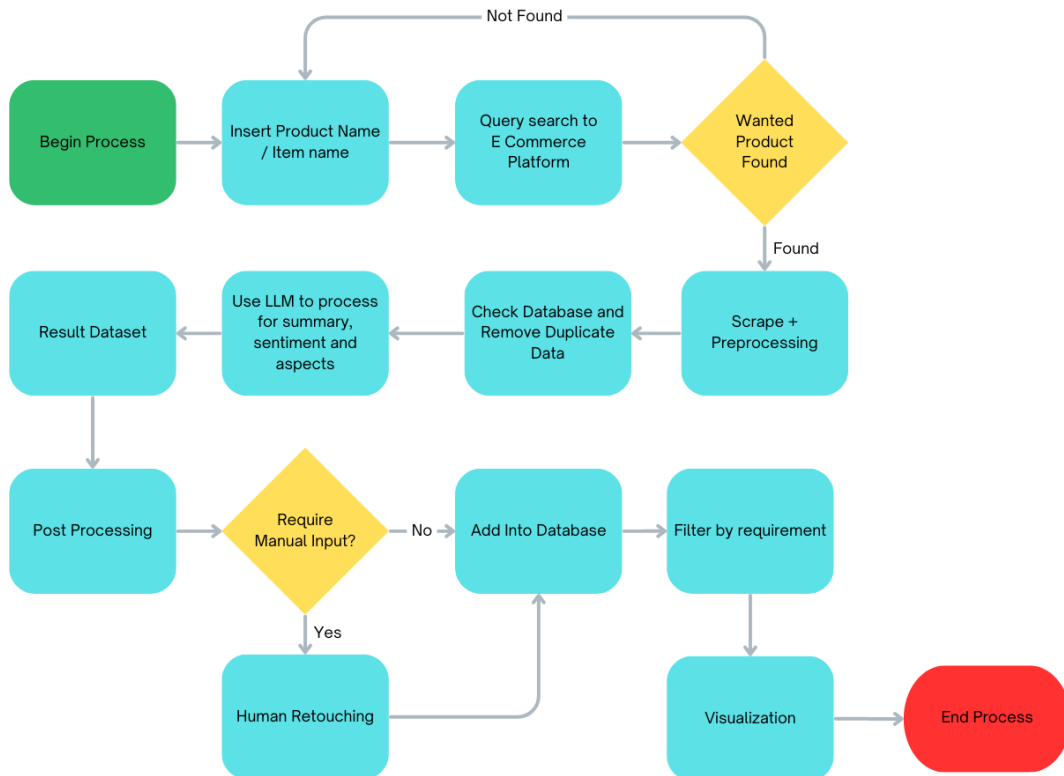


Figure 6.1 System Flowchart

After going through the process of fine-tuning, the machine can now be used using the existing workflow in Figure 6.1.

This is a schematic of a possible implementation of our model. Users would insert their product of desire, and the system would query the internet to search for e-commerce platforms and find the product. After that, the system would scrape and preprocess review data according to our model requirements.

As there are possibilities of existing processed data, data should be checked if there is existing data that we do not need to process again. After LLM has inferred the result, we can do post processing, and add additional human touches if required, and store everything into a database, where later can be used for Visualization purposes.

6.2. Visualization

For Visualization purposes, we inferred a new dataset, that consist of user reviews, which the following are the results :

a. Sample of the Dataframe

	ProductId	ProductTitle	prompt_and_question	length	Brand	Sentiment	Positive	Negative
0	B002ACI89Q	9GreenBox - Venus Fly Trap w/ Gift Box Packing...	Summarize the following Amazon reviews into 70...	255	9Greenbox	Negative	x	very disappointed ,smaller size ,black spots ,...
1	B0019QG5CG	9Greenbox Best Gift Bonsai Juniper Tree, 4 Pound	Summarize the following Amazon reviews into 70...	263	9Greenbox	Positive	packaging, quality	plastic wrap ,younger ,look like ,green bottle...
2	B0019QG5CG	9Greenbox Best Gift Bonsai Juniper Tree, 4 Pound	Summarize the following Amazon reviews into 70...	214	9Greenbox	Negative	x	dissappointment ,smaller ,baby bonsai ,packagin...
3	B0012YEKCM	9GreenBox - 5 Money Tree Plants Braided into 1...	Summarize the following Amazon reviews into 70...	375	9Greenbox	Positive	arrived promptly, well packed	x
4	B0012YEKCM	9GreenBox - 5 Money Tree Plants Braided into 1...	Summarize the following Amazon reviews into 70...	215	9Greenbox	Positive	fast delivery, healthy looking	x

Figure 6.2.1. Sample Dataframe

- b. Assess the desired brand products in our dataframe, in this case starbucks

Starbucks has 6 unique products
 Starbucks Ground Coffee, French Roast, 16-Ounce Bags (Pack of 2)
 Starbucks VIA Ready Brew Colombia Coffee, 50-Count
 Starbucks Flavored Ground Coffee, Vanilla Flavor, Velvety & Rich, Made without Artificial Flavors, Flavor Lock Packaging (11 Ounce (Pack of 6))
 Starbucks Blonde Veranda Whole Bean Coffee, 40 Ounce
 Starbucks frappuccino mocha 9.5 fl oz, Pack of 12
 Starbucks Dark French Roast Coffee, Whole Bean, 12-Ounce Bags (Pack of 3)

Figure 6.2.2. Assessed Products of Starbucks

- c. Analysis of the Brand Sentiment from Dataframe

Sentiment Distribution for Starbucks Brand

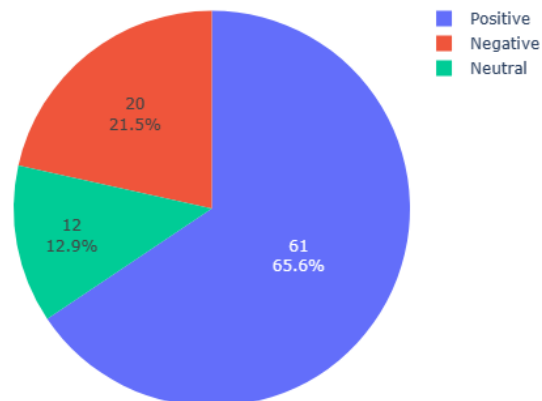


Figure 6.2.3. Sentiment Distribution for the Brand “Starbucks”

- d. Extracting Aspects such as Positive Sentiments with Positive Aspects, Positive Sentiments with Negative Aspects, Negative Sentiments with Positive Aspects, Negative Sentiments with Negative Aspects

Top 10 Term Frequency in posPos Sentiments

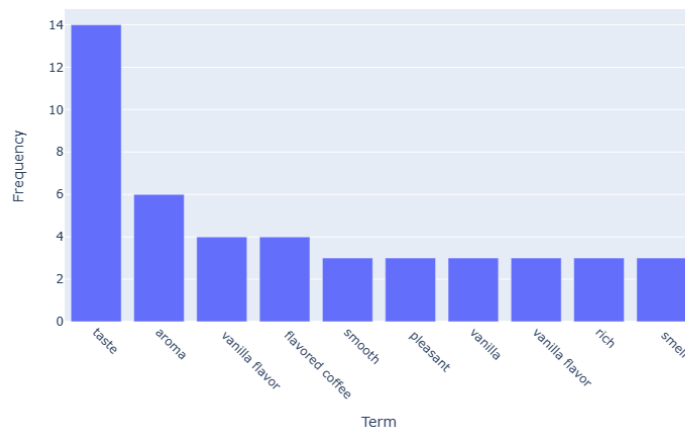


Figure 6.2.4. Positive Aspects of Positive Sentiments Word Frequency

Top 10 Term Frequency in posNeg Sentiments

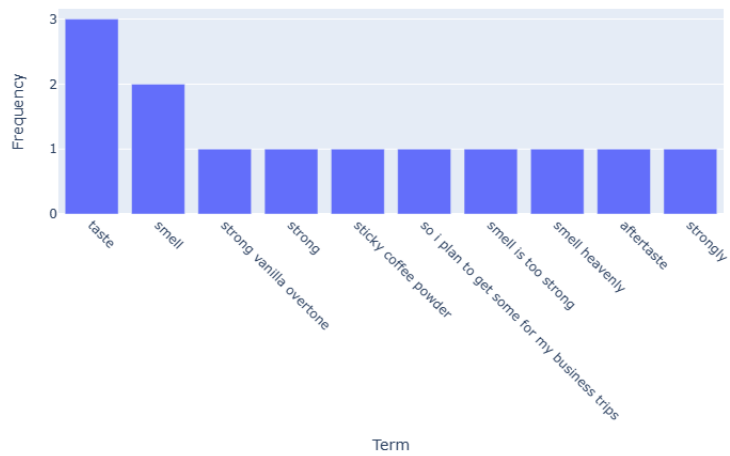


Figure 6.2.5. Negative Aspects of Positive Sentiments Word Frequency

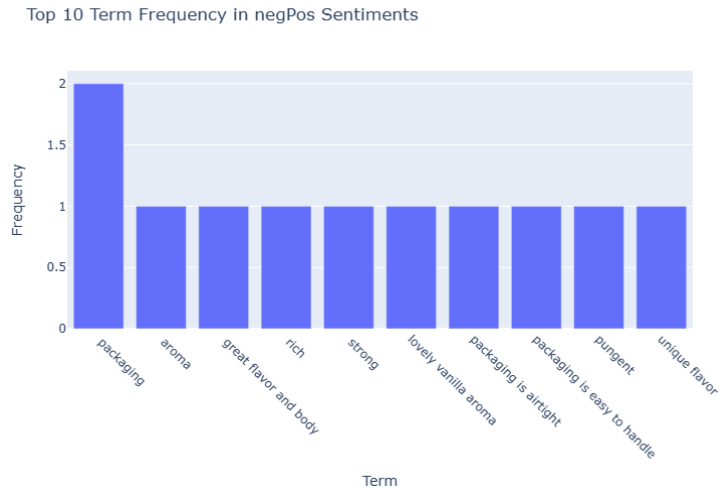


Figure 6.2.6 Positive Aspects of Negative Sentiments Word Frequency

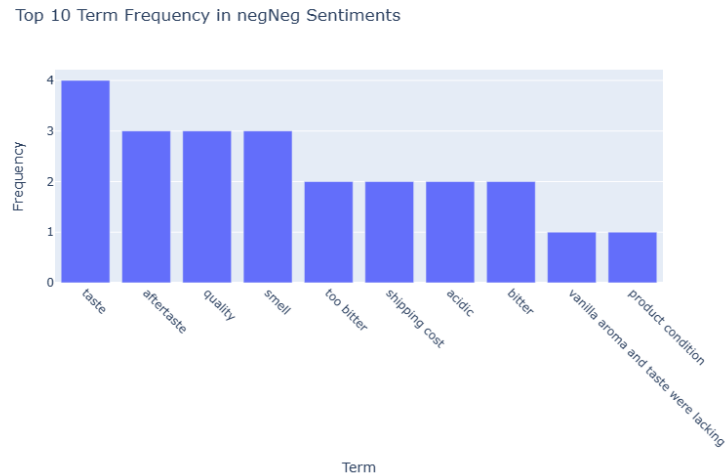


Figure 6.2.7 Negative Aspects of Negative Sentiments Word Frequency

With the data extracted from our machine, users can now easily identify where the strength and weaknesses lie in the brand. Term Frequency is divided into positive-positive, positive-negative and negative-positive, negative-negative, for more detailed feedback of the product. By grouping, customers' needs are more highlighted, allowing businesses to know what their customers love, and what they need to improve on.

7. Conclusion

Throughout System Integration Implementation II, our team had successfully incorporated machine learning and natural language processing into our project. In order to create a fully functioning product review summarizer and sentimental analysis that yields a good result, there are a couple key aspects that we tackled.

There are a lot of models that we tried and used for this project. However, the base models alone could not generate the result that we desire. This pushed us to fine-tune all the base models that we use in order to obtain the desired response. This led us to gather a large amount of robust dataset to obtain the necessary trained models.

After gathering and searching for data for the first six months of our System Integration Implementation, we concluded that it is not enough for us to gather a dataset, but to make one for our own. It is important for the people in machine learning to understand the correlation between training a model and the accuracy of the dataset that was used, this realization served as a breakthrough for us.

Going into the second part of our System Integration, our group decided that it was important for us to try out a larger language model. The newer models that we used such as Llama2, Mistral, and Zephyr helped us overcome the limitation of a smaller model. For instance, our older fine-tuned model, Pegasus, was only able to provide a summary without the sentimental analysis and key aspect of the product review.

Other than fine-tuning, we also utilize other methods to further improve the accuracy of our result which include methodology such as prompt generation and post-processing. All of this combined means that we need to efficiently utilize our GPU resources to train more data at the same time.

To measure the accuracy of our models, we also need to implement a grading criteria for our models. The evaluation metrics that along with our own human analysis helped us decide the best model to use for our final product.

We conclude our project by creating a compact visualization. Our Machine Learning Project can be used as an evaluation tool for the users to evaluate sentiment and key-point identifications. From a large amount of datasets, the user can filter out and gain insights to only a specific category of reviews, saving them a large amount of time.

Bibliography

- Abigail See, Peter J. Liu, Christopher D. Manning, Get To The Point: Summarization with Pointer-Generator Networks (2017)
<https://arxiv.org/pdf/1704.04368v2.pdf>
- Akhil Kedia, Sai Chetan Chinthakindi, WonHo Ryu, "Beyond Reptile: Meta-Learned Dot-Product Maximization between Gradients for Improved Single-Task Regularization" (2021)
<https://paperswithcode.com/paper/beyond-reptile-meta-learned-dot-product>
- Alpin, The Novice's LLM Training Guide (2023)
https://reentry.org/llm-training#low-rank-adaptation-lora_1
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, Attention is all you need (2017) <https://arxiv.org/abs/1706.03762>
- Ayub, Nafees, Muhammad Ramzan Talib, Muhammad Kashif Hanif, and Muhammad Awais. "Aspect Extraction Approach for Sentiment Analysis Using Keywords." Computers, Materials & Continua 74, no. 3 (January 1, 2023): 6879–92. <https://doi.org/10.32604/cmc.2023.034214>.
- Currie. "Summarizing Text with Amazon Reviews." Kaggle, May 10, 2017. <https://www.kaggle.com/code/currie32/summarizing-text-with-amazon-reviews>.
- Dorian Lazar, Medium, Scraping Medium with Python & Beautiful Soup (2020)
<https://dorianlazar.medium.com/scraping-medium-with-python-beautiful-soup-3314f898bbf5>
- Federico Pascual, Hugging Face, Getting Started with Sentiment Analysis using Python (2022) <https://huggingface.co/blog/sentiment-analysis-python>
- Felix Revert, Medium, An overview of topics extraction in Python with LDA Using LDA (Latent Dirichlet Allocation) for topics extraction from a corpus of documents (2018)
<https://towardsdatascience.com/the-complete-guide-for-topics-extraction-in-python-a6aaa6cedbbc>
- Fetahu, Besnik, Zhiyu Chen, Oleg Rokhlenko, and Shervin Malmasi. "Papers with Code - InstructPTS: Instruction-Tuning LLMs for Product Title Summarization." PapersWithCode, October 25, 2023. Accessed November 9, 2023.
<https://paperswithcode.com/paper/instructpts-instruction-tuning-llms-for>.
- François St-Amant, Towards Data Science, Fine-Tuning the BART Large Model for Text Summarization (2021)
<https://towardsdatascience.com/fine-tuning-the-bart-large-model-for-text-summarization-3c69e4c04582>
- GenAI, META, Llama 2: Open Foundation and Fine-Tuned Chat Models (2023) <https://arxiv.org/pdf/2307.09288.pdf>
- Hugging Face, Fine-tune a pretrained model
<https://huggingface.co/docs/transformers/training>
- Hujia Yu, Chang Yue, Chao Wang, Stanford University, News Article Summarization with Attention-based Deep Recurrent Neural Networks (2018)
<https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2746634.pdf>

Isabel Cachola, Kyle Lo, Arman Cohan, Daniel S. Weld, University of Washington TLDR: Extreme Summarization of Scientific Documents (2004) <https://arxiv.org/pdf/2004.15011.pdf>

Laskar, Md Tahmid Rahman, Xue-Yong Fu, Cheng Chen, and Shashi Shashi TN. “Building Real-World Meeting Summarization Systems Using Large Language Models: A Practical Perspective.” arXiv.org, October 30, 2023. <https://arxiv.org/abs/2310.19233>.

Lewis Tunstall, Leandro von Werra, Thomas Wolf, Natural Language Processing with Transformers, Revised Edition ISBN: 9781098136796 (2017)

Maryia Stsiopkina, Oxylabs, Scraping Amazon Product Data With Python: A Complete Guide (2023) <https://oxylabs.io/blog/how-to-scrape-amazon-product-data>

Maryia Stsiopkina, Scraping Amazon: A Step-by-Step Guide to Creating Your Own Scraper (2023) <https://dataprot.net/articles/how-to-scrap-amazon/>

Mathieu Busquet, Fine-Tuning LLaMA 2 Models using a single GPU, QLoRA and AI Notebooks (2023) <https://blog.ovhcloud.com/fine-tuning-llama-2-models-using-a-single-gpu-qlora-and-ai-notebooks/>

Mauro Di Pietro, Text Summarization with NLP: TextRank vs Seq2Seq vs BART (2022) <https://towardsdatascience.com/text-summarization-with-nlp-textrank-vs-seq2seq-vs-bart-474943efeb09>

Mike Ravine, Vinicius Arruda, GPU Utils, (2023) What’s the prompt template best practice for prompting the Llama 2 chat models? <https://gpus.llm-utils.org/llama-2-prompt-template/>

N. Moratanch; S. Chitrakala, A survey on abstractive text summarization (2017) <https://ieeexplore.ieee.org/abstract/document/7530193>

Omar Sanseviero, Pedro Cuenca, Victor Mustar, Inference for PROs (2023) <https://huggingface.co/blog/inference-pro>

OpenAI, Improved language understanding with unsupervised learning (2018) <https://openai.com/research/language-unsupervised>

Pragati Baheti, V7, Overfitting vs. Underfitting: What's the Difference? (2022) <https://www.v7labs.com/blog/overfitting-vs-underfitting>

Rachel Wolff, Visualize Your Sentiment Analysis in Tableau (2020) <https://monkeylearn.com/blog/tableau-sentiment-analysis/>

Rada Mihalcea, Paul Tarau, University of South Texas, TextRank: Bringing Order into Texts (2004), <https://web.eecs.umich.edu/~mihalcea/papers/mihalcea.emnlp04.pdf>

Rahul Katarya, Anjum, Anushka Gupta, Diksha Chugh, Automated News Summarization Using Transformers (2022) https://link.springer.com/chapter/10.1007/978-981-16-9012-9_21

Sebastian Raschka, Finetuning LLMs with LoRA and QLoRA: Insights from Hundreds of Experiments (2023) <https://lightning.ai/pages/community/lora-insights/>

Sergei Savvov, TLDR; techniques to speed up inference of LLMs to increase token generation speed and reduce memory consumption: mixed-precision, Bfloat16, quantization, fine-tuning with adapters, pruning, continuous batching, and multiple GPUs (2023) <https://betterprogramming.pub/speed-up-llm-inference-83653aa24c47>

Shivali Bhadaniya, Get Unique values in Pandas DataFrame Column (2022)
<https://favtutor.com/blogs/pandas-unique-values-in-column>

Stanford Network Analysis Project. “Amazon Fine Food Reviews.” Kaggle, May 1, 2017.
<https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews>.

Sthanikam Santhosh, Understanding BLEU and ROUGE score for NLP evaluation (2023)
<https://medium.com/@sthanikamsanthosh1994/understanding-bleu-and-rouge-score-for-nlp-evaluation-1ab334ecadcb>

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, Yoav Artzi, BERTScore: Evaluating Text Generation with BERT (2019)
<https://arxiv.org/abs/1904.09675>

Vincent Tan, Towards Data Science ,Fine-tuning pretrained NLP models with Huggingface’s Trainer (2021)
<https://towardsdatascience.com/fine-tuning-pretrained-nlp-models-with-huggingfaces-trainer-6326a4456e7b>

Waleed Kadous, Llama 2 is about as factually accurate as GPT-4 for summaries and is 30X cheaper (2023)
<https://www.anyscale.com/blog/llama-2-is-about-as-factually-accurate-as-gpt-4-for-summaries-and-is-30x-cheaper>

Zhang, Jingqing, Yao Zhao, Mohammad Saleh, and Peter J Liu. “PEGASUS: Pre-Training with Extracted Gap-Sentences for Abstractive Summarization.” arXiv.org, December 18, 2019.
<https://arxiv.org/abs/1912.08777>.

Zhang, Wenxuan, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. “Sentiment Analysis in the Era of Large Language Models: A Reality Check.” arXiv, May 24, 2023. Accessed November 9, 2023.
<https://arxiv.org/abs/2305.15005>.