

A Multi-Model Approach to Sentiment and Topic Analysis of Amazon Fine Food Reviews

1. Introduction

The rapid growth of e-commerce has led to a surge in product reviews. Prospective buyers often rely on experiences shared by previous customers to gauge product quality. But manually reading and understanding the large number of reviews is labour intensive and inefficient. Text analytics methods, including sentiment analysis and topic modeling, provide insights that can help businesses and customers quickly gauge the overall quality of various products.

This project uses the Amazon Fine Food Reviews dataset obtained from Kaggle (McAuley et al., 2013) containing 568,454 reviews of 74,258 Amazon food products from October 1999 to 2012. The project focuses on two analytics tasks: (1) Conduct **sentiment analysis** using logistic regression, DistilBERT, and Llama 3.2 Large Language Model, (2) Perform **topic modelling** using Latent Dirichlet Allocation (LDA) and BERTopic methodologies, including time series and hierarchical analysis to uncover evolving themes and subtopics within reviews.

By understanding the sentiment patterns and key themes, businesses such as Amazon and product manufacturers can better understand customer perceptions and rapidly address concerns. For instance, Amazon can optimize the product catalog and ensure that the platform carries products that align with consumer preferences. Product manufacturers can benefit from insights from negative reviews, such as that on packaging flaws, product quality or pricing, enabling manufacturers to respond accurately and quickly to customer demands.

2. Project Objectives

Manually examining over 500,000 Amazon Fine Food reviews is time consuming and inefficient. This project addresses these challenges by employing sentiment analysis and topic modelling techniques to systematically uncover insights to drive catalog optimization, product improvements and customer satisfaction. The two main objectives are to (1) classify and interpret customer sentiments - both positive and negative, and (2) uncover the dominant discussion themes in Amazon Fine Food reviews dataset.

To achieve this, the approach integrates multiple sentiment analysis techniques - logistic regression to provide a baseline, followed by DistilBERT that may better capture contextual nuances more effectively, and exploring Llama 3.2 that may provide better performance. By exploring these three different models, we aim to understand which type of models are most suited for sentiment analysis.

Topic modeling will first be performed using LDA as a foundational approach for identifying themes. LDA has its limitations (e.g., unable to understand word context), and hence, BERTopic will be used to extend this analysis with transformer-based embeddings for context-aware topic extraction. In addition, the BERTopic analysis would include time series analysis to observe how consumer review topics evolve over time and hierarchical modeling to reveal specific subtopics for a more nuanced view of consumer feedback.

3. The Data Set

The dataset used in this project is the **Amazon Fine Food Reviews dataset** obtained from Kaggle (McAuley et al., 2013). It includes 568,454 product reviews written by Amazon customers of Amazon food products from October 1999 to 2012 from 74,258 unique products. Each review includes a star rating (on a scale of 1 to 5), textual descriptions (covering product quality, packaging, taste, overall impressions), and metadata (e.g., product identifiers, timestamps).

This dataset has been widely utilized in research on sentiment analysis, product recommendation systems and topic modeling. The overall statistics and sample reviews from the Amazon Fine Food Reviews dataset are presented in **Appendix, Table A1**.

The dataset poses several key challenges for text analytics. First, reviews vary in length from fewer than 10 words to over 500 words (**Figure 1 in Appendix**) and contain various writing styles, vocabulary (including slang, emojis) and spelling errors. Second, the class imbalance complicates the analysis, as the distribution is skewed towards 5-star reviews (**Figure 2 in Appendix**), and 3-star reviews tend to represent a neutral sentiment that is harder to classify distinctly. Third, the project must balance computational efficiency with depth of understanding. Traditional models like logistic regression and LDA can handle large datasets efficiently but may not consider contextual nuances (Ali et al., 2024) such as sarcasm and negation. More advanced models such as DistilBERT and Llama 3.2 provide deeper contextual understanding (Sayeed, 2023), but it requires more computational resources.

4. Methodology

Figure 3 shows the proposed workflow for sentiment analysis and topic modeling. The proposed methodology starts by removing missing and duplicate reviews, followed by sentiment analysis and topic modeling.

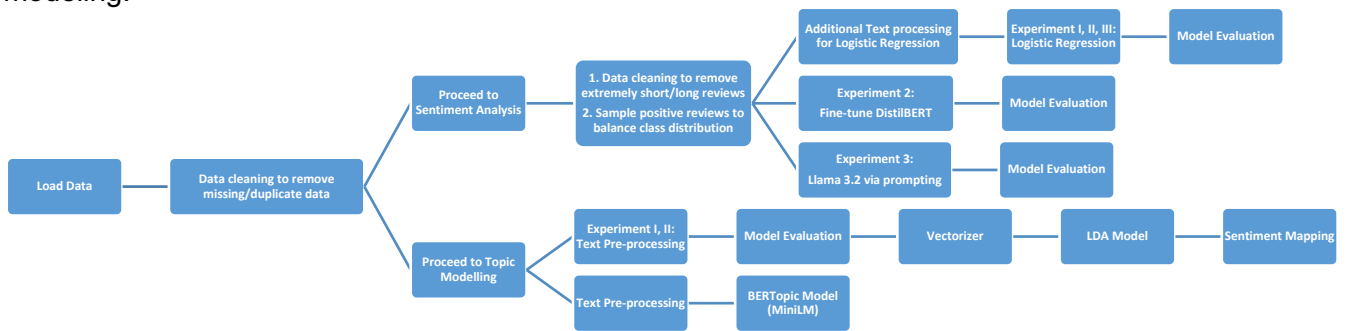


Figure 3: Sentiment Analysis and Topic Modelling Workflow

Sentiment Analysis

For data processing in sentiment analysis, the first step involved examining the distribution of reviews by word length (i.e., <10, 10-50, 50-100, 100-200, 200-400, 400-500, >500 words). Extremely short reviews (<10 words) often contain insufficient detail without explanation on why review is negative or positive. Conversely, reviews that are too long (>500 words) likely contain irrelevant information. To ensure data consistency, reviews that are in these extreme length categories were removed (Amat et al., 2023).

The second step focused on ensuring a balanced distribution of positive and negative labels was important to optimize model's performance (Ahmed et al., 2021). Reviews with score of 1-2 were classified as negative, and reviews with scores of 4-5 as positive. As shown in **Figure 4 in Appendix**, the initial dataset contained 56,817 negative reviews and 305,574 positive reviews. To achieve equal representation, 56,817 negative reviews were retained, and an equivalent number of positive reviews were randomly sampled. This resulted in 56,817 negative reviews and 56,817 negative reviews in the final dataset. The dataset was split into training and test sets in 80-20 ratio and used for subsequent experiments. Sentiment analysis with logistic Regression, DistilBERT and Llama 3.2 were performed. In all experiments, model performance was compared using accuracy, precision, recall, and F1-score.

Topic Modelling

The challenge for topic modelling is to extract coherent themes from a large dataset of unstructured reviews. Unlike the sentiment analysis pipeline, reviews in extreme length categories were not removed but having broader content can help identify more nuanced topics. However, for topic modeling, we restricted reviews to those with word counts between 10 and 500 to reduce noise from extremely short or excessively long texts. Judging from the distribution of reviews from 1999 to 2012 presented in **Figure 5 in Appendix**, there is a consistent increase in the number of reviews starting from 2006. Therefore, a filter was applied to include only data from 2006 onwards, resulting in 390,353 reviews for model training. The text preprocessing for LDA involved basic cleaning such as lowercasing and removing elements with no topical value including HTML tags, URL, stop words and non-alphanumeric characters. After text preprocessing, the LDA model is trained for topic discovery.

For BERTopic, only minimal text preprocessing was performed by removing special characters, extra spaces on the filtered dataset from 2006 onwards as used for LDA topic modeling. Typical text preprocessing cleaning steps such as removing stop words are not required to preserve contextual signals that are crucial for generating cohesive and context-aware clusters (Egger et al, 2022). The outcomes from both LDA and BERTopic will be compared based on topic coherence and diversity measures and qualitative assessments.

5. Solution Details

For sentiment analysis, logistic regression was selected as a baseline sentiment classifier due to its computational efficiency. However, it requires further text preprocessing such as lowercasing and removing non-informative elements (e.g., stop words and punctuations) and is limited by its ability to capture contextual nuances especially if the reviews contain sarcasm, negations, typographical errors, or slang. To overcome these limitations, we fine-tune transformer-based models, DistilBERT, a distilled variant of Bidirectional Encoder Representations Transformers (BERT). DistilBERT requires minimal text preprocessing because removing stopwords can reduce the models' ability to capture contextual nuances (Dorssers, 2024). It preserves the contextual nuances through minimal preprocessing and strikes a balance between performance and computational cost. Its reduced parameters count offers a faster alternative while maintaining its superior language understanding (Sanh et al, 2019). Lastly, Llama 3.2 was evaluated using prompting-based classification. Each review was passed to the model as a prompt and the model return a sentiment prediction. Llama 3.2 was selected as it can capture subtle linguistic features with higher precision as it is a larger model and trained on a much larger dataset. However, it is more computationally demanding (Roumeliotis, 2025).

For topic modeling, LDA is selected as a baseline for topic modelling due to its computational efficiency. However, it uses a bag-of-words representation which ignores contextual nuances and may produce less nuanced topics. To overcome this limitation, we use BERTopic which uses transformer embeddings to generate high dimensional representation of the text and clusters the embeddings to form context-aware topics that capture subtle linguistic features and word associations (Grootendorst, 2022).

To gain further insights, time series analysis (Grootendorst, 2022) will track how topics evolve over the years which can reveal the shifts in concerns and trends. Hierarchical analysis (Grootendorst, 2022) will break down broader themes into subtopics and offer deeper insights to customer reviews.

6. Experiments

Sentiment Analysis

There were three experiments conducted for Logistic Regression to understand the performance and limitations of the methods applied (**Figure 6**).

Logistic Regression Experiment I: Text was only vectorized by TF-IDF as input for training.

Logistic Regression Experiment II: Minimal Text processing was performed by lowercasing the text, and remove punctuation except for “!”, “...” and “?”, followed by TF-IDF.

Logistic Regression Experiment III: Extra preprocessing steps include converting text into lowercase, removing HTML tags, URL, stopwords, certain punctuations, converting emoticons and emoji into descriptive words, lemmatization, and TF-IDF.

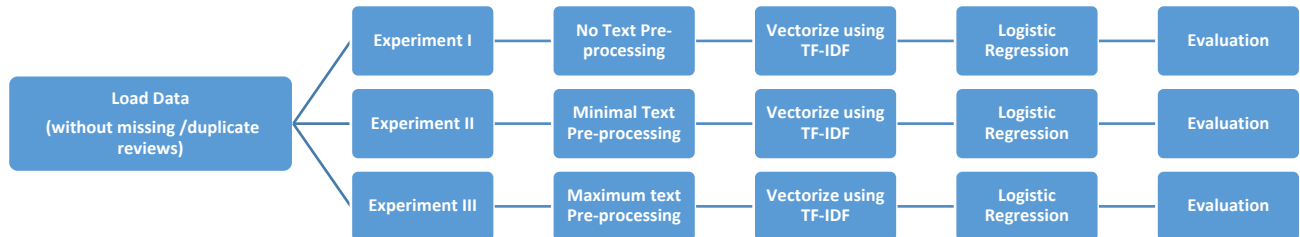


Figure 6: Experiments Conducted for Logistic Regression

Data used for DistilBERT and Llama 3.2 were not processed further to preserve contextual nuances. Experiments on DistilBERT was run on PyTorch while Llama 3.2 was run through Ollama. Both models require GPU (MPS) resources. For Llama 3.2, a prompt-based approach was used to perform sentiment analysis, and 12 representative examples (positive and negative sentiments, mixed and nuanced reviews) from the training set were provided. The prompt can be found in **Figure 7 of the Appendix**.

Model performance for logistic regression, DistilBERT and Llama 3.2 will be evaluated through classification report (accuracy, precision, recall, F1 score) and confusion matrix to identify the best performing sentiment classifier. Additionally, misclassified reviews will be further examined on whether certain features such as sarcasm, negations, typographical errors, or slang, contribute to errors.

Topic Modelling

For topic modelling, the preprocessed data will be used for LDA and BERTopic analysis. Pre-processing involves tokenizing words, converting them to lowercase, and removing HTML tags, URLs, stop words, and non-alphanumeric characters. These steps normalize the text corpus and eliminate case sensitivity issues that could distort the distribution of topics. Additionally, removing HTML tags, URLs, stop words (such as “this”, “is”, “the”, which contains little content on the topic), and non-alphanumeric characters helps to concentrate on the meaningful content of the text and minimizes any potential noise in the analysis.

Experiments on BERTopic will be trained using PyTorch on GPU resources (CUDA). Topic coherence and topic diversity scores (Grootendorst, 2022) will be used to compare the performance of both models, and qualitative inspection of generated topics and subtopics will be evaluated to identify the most coherent and dominant topics.

Topic Modeling – Latent Dirichlet Allocation (LDA)

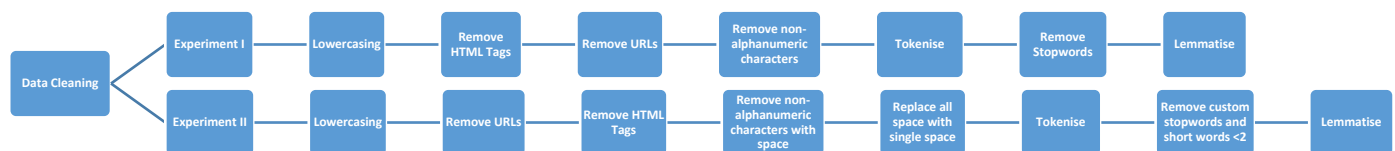


Figure 8: LDA Text-Preprocessing Process

LDA was applied using two different text pre-processing methods to measure and compare topic coherence and topic diversity scores. A count vectorizer was used to construct the document-term matrix before training the LDA model.

LDA Experiment I: The basic pre-processing steps included converting text to lowercase, removing HTML tags, URLs, and non-alphanumeric characters, tokenizing, removing stop words.

LDA Experiment II: Extra steps have been applied to pre-processing which include converting text to lowercase, replacing all punctuations with space, removing URLs, replacing white space with single space, removing non-alphanumeric characters while keeping spaces, tokenizing the text, defining custom stop words, remove stop words and filter out short words and lemmatize words then join words into a processed string. Custom stop words such as 'get', 'got', 'make', 'one', 'product,' 'would', 'really', 'also', 'even', 'since', 'never', 'ive', 'per', 'dont', 'ya', 'etc', 'ever' had been added after evaluating the result from Experiment I.

From the two experiments, it is worth noting that an overlooked detail such as removing HTML
 tags too early in the sequence would result in removing the '<' and '>' tags but keeping 'br' in the text. The regular expressions for removing non-alphanumeric characters have to be carefully applied to catch tags that are in HTML syntax as well. The result has shown that if basic pre-processing has been applied (LDA Experiment I), it is not enough to generate meaningful outputs.

Topic Modeling – BERTopic

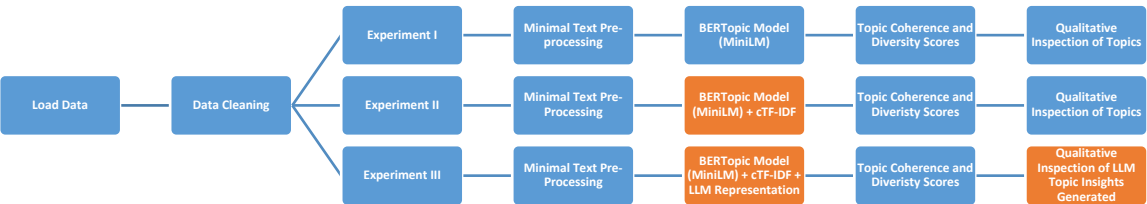


Figure 9: BERTopic Process

BERTopic Experiment I: In this experiment, we applied the BERTopic framework to uncover latent topics from the dataset. The documents were minimally preprocessed by lemmatization and basic cleaning deliberately retaining stopwords to allow the transformer-based model to capture more contextual nuance. We have used the all-MiniLM-L6-v2 embedding model for semantic representation, followed by dimensionality reduction with UMAP and clustering with HDBSCAN. BERTopic was then configured with a custom CountVectorizer to extract meaningful topic representations.

The model successfully identified 59 topics with a coherence score of 0.6906 and diversity score of 0.8169, indicating both semantic consistency and lexical variety in the topic keywords. Additionally, we visualised topic relationships with a hierarchical dendrogram and explored how selected topics evolved over time with a temporal breakdown of the dataset.

BERTopic Experiment II: In the second experiment, we enhanced topic representation by incorporating class-based TF-IDF (cTFIDF) by reducing frequent words. This approach down weights overly common terms and boost words that are more distinctive within each topic cluster, while the TC score decreased slightly to 0.6804, the TD score improved significantly to 0.9372 which indicates a broader lexical spread across topics. This would provide a clearer segmentation of customer reviews into unique themes – supporting more targeted product improvement, marketing personalization on the nuanced consumer insights. The topic representation would be further compared in section 7.

BERTopic Experiment III: In the third experiment, we enhanced topic labeling by using a LLM (Zephyr 7B Alpha - GGUF) for inference through llama.cpp in Python. The LLM was used to generate a topic label based on the keywords in each document passed through BERTopic. The TC improved significantly to 0.7722 while TD decreased slightly to 0.9224 with 156 topics generated.

7. Results and Analyses

Sentiment Analysis

For all experiments conducted for sentiment analysis, they demonstrated robust performance on 22,727 labelled reviews (11,301 negative and 11,426 positive), achieving an overall accuracy, F1-score, precision and recall of between 0.88 to 0.95, as shown in **Table 1**. Each of the models have similar precision, recall, and F1 scores for both sentiment classes, suggesting that it is not biased towards any class, which is essential in maintaining fairness and robustness in predictions. The detailed analysis of some sample predictions for all three models are provided in the **Appendix (Figures 8-12 and Table A2-A7)**.

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression (No Text Preprocessing)	0.8967	0.89	0.90	0.90
Logistic Regression (Minimal Text Preprocessed)	0.8968	0.89	0.90	0.90
Logistic Regression (Maximal Text Preprocessed)	0.8886	0.88	0.89	0.89
DistilBERT	0.95	0.95	0.95	0.95
Llama 3.2 (1B)	0.91	0.91	0.91	0.91

Table 1: Performance Metrics of Sentiment Analysis Models

Logistic Regression

The logistic regression on unprocessed data (accuracy: 0.8967; Logistic Regression Experiment I) outperforms maximally processed data (accuracy 0.8886; Logistic Regression Experiment III) but slightly worse than minimally processed data (accuracy: 0.8968; Logistic Regression Experiment II). There are several explanations for this outcome. First, some removed stop words (e.g., “not) may have sentiment meaning. Second, when transforming the emoji into words, word-based sentiment may not fully capture the emotion intensity. Third, preprocessing could disrupt word order, altering meaning and cause a loss of sentiment. Although the logistic regression model detects direct sentiment words well and has an accuracy of around 0.89, but it misses contextual nuances such as sarcasm, negation, slang. Overall, this analysis serves as baseline for comparison with DistilBERT and Llama 3.2.

DistilBERT and Llama 3.2 (1B)

The DistilBERT model excels at detecting contextual nuances such as sarcasm, negation, slang, and spelling errors. The confusion matrix displays a moderate number of misclassifications which could be attributed to multiple factors such as ambiguity or mixed reviews. As shown in **Table A3**, DistilBERT is able to correctly predict the sentiment even when there are typographical errors, slang and some mixed sentiments. However, when the review has an even mix of both negative and positive sentiments as seen in **Table A4**, the model struggles to predict accurately.

The Llama 3.2 model consistently identifies the primary sentiment correctly even in reviews containing subtle sarcasm or mixed elements as shown in **Table A5**. These examples also demonstrate the model can effectively handle contextual nuances such as negation, sarcasm, slang and typographical errors. This robust performance is valuable for e-commerce platforms seeking efficient and reliable assessment of customer feedback. However, **Table A6** shows that most misclassifications arise when reviews mix positive and

negative elements. For example, a reviewer praised olive oil as ‘great value’ yet found it ‘on the young side’, leading to a negative classification. Another reviewer endorsed Fromm’s cat food but mentioned unrelated negative experiences, causing the model to classify the review as negative. These cases highlight the model’s difficulty in discerning which statements directly apply to the product’s overall evaluation.

Topic Modeling

To evaluate the effectiveness of LDA and BERTopic, we utilized two commonly used metrics: Topic Coherence (TC) and Topic Diversity (TD). Topic coherence was measured using c_v coherence metric (Röder et al., 2015), which combines normalized pointwise mutual information (NPMI), cosine similarity and a sliding window approach to assess the semantic consistency of topics. Topic Diversity was calculated as the proportion of unique words across all topics, indicating the model’s ability to generate non-redundant topics (Dieng et al., 2020). In LDA, perplexity score has also been measured then the models were trained on 10, 15, 20, 25, and 30 topics for comparison.

	Experiment 1			Experiment 2		
Topics	TC	TD	Perplexity	TC	TD	Perplexity
10	0.4762	0.5600	1928.7037	0.5169	0.5900	1679.3177
15	0.4997	0.5467	1884.1367	0.5364	0.6133	1638.7873
20	0.5239	0.6050	1862.2900	0.5439	0.6150	1622.8851
25	0.5129	0.6000	1867.4068	0.5441	0.6080	1632.0438
30	0.5428	0.6233	1845.4796	0.5599	0.6200	1616.4570

Table 2: LDA Experiment Result

With additional pre-processing applied, all topic models in LDA’s Experiment II achieved higher evaluation scores compared to those in Experiment I. Based on the metrics of perplexity, topic coherence, and topic diversity, 20 topics in Experiment II demonstrate the most balanced performance. This model achieves a low perplexity score of 1622.88, while maintaining a coherence score of 0.5439 and a topic diversity score of 0.6150.

Experiment	TC	TD
BERTopic Experiment I [MiniLM]	0.6906	0.8169
BERTopic Experiment II [MiniLM (with cTF-IDF)]	0.6804	0.9372
BERTopic Experiment III [MiniLM with cTF-IDF & Zephyr 7B LLM]	0.7722	0.9224

Table 3: BERTopic Result

The results show that BERTopic consistently outperformed LDA in both TC and TD across all experiments. Notably, with the integration of the LLM as a representation model in the BERTopic Experiment III achieved the highest TC while maintaining a high TD, indicating that the LLM assisted topic labeling significantly enhances the semantic quality and substantially improving the interpretability and effectiveness of topic modeling.

Results & Analysis of LDA

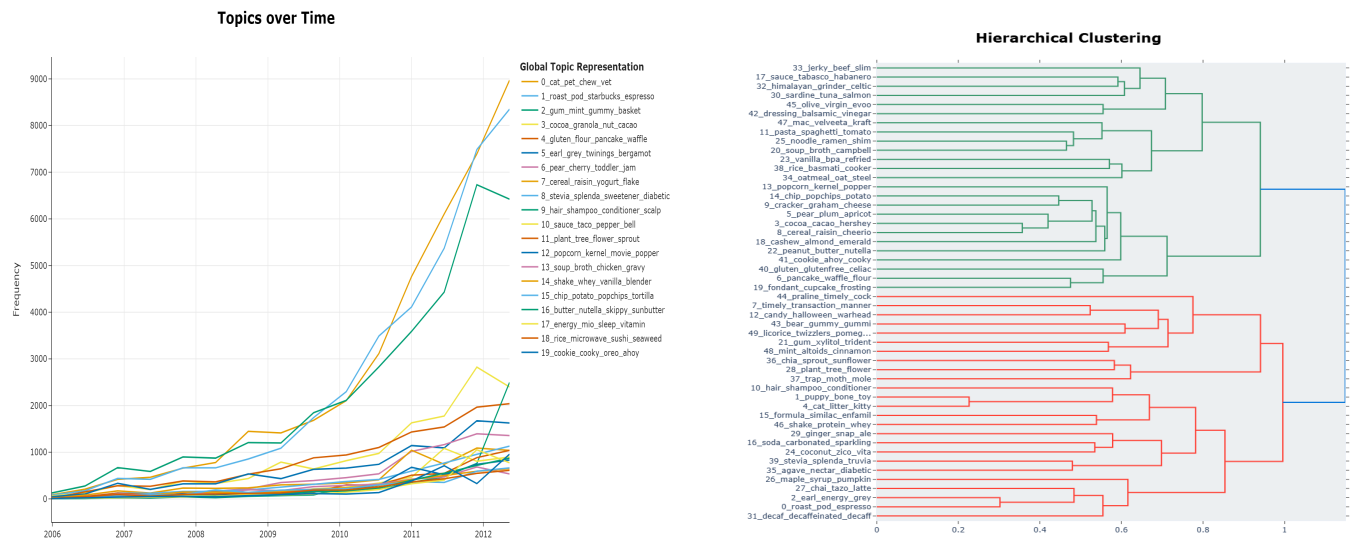


Figure 16: Temporal Analysis (left) and Figure 17: Hierarchical Clustering (right)

The temporal analysis from BERTopic reveals how different product-related themes evolved in Amazon reviews from 2006 to 2012. The result highlights notable shifts in consumer interest and market trends across the years. Topics which were of particular interest were: Topics 0, 1 and 2 experienced sharp growth starting around 2010, with topics like cat_pet_chew_vet, ‘roast_pod_starbucks_espresso’ and “gum_mint_gummy_basket”. This shows rising consumer interest in coffee-related products, confectionary, and gluten-based goods. Conversely many other topics maintained relatively low and stable frequencies throughout the analysis period.

The hierarchical clustering displays the similarity structure among 59 topics, grouped based on their content representations. The dendrogram branches reveal two main clusters (red and green) with suggesting two distinct themes or groupings. The cluster predominantly includes topics related to savory and staple food items, such as “jerky_beef_slim”, “sardine_tuna_salmon”, “rice_basmati_cooker” while the red cluster includes more sweet, snack and health related products such as “bear_gummy_gummi”, “gum_xylitol_trident”.

Index	Topic	Keywords	Insight
0	0	cereal, bar, cup, nut, blend, cat, snack, breakfast, chocolate, roast	Answer: Topic 0 is about snack foods for cats that are not necessarily traditional cat food, but rather breakfast or snack options for cats that are also suitable for human consumption. This topic includes a variety of ingredients, such as cereal, nuts, and roasted items, which provide a balanced blend of nutrients for cats. This topic might be useful for Amazon in expanding their product line for pet food, especially for niche markets that cater to health-conscious cat owners who prefer feeding their cats with human-grade food. The Fine Food Departments, on the other hand, can use this topic to identify potential ingredients for their fine food products that are also suitable for cats, as there is an increasing demand for pet-friendly fine food options. This can lead to the development of new product lines that cater to both human and pet food needs.
1	1	dog food, dog, food dog, beneful, puppy, kibble, coat, stool, dry dog, feeding	This topic focuses on dog food, specifically dry kibble for puppies and adult dogs. The coherence score suggests that the documents in this topic are closely related to this topic and have a clear focus. This information may be useful for Amazon and the Fine Food Departments in terms of product development and marketing. The topics suggest that beneful is a popular brand and that coats and stool are important concerns for dog owners. Amazon can use this information to improve their product offerings, marketing strategies, and customer support to meet the demands of dog owners.
2	2	chip, potato chip, potato, popchips, tortilla, love chip, salt vinegar, tortilla chip, kettle, pop chip	Topic 2 is primarily focused on chips, with an emphasis on potato chips and brands such as Popchips and Kettle. This topic is about consumers' preferences for and experiences with various brands of chips. The representative document provides an example of a consumer who had a problem with a product and had to contact the seller. This topic could be useful for Amazon and the Fine Food Departments in terms of understanding consumer preferences for certain brands of chips, identifying common issues and addressing them, and improving overall customer satisfaction. By analyzing this topic, Amazon can also identify which brands and flavors are most popular and ensure that they have a strong supply chain to meet demand.
3	3	email, wa exactly, timely, business, seller, vendor, transaction, item wa, product arrived, great service	1. This topic is about emails, timely transactions, business-related activities, and interactions with vendors and sellers, regarding the purchasing of products such as carnival theatre kettle style popcorn popper, where the focus is on its great service and how it compares to movie theatre popcorn. 2. For Amazon, this topic represents the customer's purchasing process, from email inquiries to the successful transaction and delivery of the product. The Fine Food Departments can utilize this topic to better understand the customer's needs and preferences, as well as identify any areas for improvement in the purchasing process. This insight can help Amazon to optimize its product offerings, improve customer satisfaction, and ultimately increase sales. Additionally, the Fine Food Departments can use this topic to identify trends and patterns in customer demands, such as a preference for popcorn products, and leverage these insights to develop new products, improve existing ones, and optimize product pricing.

Figure 18: Topic Insights - Prompted and generated from LLM – Zephyr 7B

Using the keywords generated from each topic, the Zephyr 7B LLM representation model produced concise and semantically coherent summaries for each topic. In the prompt, we included an instruction to the

model passing it as an ‘You are an NLP expert working for Amazon under the Fine Food Departments.’ and also for the model to (1) Briefly explain what each topic is about.
(2) Suggest how each topic might be useful for Amazon and how the Fine Food Departments, in terms of product characteristics or demands.

Notably the model demonstrated strong alignment between keywords and the generated descriptions by effectively contextualizing customer preferences (e.g. Topic 0 – increasing demand for pet-friendly fine food), product segments (e.g. Topic 2 – snacks, but specifically brand and flavours are popular to address supply issues). The relevance and clarity would provide outputs sufficient for a headline analysis allowing for the relevant departments to look further into. This reinforces the earlier quantitative results where the LLM-enhanced BERTopic model achieved the highest TC with a strong TD.

8. Discussions and Gap Analysis

Comparison of Logistic Regression, DistilBERT and Llama 3.2 in Sentiment Analysis

DistilBERT achieved the highest accuracy and F1-score of 0.95 due to its transformer-based architecture and fine tuning on sentiment data. Logistic regression, used as baseline, achieved accuracy of 0.89 and F1-score of 0.90 but struggled with complex nuances. Both DistilBERT and Llama 3.2 struggle with reviews containing mixed sentiments, as shown in **Table A4** and **Table A6**, yet DistilBERT more reliably identifies the dominant sentiment. This is likely because DistilBERT was fine-tuned on sentiment data, while Llama 3.2 depends on prompting with examples.

Comparison of LDA and BERTopic

The comparison between LDA and BERTopic highlights key methodological and performance differences. LDA, while effective in extracting general themes, is highly sensitive to preprocessing and lacks the ability to capture deep contextual relationships, resulting in lower topic coherence and diversity. In contrast, BERTopic consistently produced more coherent and diverse topics through transformer-based embeddings and clustering. Enhancements such as cTF-IDF and LLM-based labeling further improved BERTopic’s interpretability and semantic richness. The primary gap identified in LDA is its limited capacity for nuanced contextual meaning, where BERTopic addresses more robustly. Although it may require additional computational resources and parameter tuning for optimal performance, its performance makes it the more suitable choice for complex, real-world text data.

9. Future Work and Conclusion

In this project, we performed sentiment analysis on product reviews using logistic regression, DistilBERT and Llama 3.2. DistilBERT achieved the best performance with accuracy and F1-score of 0.95, although all models struggled with mixed sentiment. Future work should focus on further fine-tuning Llama 3.2 with larger sentiment datasets to resolve conflicting cues. This can benefit businesses by providing more reliable insights into customer opinions, which can improve product development and marketing strategies to improve customer satisfaction.

In addition to sentiment analysis, we explored topic modeling using BERTopic to extract key themes from customer reviews. BERTopic, enhanced with transformer-based embeddings, clustering, class-based TF-IDF generated a high topic coherence (0.77) and topic diversity score (0.92). While the LLM representation improved topic coherence and also improved interpretability of results. Future work could include multi-aspect topic modeling using BERTopic to analyse how topics vary across different dimensions such as product categories, time periods or customer segments. This may reveal deeper insights, such as how customer

concerns evolve over time or differ between product types. Additionally, integrating sentiment scores with topic clusters could provide a more comprehensive view of customer feedback. This combination can support more targeted improvements in product marketing and customer experience strategies.

References

- McAuley, J., & Leskovec, J. (2013). From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews. In Proceedings of the 22nd International Conference on World Wide Web (pp. 897–908). <https://doi.org/10.1145/2488388.2488466>
- Ali, H., Hashmi, E., Yayilgan, S. Y., & Yamin, M. M. (2024). Analyzing Amazon products sentiment: A comparative study of machine and deep learning, and transformer-based techniques. *Electronics*, 13(7), 1305.
- Sayeed, S. (2023). BERT: A review of applications in sentiment analysis. *HighTech and Innovation Journal*, 4(2), 423–433.
- Amat-Lefort, N., & Barnes, S. J. (2023). Towards more convenient services: A text analytics approach to understanding service inconveniences in digital platforms. Proceedings of the 56th Hawaii International Conference on System Sciences (HICSS), 1346–1355.
- Ahmed, H. M., Awan, M. J., Khan, N. S., Yasin, A., & Shehzad, H. M. F. (2021). Sentiment analysis of online food reviews using big data analytics. *Elementary Education Online*, 20(2), 827–836.
- Dorssers, O. P. W. (2024). Sarcasm Detection in Structured Text using DistilBERT: Evaluating the Impact of Text Normalization on Model Performance (Master's thesis, Tilburg University). <http://arno.uvt.nl/show.cgi?fid=171929>
- Egger, R., & Yu, J. (2022). A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. *Frontiers in Sociology*, 7, 848150.
- Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv*. <https://arxiv.org/abs/1910.01108>

Roumeliotis, K. I., Tselikas, N. D., & Nasiopoulos, D. K. (2025). Optimizing airline review sentiment analysis: A comparative analysis of LLaMA and BERT models through fine-tuning and few-shot learning. *Computers, Materials & Continua*, 74(1), 1015–1033. <https://doi.org/10.32604/cmc.2025.059567>

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure [arXiv preprint arXiv:2203.05794]. <https://arxiv.org/abs/2203.05794>

Appendix

Statistic	Description/Value
No. of reviews	568,454
No. of products	74,258
Review Length	Varies from fewer than 10 words to more than 500 words.
Sample Review (5 star)	"Great product, nice combination of chocolates and perfect size! The bags had plenty, and they were shipped promptly. The kids in the neighborhood liked our candies!"
Sample Review (4 star)	"good flavor! these came securely packed... they were fresh and delicious! i love these Twizzlers!"
Sample Review (3 star)	"I LOVE spicy ramen, but for whatever reasons this thing burns my stomach badly and the burning sensation doesn't go away for like 3 hours! Not sure if that is healthy or not and you can buy this at Walmart for \$0.28, way cheaper than Amazon."
Sample Review (2 star)	"I can't eat these oats, they have a funny taste to them. My kids also think they taste funny. My husband thinks they are OK, so he has 50 pounds of oats to eat on his own!"
Sample Review (1 star)	"This is an awful coffee. The flavor is extremely weak and tastes more like a lemon grass tea than coffee."

Table A1: Key Statistics and Examples from the Amazon Fine Food Reviews Dataset

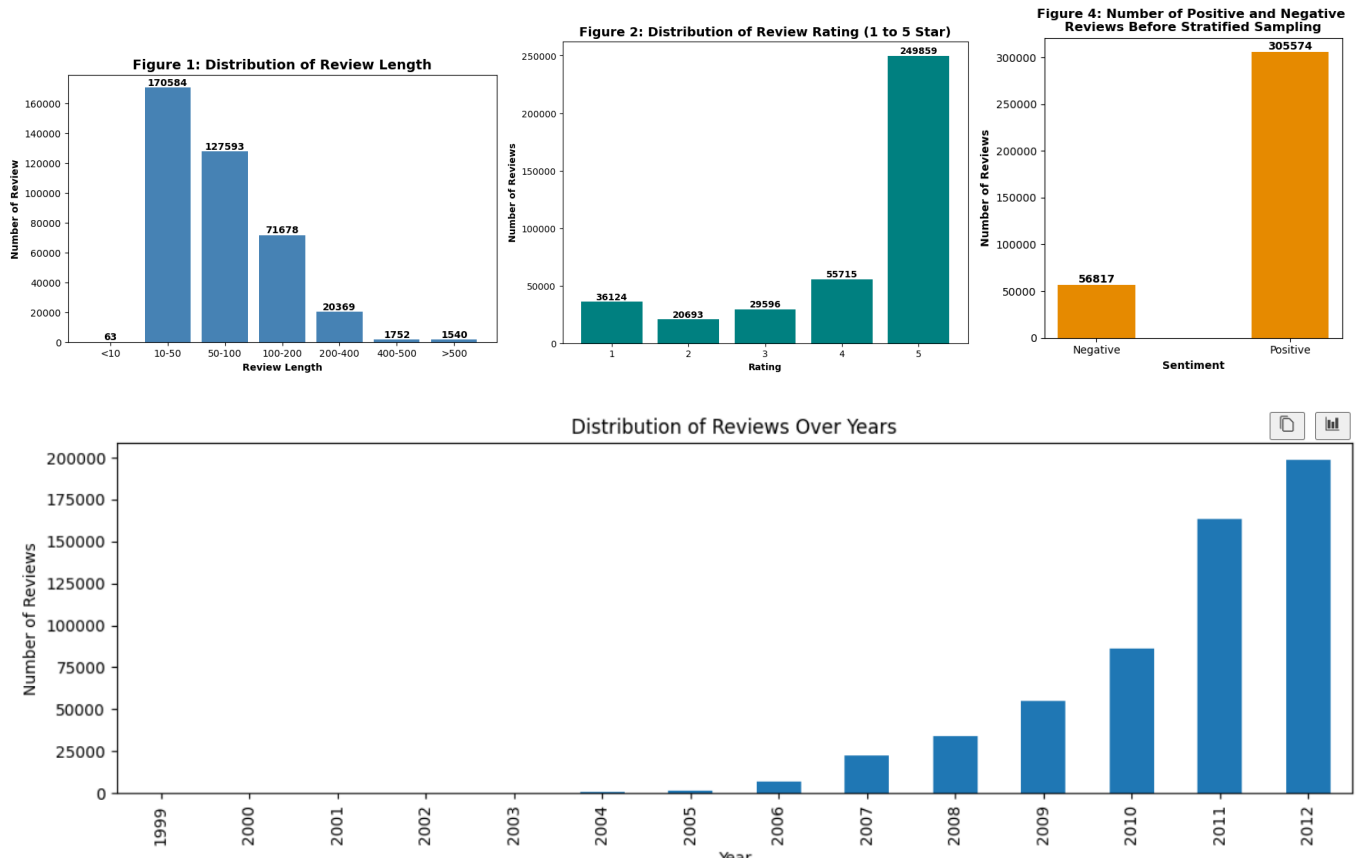


Figure 5: Distribution of Reviews over Years

You are a sentiment classifier for Amazon Fine Food reviews.

Output exactly '0' (no quotes) if the review is negative, or '1' if the review is positive.

No explanation, no extra words, just the single digit '0' or '1' (without quotes).

For the review below, determine the overall sentiment by weighing all positive and negative aspects. If the majority of the review is positive, return 1; if it's negative, return 0. Even if the review contains some negative points, if the overall tone is positive, output 1.

Here are some examples:

Review: Great buy for the money. Almonds are large and whole. I am planning to order more almonds plus other nuts from this company.

Sentiment: 1

Review: I was very disappointed with my order. The first thing, I did not receive the complete order. When I called, they said they could not help me. I only received one Barrel of Pretzels. Second, the one Barrel I did receive looked just awful. I do not even want to open it. I am a Pretzel Lover, and Utz is one of my Favorites. I do not know what happened, but I will never buy this item again through Amazon. I am even concerned about ordering any food item through Amazon. I have ordered many items through Amazon and I never had a problem until now.

Sentiment: 0

Review: Very Convenient, can take with you to work, on trips, etc. Excellent product.

Sentiment: 1

Review: As much as our dogs enjoyed Lickety Stik, the cloying stench of "savory chicken" repulsed the rest of the family. The odor wafted off the stick and filled the room like a musky air freshener. We called it "Icky Stik" or "Stinky Stik" as the experience was so nauseating that no one volunteered to reopen the stick, despite pleas from our dogs.

Sentiment: 0

Review: This tea is delicious! I've noticed my skin is starting to look a lot better since I have been drinking it. I have a cup almost every day. I will be purchasing more when I run out!

Sentiment: 1

Review: I am unhappy with the purchase on the grounds of deceptive labeling. This tea is NOT matcha, but rather a matcha blend that is mostly sencha. I should have read the product description more carefully, but was duped by the product listing title. Amazon should change the listing to Rishi Tea... Matcha blend.

Sentiment: 0

Review: I would give this company zero stars if I could. I purchased these flowers for my grandmother's birthday, but she never received them. When I contacted the company's customer service, I was told that the flowers were delivered, but no one was home, so they were left at the front door. I was also told that I cannot receive a refund or a delivery with signature. Hopefully, the Better Business Bureau resolves this issue.

Sentiment: 0

Review: I got these for a reason. My eight year old Elizabeth and I make our Christmas cookies (at least one kind) every year and they need decorations; now we have a lot!
Sentiment: 1

Review: The item was delivered in great condition and my dog enjoyed the tug toy, but after a few uses a small piece chipped off and it could have been dangerous if ingested.
Sentiment: 0

Review: I wanted so badly to love these peanuts. The texture is absolutely perfect; however, the salt quickly sneaks up on you and makes them painful to eat.
Sentiment: 0

Review: These brownies are really, really good and remind me of homemade box-mix brownies. However, they only make an 8x8 pan and the consistency isn't ideal for a sundae or layered dessert.
Sentiment: 1

Review: I generally like this brand's flavored coffees; however, the Turtle Sunday flavor I bought did not suit my taste, even though the overall quality cannot be beat.
Sentiment: 1

Now classify the following review:

Review: "{review_text}"

Answer:

Figure 7: Prompt used for Llama 3.2 Sentiment Analysis

Logistic Regression - Unprocessed Data

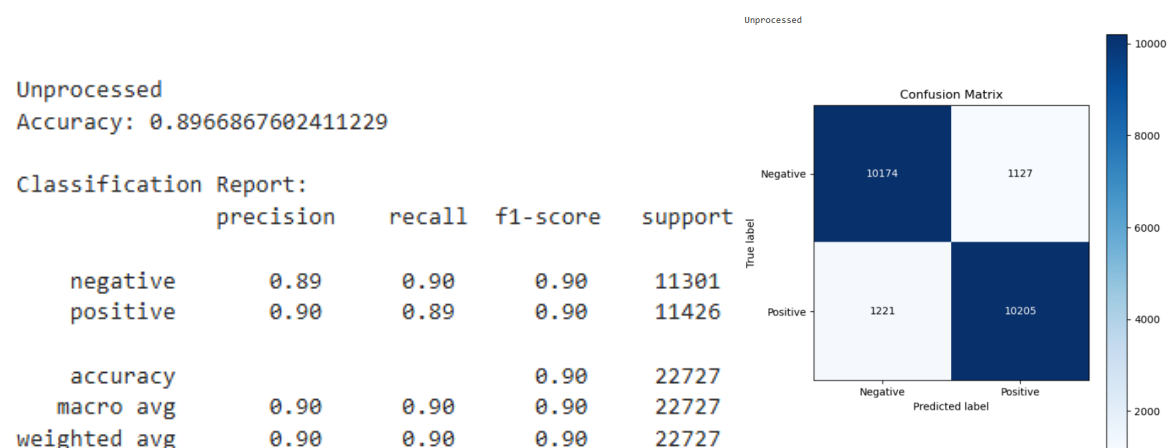


Figure 8: Classification Report and Confusion Matrix for Logistic Regression (Unprocessed Data)

Logistic Regression – Minimally Processed Data

Minimally processed

Accuracy: 0.8967747612971355

Classification Report:

	precision	recall	f1-score	support
negative	0.89	0.90	0.90	11301
positive	0.90	0.89	0.90	11426
accuracy			0.90	22727
macro avg	0.90	0.90	0.90	22727
weighted avg	0.90	0.90	0.90	22727

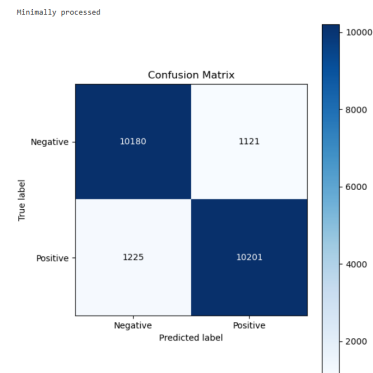


Figure 9: Classification Report and Confusion Matrix for Logistic Regression (Minimally processed Data)

Logistic Regression – Maximally Processed Data

Maximally processed

Accuracy: 0.8885906630879571

Classification Report:

	precision	recall	f1-score	support
negative	0.88	0.89	0.89	11301
positive	0.89	0.88	0.89	11426
accuracy			0.89	22727
macro avg	0.89	0.89	0.89	22727
weighted avg	0.89	0.89	0.89	22727

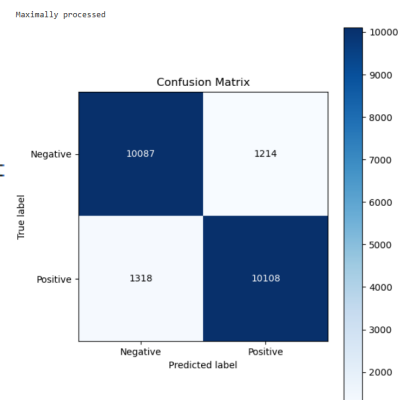


Figure 10: Classification Report and Confusion Matrix for Logistic Regression (Maximally Processed Data)

DistilBERT

Classification Report:

	precision	recall	f1-score	support
0	0.9500	0.9533	0.9516	11301
1	0.9536	0.9504	0.9520	11426
accuracy			0.9518	22727
macro avg	0.9518	0.9518	0.9518	22727
weighted avg	0.9518	0.9518	0.9518	22727

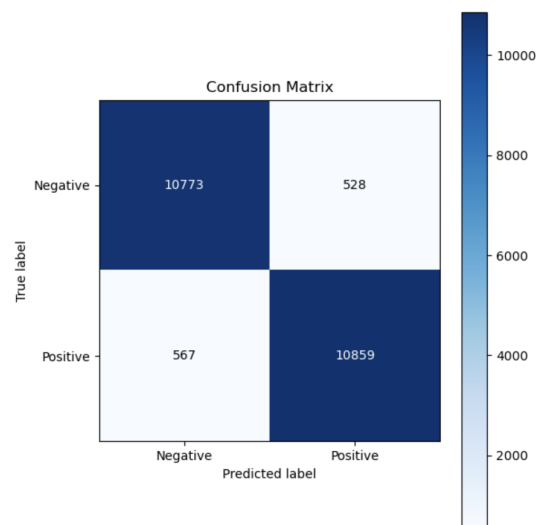


Figure 11: Classification Report and Confusion Matrix for DistilBERT

Llama 3.2

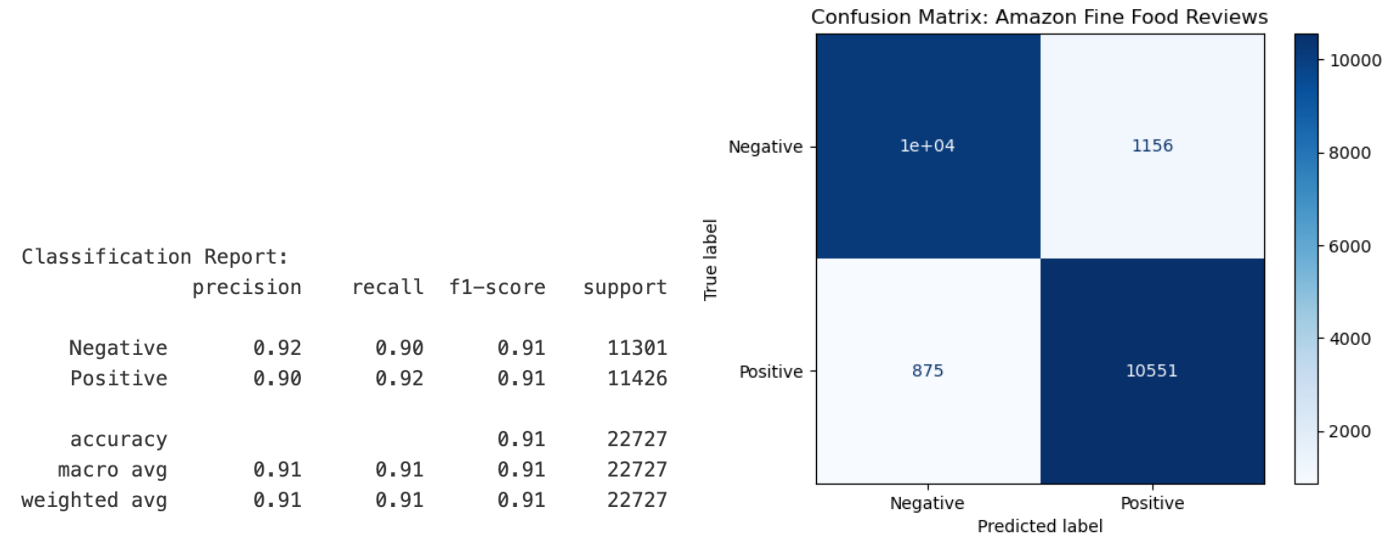


Figure 12: Classification Report and Confusion Matrix for Llama 3.2

Review	Ground Truth Label	Unprocessed Predicted Label	Minimal unprocessed Predicted Label	Maximal unprocessed Predicted Label	Reason for Mismatched
I lost faith in my long time cat food, Science Diet, after the massive pet food recall. Fromm claims this dry food is made in the US (Wisconsin) from meat, cheese and vegetables bought from local growers. Quality meats, no glutens, fresh vegetables and human quality vitamins. The cats like it. It is more expensive than Science Diet, but not prohibitively so. Note: the Milwaukee newspaper advises the canned Fromm is made in China. I have absolutely no faith in the Chinese quality and safety control standards so I would not buy the canned foods.	Positive	Negative	Negative	Negative	The text contains more negative sentiments than positive, however, the main sentiment is with the positive sentiment Note: Text with positive sentiment is highlighted in green and negative sentiment in red.
My daughter loves these! I know the box says toddler, but like the previous reviewer said, there is little difference between these and the baby variety. These might be slightly thicker and they are a round. I love Mum Mums. Such a great alternative to Wagon Wheels and Biter Biscuits, which can be a choking hazard . And the previous reviewer who calls these " expensive " needs to learn to read, or at least, read more carefully before leaving a review. There are 6 boxes in this pack. They sell for about	Positive	Positive	Positive	Negative	The text contains strong positive sentiments but when maximally processed with stop words removed, these sentiments are lost

\$3.50 per box in the store, so it's not a bad deal at \$18 for a 6 pack.					
Got this for Christmas last year. It drains electricity like crazy - my bill was almost doubled! In addition, the lights are on in the middle of the night and are nearly blinding. Not worth it for herbs you can easily grow outside in pots.	Negative	Positive	Positive	Positive	Some of the negative sentiments (red) may be lost when texts were processed leading to incorrect prediction
HP Sauce is my favourite "brown sauce" and this deal was cheaper then buying it locally at the supermarket, so what's not to like ;)	Positive	Negative	Positive	Positive	Unprocessed was not able to pick up the positive sentiments and likely considered the "not" as negative
This large-leaf black tea has bits of orange peel and "other citrus" mixed in, something I didn't expect from the product description. My preference is for straight unflavored tea, but as additives go, a little tangy fruit flavor isn't bad . The underlying tea is of excellent quality - as have been all my Hyley's purchases - so I'll look over their other products to see if I can find the same tea without the fruit.	Positive	Negative	Negative	Positive	Some of the sentiment colored in red is supposed to be positive sentiment but may have been misinterpreted by the model when the data was not processed and stop words not removed

Table A2: Analysis of Mismatch Reviews from Logistic Regression model

Review	Sentiment (Ground truth and DistilBERT)	Analysis
It does have a very good taste!!!! However , when I read the process of these beans for sumatra, ugh , I am done for good !!!!	Negative	Although the user has indicated that the product has very good taste, the model is able to rightly identify that the overall sentiment is negative. Note: Text with positive sentiment is highlighted in green and negative sentiment in red .
I bought a box, waist'd my \$\$ but people love me that I could give them this!!! AFTER having a bough of HAVING to rush to bathroom I re-read the posts! AND looked at ingredients closer! Gluten Free Whole Grain Dextrins IF you don't have a problem GREAT ! If you do you may end up re-reading the 1's later -} I like the idea of Bites had them before. going to try out Pamela's. I Love the Nana's Cookies but IT's VERY hard Not to eat the Whole cookie! The Nana's No Gluten Lemon Cookie, 3.5-Ounce Cookies (Pack of 12) & <a	Negative	Even though there is a typographical error where user indicated that they wasted (waist'd) their money and there are plenty of positive comments, DistilBERT classified this as a negative review.

<p>href="http://www.amazon.com/gp/product/B000FNB3BC">Nana's No Gluten Ginger Cookie, 3.5-Ounce Packages (Pack of 12) It's TWO servings!! Plus the texture is soooo smooth it go's down VERY easy! I don't have Coffee or Chocolate anymore! Have Green Tea Decaf.
NEED to go to the Nana's Banana Bars + modify or delete my post!
Only Bites I hope Pamela's ginger works out for me!
I Do LOVE this TOO Lundberg Eco-Farmed Honey Nut Rice Cake, 9.5-Ounce Units (Pack of 12) +Lundberg Eco-Farmed Buttery Caramel Rice Cake, 9.5-Ounce Units (Pack of 12) + Lundberg Eco-Farmed Brown Rice Cake, Salt Free, 8.5-Ounce Units (Pack of 12)</p>		<p>Note: Text with typographical error is highlighted in golden.</p>
<p>Tres Rios Valdivia was excellent coffee, I was in Costa Rica in May and fell in love with Brit coffee it just tates like the country..the air, the earth, the water fantastico!!
Steven Kilts</p>	<p>Positive</p>	<p>Model did not misclassify when slang words were used and there are typographical errors.</p> <p>Note: Text with slang is highlighted in pink. Text with typographical error is highlighted in golden.</p>
<p>On Sept.12 I posted the following seller review:
 3/5 stars:
 "I'm sorry to say that this item arrived with some slight cosmetic damage. On the top front edge of the rack there is what I would call a dent, but it's also rough. From the way it was packaged I would have to say that it was that way when it was put into the box. I intended this as a gift, and I don't have time to return it, so I'll just try to sand it down so it's at least smooth."
 Jammies, September 12, 2012
Later Amazon added:
 Message from Amazon: "This item was fulfilled by Amazon, and we take responsibility for this fulfillment experience."

Shortly after I posted it, I received an email from The Spice Lab Inc. offering to send a new block. I sent them a couple of pictures of the damaged area. They sent me a new block and two little spoons to replace the one that was missing from my original order. The replacement block arrived yesterday in plenty of time for my trip to visit my son and daughter-in-law.

In hindsight, I should have contacted The Spice Lab before I posted the review, so they could have had time to resolve the problem. I had jumped to the conclusion that I'd need to return the item in order to get a replacement, which wasn't the case.

Since it's a gift, I can't speak to the quality of the salt, but I can say that the block is very sturdy and attractive, and the little spoon is a charming addition.

I greatly appreciate the efforts of Brett at The Spice Lab, and I would happily buy from them again.</p>	<p>Positive</p>	<p>The first half of the review is wholly negative since it is referring to initial purchase, but the model has accurately classified the overall sentiment as positive based on the remainder of the review.</p>
<p>I lost faith in my long time cat food, Science Diet, after the massive pet food recall. Fromm claims this dry food is made in the US (Wisconsin) from meat, cheese and vegetables bought from local growers. Quality meats, no glutens, fresh vegetables and human quality vitamins. The cats like it. It is more expensive than Science Diet, but not prohibitively so. Note: the Milwaukee newspaper</p>	<p>Positive</p>	<p>Despite the review containing more negative phrases, the model can correctly identify the overall sentiment to be positive.</p>

advises the canned Fromm is made in China. I have absolutely no faith in the Chinese quality and safety control standards so I would not buy the canned foods.		
--	--	--

Table A3: Reviews with DistilBERT predictions aligning with ground truth labels

Review	Ground Truth Label	Llama 3.2 Predicted Label	DistilBERT Predicted Label	Reason for Mismatched DistilBERT Predictions
I bought this oil after reading all of the reviews that mentioned just how strong it is. However, it can't be understated just how powerful it is. I have been using it for making cinnamon hard candy. I put a half teaspoon of the cinnamon oil in a blue, Kitchenaid measuring spoon and set it aside for a few moments. When I picked up the spoon again, I noticed that the oil was blue colored. Upon further inspection, I discovered that the oil had actually eaten away at the plastic--it was like I had put nail polish remover (acetone) in the spoon. Do not use this product with plastic measuring spoons! ... This product is good stuff; just be very careful when using it.	Positive	Positive	Negative	DistilBERT focuses on the strong cautionary statement that the oil could leech material from plastic spoon although the review is an overall endorsement of the product that it is 'good stuff'.
Besides feeling like chalk and not very wintergreen minty tasting the mints do last a bit longer than you would think. The little pieces of fiber get on my nerves, too. Note: 'chalk' is a typographical error. It should be 'chalk'.	Negative	Negative	Positive	DistilBERT focuses on the slightly positive statement even though the review is mainly negative.
This is a very good food supplement. It strengthens your nails and hair and connective tissue. I have been using it for over 10 years. However the price for shipping and handling is way too much!!! You can get this in Walgreens for \$7.24 8 oz I would not pay this shipping price in my opinion it is not right. Try Walmart.com. They are premeasured in the 8 oz box 32 packs so you can just pour it in your tea or hot water. This is powered it should be a lot less for not having to package it.	Negative	Positive	Positive	Distilbert focuses on the positive segment on the product but fails to recognize the high shipping charges that ultimately makes the user rate the product lowly with a score of 2.
Approximately 20 years ago, I began to exclusively feed our cats Iams or Science Diet. Those brands were recommended by our Veterinarian after our first cat showed signs of Feline Urinary Syndrome (FUS). I had previously fed the cats the typical brands such as 9 Lives, Purina and Friskies. I continued with IAMS and Science Diet, because the other brands caused FUS, caused other digestive problems and the good reputation of IAMS and Science Diet. Then last year, one of our cats developed bladder disease and died. Not long afterward, the pet food scandal broke. Cat and dog food from nearly every pet food company, including premium foods like IAMS and Science, were tainted	Positive	Positive	Negative	DistilBERT happened to focus on the negative portion of other non-organic brands that the user had tried and has many negative terms such as "problems", "scandal", "recalled" and "betrayed". All 70

by melamine. More than 150 brands of cat and dog food were recalled. IAMS and Science Diet were two of the recalled brands, as was natural sounding brands such as Nutro. I vowed to never again purchase pet food from the brands that had betrayed my trust. I began searching for pet food that was all natural and organic. Newman's Own is both natural and organic. It was also one of the few brands not recalled. Although it was difficult to find a local source, I finally found a few, including Amazon. My cats love Newman's Own canned and dry food. They also love Innova EVO. I love it because I don't have to worry about tainted food poisoning my cats. Pros: 1. Peace of mind 2. Does not cause FUS, as did Purina and other similar brands 3. Much less likely to be tainted by lethal contaminants 4. The cats love it 5. Does not cause the other digestive problems. Cons: 1. It is more expensive than the non-premium brands 2. It can be difficult to find a nearby store that sells Newman's Own. Overall, the pros outweigh the cons. And I can now add Amazon to my list of sources for the food.				people who had responded to the usefulness of the review had found it useful and yet DistilBERT did not identify the correct sentiment. In contrast, all other models identified the sentiment accurately.
---	--	--	--	--

Table A4: Reviews with mismatched DistilBERT predictions

Review	Sentiment	Analysis
It is almost impossible to find chocolate that is low carb, and doesn't contain sucralose or malitol. This candy is completely unique in that the sweetener is erythritol - no carbs, no toxic artificial junk, no sickening malitol. I can eat these without raising my blood sugar or making my tummy complain violently. That is the good part. Here are the negatives: My family likes them, too, and I cannot make a box last long around my house - and these ARE NOT CHEAP! That is why I gave it only 4 stars - I think the people who make these bars are raking in an obscene profit because they can - this product is one of a kind - No Competition and so the price is sky high. I don't think the chocolate is any better quality than comparable chocolate bars for half the price or less, but this company has inflated the cost of ChocoPerfection to indecent levels because they have the market cornered on a specialty product. Not uncommon, but kind of unfair to their customers. All the same, treats like this are so rare in my low carb, diabetic world, that I will buy these anyway. And I will watch the pantry like a hawk to keep my non diabetic kin-folk out of the goodies!	Positive	Although reviewer criticized on the high price, the overall sentiment is positive. This demonstrates the model's ability to handle nuanced reviews well. Note: Text with positive sentiment is highlighted in green and negative sentiment in red.
My work takes me on the road often, and here I am stuck again with a Tassimo in my host's house. I thought I'd contribute by buying some pods. Not unlike the other one-star reviewer on here, against my better judgement I bought the Starbucks Pods BECAUSE I love a bold Italian roast, thinking it would be better than what we get at Starbucks shops. But this is not only over-roasted, this is beyond burnt. Experiencing this allegedly great cup of coffee is akin to drinking crude oil mixed with sea water in the Gulf of Mexico. If you read through all the five-star reviews it is obvious what people like about SB: It's consistency. And consistent it is; consistently burnt. Several local roasters on my trips told me that what SB does is over-roast the coffee beans to an almost burnt stage so that it	Negative	Despite the sarcastic phrases, model can correctly identify overall sentiment as negative. Note: Text with sarcasm is highlighted in blue.

ALWAYS tastes the same, anywhere in the world. And that's what they really sell, not coffee, but predictability... just like McDonald's. And that's usually what the mob goes for; knowing what you're gonna get is far more comfortable, even if you know that it is bad. It is like voting for a bad president for a second term. Thank God we have term limits. I should have remembered my roasting friends' advice and bought anything other than SB.		
Not as good as home made, but better than any other brand that I have tried. It actually tastes like a queso dip and does not have any chemical off flavor. It is spicy but not hot. I would say that this was on the high end of mild. It is best served hot, but stays dippable when cold.	Positive	Even though product is 'Not as good as home made' but model is able to detect overall sentiment as positive. Note: Text with negation is highlighted in grey.
I made Chebe rolls this evening for dinner. They needed to be cooked an extra 10 minutes and they still were gooey and sticky in the middle. They tasted somewhat okay, but the texture and color (clear) of the middle was very unappitizing. I won't buy them again.	Negative	Even though the 'unappitizing' was spelled wrongly, model was able to detect that review is overall negative sentiment. Note: Text with typographical error is highlighted in golden.
These taste soooooo good, kind of like a candy bar without all the calories. Nice in between meals snack, very satisfying, Be careful, they taste so good, it's hard to each just one! I totally recommend this product, AWESOME!	Positive	Model did not misclassify when slang words were used. Note: Text with slang is highlighted in pink.
My son LOVES graham crackers. In fact, they are a staple at my house. These fresh stacks are a great size for throwing in a purse or diaper bag. The smaller packages keep the crackers fresh and are great for busy moms on the go. We've really enjoyed them at my house. My son gives them two thumbs up! I keep a pack on me at all times. :-)	Positive	Model correctly classified text with slang and emoticon.

Table A5: Reviews which Llama 3.2 predict correctly

Review	Ground Truth Label	DistilBERT Predicted Label	Llama 3.2 Predicted Label	Reason for Mismatched Llama 3.2 Predictions
I've just finished my first bottle. It's a 4/5, great value, pure simple extra virgin goodness. Not earthy, a bit on the young side.	Positive	Positive	Negative	Although user praised the product, the Llama model focused on mild criticism.
I was really hoping this would be good. Unfortunately it is not. Overall, the nutritional composition of peanuts is extremely good, with most of the peanut's fat being monounsaturated. They also provide an excellent source of protein, vitamin E, and fiber. So make your own peanut butter, or just buy the natural with no additives.	Negative	Negative	Positive	Although user is dissatisfied with product, the Llama model focused on positive phrases unrelated to the reviewed product.
I lost faith in my long time cat food, Science Diet, after the massive pet food recall. Fromm claims this dry food is made in the US (Wisconsin) from	Positive	Positive	Negative	Although the user favors Fromm's dry cat food, praising its positive

meat, cheese and vegetables bought from local growers. Quality meats, no glutens, fresh vegetables and human quality vitamins. The cats like it. It is more expensive than Science Diet, but not prohibitively so. Note: the Milwaukee newspaper advises the canned Fromm is made in China. I have absolutely no faith in the Chinese quality and safety control standards so I would not buy the canned foods.				qualities, the Llama model focused on negative phrases that were unrelated to the reviewed product (cat food from another brand, Science Diet and canned food).
---	--	--	--	---

Table A6: Reviews with mismatched Llama 3.2 predictions.

Review	Ground Truth Label	Logistic Regression Predicted Label	Llama 3.2 Predicted Label	DistilBERT Predicted Label	Reason for Mismatched DistilBERT Predictions
I love the product, but the stock the Amazon sent me was about 8 months passed the best before date stamped on the box. Make sure that your product is fresh and safe for consumption by verifying the BEST BEFORE DATE!	Negative	Positive	Negative	Negative	Although user has commented that they love the product and used positive terms such as fresh and safe for consumption, both Llama 3.2 and DistilBERT were able to correctly identify that the sentiment is negative.
Oh sweet Magnum Kona, you were my go-to caffeine pusher for 10 months. You taste so smooth, you smell so good, you were the best part of waking up in the way foldger's never could be. Now you slip in this 25% price hike to \$10/lb? No thank you. Why have you deserted me?	Negative	Positive	Negative	Negative	Logistic regression model focused on the positive elements of the product but failed to capture the negative sentiment of the high price.
this is my favorite flavor i use this for rock candy for christmas for my grand children they llove it	Positive	Negative	Positive	Positive	Both of the positive words in this review have been misspelled but DistilBERT and Llama 3.2 models still accurately identify the correct sentiment.
Honestly, at first, I didn't think this would be a very good product. I decided to take a chance on it, and I was NOT disappointed. I thought maybe it would be a complete different product then from the picture, but I was wrong. It looks just like the picture. It was an awesome	Positive	Negative	Positive	Positive	Logistic regression fails to identify the positive sentiment while both Llama and DistilBERT are better at handling the negation.

gift for my mother, and from what I tasted, it was delicious as well. It also came surprisingly fast in the mail! Bravo, you people have impressed me!					
If you want a cereal that has lots of fiber and protein , and will keep you full for hours.... this is not that cereal . If you want a cereal that is just plain delicious , to be eaten for the pure joy of eating it... this is not that cereal . If you want a cereal that tastes like stale kix rolled in the crumbs of some honey roasted peanuts, and is satisfying in neither taste nor quality.... then this is the breakfast cereal for you . I think i've made myself clear.	Negative	Positive	Positive	Negative	Distilbert correctly recognizes the phrase "not that cereal" and negates the positive comments at the start of the review. However, the other models fail to identify and focus on the positive comments.
Bought a bag of the hickory smoked flavor, and this stuff was salty enough that you could throw it on the driveway to melt ice. Great price if that's how you like it, but I've had just about every type of Jerky commonly available along with a few that aren't, and this is by far the saltiest of any I've tried.	Negative	Positive	Negative	Positive	Llama was the only model that could correctly predict the sentiment of this review. The other models are most likely focusing on the great price but do not take into account the user's sarcasm about the saltiness of the jerky.

Table A7: Reviews with Logistic Regression, Llama 3.2 and DistilBERT predictions