Master of Science for Business Analytics

NUS Business School & School of Computing

BT5151: Project Proposal for Group 2

Members:
Alex Chen Chen (E1148800)
Laura Ngoc Ha Do (E1148768)
Lim Ciwen Brendan (E0540070)
Toh Yu Qi Chermaine (E0201966)
Wong Cheuk Wah (E1148763)

1. Recap of the importance of customer sentiment analysis and project objective

1.1. The importance of customer-centric approaches for today's businesses

In today's competitive online market, it is essential for businesses to understand their customer's satisfaction to thrive and allocate their resources in order to retain their customer base. Customer-centric approaches can help enhance brand loyalty by tailoring the company's product innovation in the direction that meets the expectations of their consumers. Due to the growth of online shopping, product ratings and reviews have become a major target for businesses to gain deeper insights into both the positive and negative aspects of their products. Through customer-centric machine learning approaches such as sentiment analysis, businesses are able to instantly analyze the overall sentiment in their products extracted from customer reviews, and allowing them to make better business decisions in the future.

1.2. Objectives of the sentiment analysis project

The main objective of sentiment analysis is the use of natural language processing (NLP) techniques to gain insights from the sentiments of customer reviews, in this case from the fine food products listed on Amazon. Sentiment analysis is not only able to measure the overall sentiment of each customer review, but also extract the words with sentiment association that led to the classification on whether the product was predicted to be positive, neutral, or negative. We implement increasingly complex methods from linear models up to transformers to obtain sentiment from reviews and compare the accuracies of these models. Subsequently, we segment overall sentiment into the various Key Purchasing Criterias (KPCs) to assess each review for different attributes important to a customer, contributing the chance they purchase again. Finally, we evaluate the appropriateness of each model and recommend the suitability of implementation based on business requirements and available resources.

1.3. Overview of the methodology proposed in the preliminary report

To classify the sentiment of reviews, we use and benchmark the following models. A more detailed section on each model will be discussed in Section 2 of the report.

	Model (in increasing complexity)	**Sentiment granularity**
1	Logistic regression	Overall sentiment
2	Naive Bayes	Overall sentiment
3	DistilBERT	Overall sentiment
4	RoBERTa	Overall sentiment
5	Multi-head attention	Overall sentiment
6	SpaCy Aspect Classifier	Aspect based sentiment
7	PyABSA Transformer	Aspect based sentiment

Table 1. Models implemented

2. Data Description, Preprocessing and Data Labeling

2.1. Amazon fine food reviews dataset and its unstructured data for sentiment analysis

We used an Amazon Fine Food Reviews Kaggle dataset, containing about 500,000 food reviews from a span of more than 10 years up to October 2012. Data included details on the products, users, their rating and the natural language summary as well as review. These reviews are individually tagged with a rating between 1 to 5, representing the overall satisfaction of the buyer.

2.2. Data cleaning, preprocessing, and feature engineering, incl. challenges

Due to the large dataset size, only reviews from 2011 were utilized in training the model. Data collected in 2012 were reserved as unseen data for further analysis in the last section. To preprocess the data, rating scores were first converted from a range of 1 to 5 to multiclass labels of +1, 0, and -1. Reviews with ratings of 4 or 5 were assigned a sentiment score of 1 (positive), those with a rating of 3 were assigned a sentiment score of 0 (neutral), and reviews with ratings of 1 or 2 were assigned a sentiment score of -1 (negative). Subsequently, review columns with missing values were replaced with blank space, and duplicated rows with the same UserId and ProductId were removed. The summary and review text were then merged and cleaned by eliminating irrelevant elements such as URLs, mentions, punctuation, and stopwords using the Natural Language Toolkit. TextBlob was employed to further extract features such as review polarity and subjectivity. Finally, additional features such as the length of the text and the ratio of upvotes to downvotes of the review's helpfulness were included. These additional features were also analyzed to assess their impact on model performance. In the last step, all unnecessary columns were dropped.

3. Models deployed for customer sentiment analysis

3.1. Baseline Models - Naive Bayes and Logistic Regression

To effectively analyze the predictions of customer sentiment analysis from our dataset, models of different levels of complexity were tested in order to compare the advantages of each model and benchmark their performances with the rest. For our baseline models we decided to use two simple and efficient models to benchmark our results: Naive Bayes and Logistic Regression. These are fast and intuitive models that are able to give a general baseline performance and give an idea on how subsequent more complex models should perform and which parameters to optimize.

In the dataset train and test split, Synthetic Minority Oversampling Technique (SMOTE) was used to address the imbalance in the dataset to provide the model with more negative reviews and neutral reviews to train on. SMOTE was used to overcome overfitting problems from oversampling or losing valuable information from undersampling. It preserves the underlying characteristics of the minority class and improves the generalization capability of the model. The TfidfVectorizer was subsequently utilized on the *Summary_Text* and *Clean_Text* columns to preprocess them for the model.

3.1.1. Result and performance metrics (Precision, Recall, F1 Score)

The performance of the models were measured through the following metrics:

- Precision: Accuracy of positive predictions.
- Recall: The sensitivity of true positive rates.
- F1 Score: The balanced mean of precision and recall.

The weighted average for these metrics is used due to the presence of a class imbalance among the three different sentiments. The performance of the baseline models (Naive Bayes and Logistic Regression) are the following and are benchmarked against the method of cleaning and inclusion of additional features pertaining to user attributes, as well as the polarity and subjectivity of the reviews.

Model	Metric	1	2	3	4
Stopwords and punctuation		Retained	Retained	Removed	Removed
Additional Features		Included	Excluded	Included	Excluded
Naive Bayes	Precision	0.87	0.87	0.86	0.86
Logistic Regression		0.90	0.90	0.88	0.88

Naive Bayes	Recall	0.79	0.79	0.79	0.79
Logistic Regression		0.87	0.87	0.84	0.84
Naive Bayes	F1 Score	0.82	0.82	0.82	0.82
Logistic Regression		0.88	0.88	0.85	0.85

Table 2. Performance for Logistic regression and Naive Bayes models

3.1.2. Benefits and advantages

The results show that the baseline models perform extremely well in predicting the positive sentiment (1) of the reviews in the Amazon fine food dataset, with a precision score of around 90%. Although both models struggle more with accurately classifying reviews with neutral sentiment, they demonstrate good performance in identifying negative sentiment. Moreover, their computational efficiency, characterized by low training and prediction costs, underscores their suitability for rapid prototyping. This makes them particularly advantageous for businesses seeking to deploy a minimum viable product for proof of concept purposes.

3.1.3. Limitations of the Baseline Models and Selection of more advanced Models

The main limitations of our baseline models are the ability to handle the complexity of large texts and the overall context of the reviews. These models are able to identify the probability of positive and negative sentiments in specific words, resulting in an assigned weight in the classification of the text sentiment, however, they are not able to capture the dependencies between different words, resulting in difficulties capturing the contextual sentiment. The baseline models assume each word is independent with the other words in the prediction of the sentiment, losing a lot of contextual value that could help the predictions. The use of transformer models resolves this issue, as the higher complexity of the models result in the ability to analyze sequential data to improve accuracy and capture the contextual sentiment of the whole text other than through specific keywords.

3.2. BERT Models

BERT, a prominent large language model developed for natural language processing (NLP) tasks, has revolutionized the field of NLP by significantly enhancing performance across various NLP tasks. Consequently, we aim to evaluate the performance of optimized versions of this transformer language model, including DistilBERT, RoBERTa, and DistilRoBERTa from the Hugging Face transformers library. Initially, our proposal suggested the use of Recurrent Neural Networks (RNN). However, leveraging transformer models can address the limitations of RNNs, such as difficulty in capturing long-term dependencies. The initial evaluation involved utilizing DistilBERT as the tokenizer to extract features and train on logistic regression, which performed better compared to Naive Bayes. This was done to assess whether the baseline models would achieve better performance with this feature extraction technique. Subsequently, we also evaluated the various BERT transformer models DistilBERT, RoBERTa, and DistilRoBERT by fine-tuning the last two layers and trained them on our dataset to evaluate the performance of sentiment class prediction.

For all these models, the *Sentiment* column was used as the target and the *Summary_Text*, which is the original review text, was used as the feature. The column with no pre-processing of the review was used to preserve the contextual information that may influence the sentiment of the sentence. We then apply a 80-10-10 split for the training, validation and test splits respectively, training our data with the same 80% of the dataset as our baseline models. The random state is maintained across all models to ensure all the data is trained and tested on the same partition of the dataset, ensuring compatibility of

metric performances across all the models. The Summary_Text column was then tokenized using the tokenizers from the respective models in the Hugging Face transformer Library. A maximum length of 512 was standardised across the 3 models, which is the maximum length available for the distilled models.

3.2.1. Benefits and advantages

The primary advantage of BERT transformer models lies in their capability to analyze word sequences through the attention mechanism. It considers inputs before and after each word, enabling better context comprehension and observation of relationships between words. This aspect is crucial in sentiment analysis, where understanding relationships between different keywords is paramount. This addresses the primary weaknesses of our baseline models, potentially leading to significant enhancements in performance metrics. The attention mechanism assesses the impact of each word on all other words in the sentence, thereby capturing intricate patterns and relationships that offer comprehensive contextual information of the entire text. Additionally, being pre-trained on extensive corpus allows for higher accuracy in sentiment prediction.

DistilBERT was first evaluated as it was optimized from BERT using a compression technique crafted to train a smaller model with the aim of emulating BERT's functionality. Subsequently, RoBERTa and DistilRoBERTa were assessed for comparison. These models were optimized and trained along larger datasets than the BERT model, with 124 million parameters (DistilBERT has 67 million). Therefore, they should have a deeper knowledge of nuances that could also appear on Amazon food reviews, such as slang and abbreviations. Given their enhanced complexity and focus on sentiment analysis, RoBERTa-based models are expected to outperform BERT and DistilBERT.

3.2.2. Challenges in training and validation

The primary challenge during training and validation lies in the substantial computational demands of transformer models, even when leveraging faster and smaller models like DistilBERT and DistilRoBERTa. To mitigate this, extra computational resources were acquired on Google Colab to expedite training and prevent runtime errors. Without this optimization, training a single model could exceed 6 hours. However, the extensive training times pose obstacles to effective hyperparameter optimization, crucial for mitigating overfitting in these complex transformer models, particularly given the dataset's size and the length review texts.

3.2.3. Result and performance metrics (Precision, Recall, F1 Score)

Similar to the baseline models, the weighted average of Precision, Recall and F1 Score were used to evaluate the performance. Logistic Regression was omitted from Table 3 as there was no finetuning of the hyperparameters.

	DistilBERT	**RoBERTa**	**DistilRoBERTa**
Precision	0.86	0.88	0.88
Recall	0.87	0.89	0.89
F1 Score	0.87	0.88	0.88

Table 3. Performance for various models on Validation Data

	DistilBERT as Tokenizer (Logistic Regression)	**DistilBERT**	**RoBERTa**	**DistilRoBERTa**
Precision	0.85	0.88	0.89	0.86
Recall	0.87	0.86	0.89	0.88
F1 Score	0.85	0.87	0.87	0.88

Table 4. Performance for various models on Test Data

All the validation and test results were relatively similar, with the RoBERTa model performing slightly better compared to the rest. This also suggests there is not an overfitting problem despite the large complexity of the model. The scores ranging from 85-89% across all metrics and data partitions also suggest various BERT models adjust efficiently to the imbalance of the dataset, where negative sentiment reviews (-1) are a lot less prominent than positive sentiment reviews (1).

3.3. Multi-head Self-Attention Model

Moving on from BERT models, we also evaluated a multi-head self-attention model in Keras that uses only the encoder block of the original transformers model (encoder-decoder) designed for sequence problems. It is also a transformer model that is able to understand the relationship between sequential elements that are far from each other. It expands the model's ability to focus on different positions of the sentence or text simultaneously depending on the number of heads set in the model, creating that number of embeddings. The model then groups all the information learnt from each head and forms a classification of the overall text sentiment just like previous models.

3.3.1. Benefits and advantages of Multi-head Self-Attention Model

The main benefit of the multi-head attention model over the other transformer models is the ability to gain a deeper insight into the text as a whole, analyzing the reviews holistically. It is also able to analyze different nuances of the same word simultaneously, providing deeper context on the possible sentiments of the text.

3.3.2. Training process, incl. data partitioning

The multi-head model is trained with the same 80-10-10 train, validation, test splits to maintain the consistency across all models. The random state is also kept the same to ensure the transformer models are trained with the same dataset partitions. The model is set with the parameter of 2 heads, 32 neurons, 100 embed dimensions and a max length of 562. The max length was chosen based on the 75th percentile of the length of the *Summary_Text* column.

3.3.3. Challenges in training and validation

Due to the architecture of a multi-head attention model, the complexity and training times of the model increase quadratically as the text size of the Amazon reviews increase. Similar to the previous transformer models, computation power is one of the main challenges as figuring out the optimal hyperparameters and epochs to use can be complicated due to the limitation of GPU resources. A large amount of heads in the multi-head attention model also runs the risk of overfitting, which is why this number was simply set to 2 in order to check its effect compared to the other transformer models.

3.3.4. Result and performance metrics (Precision, Recall, F1 Score)

The precision, recall and F1 score of the model are the following:

	Test Data
Precision	0.90
Recall	0.91
F1 Score	0.91

Table 5. Performance for the Multi-head Attention model

The results show a clear improvement over all other transformer models, yielding the best precision, recall and F1 score. It demonstrates the complexity of the dataset, as the ability for the multi-head attention model to capture multiple relationships simultaneously gives it an advantage over its counterparts. Sentiment analysis is heavily weighted around understanding the context of the text as whole, and in large texts of data such as Amazon reviews, a multi-head self-attention model excels in this department.The ability to analyze data from different perspectives allows the model to also assign different weights depending on the relevance of the relationship and is able to filter out noisy data, making it one of the superior models when it comes to sentiment analysis.

3.3.5. SHAP Feature Importance

SHapley Additive exPlanations (SHAP) is tested on the multi-head attention model due to its best classification performance in sentiment analysis to extract the tokens that push the model towards any of the three classifications (positive, neutral, negative) by assigning a SHAP absolute value.

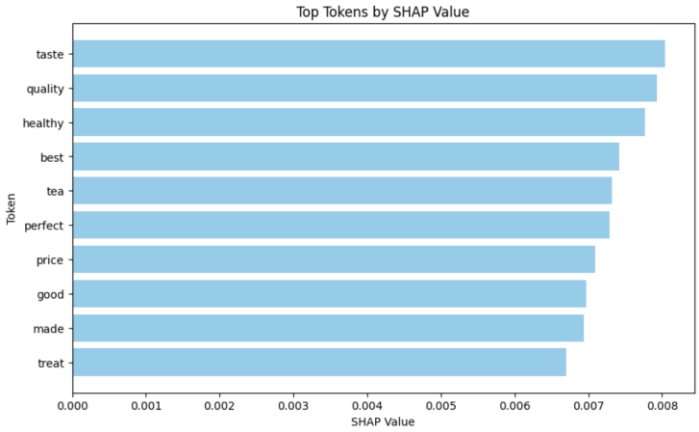


Fig.1 SHAP Importance Plot

The table above shows the top tokens that had the greatest contribution in predicting the sentiment of the Amazon dataset reviews. It is worth noting that the top words such as ‘taste’ and ‘healthy’ may lack the contextual relationships to fully understand their impact in the sentiment analysis performed by the model as transformer models’ complexity can not be fully explained through simple token importance. However, the SHAP analysis could also become an important indicator in identifying valuable KPCs such as ‘taste’, ‘quality’, ‘healthy’ and ‘price’. These words being identified with high SHAP values show the importance of KPCs in customers’ sentiment towards online products, and could provide key insights on how Amazon or other online shopping markets could include KPCs in their review criteria to allow sellers to have greater insights into the specific reasonings behind the positive and negative sentiments towards their products.

3.4. GenAI Zero code Comparison

To compare our results against what can be done by current open source methods using only natural language, we purchased a subscription for Open AI’s Chat GPT4 in order to examine its performance. We then prompted the model to try to label the review sentiments by itself, and then compared it to the ground truth in the dataset. This gives us a comparison of the value add from our project as compared to a quick prompting through Open AI’s Chat GPT4. To conduct this, only the natural language string reviews are given to GPT4, without the ground truths.

3.4.1. Benefits and advantages of GenAI Zero code

A definite benefit of this technique is that no data science or machine learning background is required for this to be done as pure natural language prompting can be used to obtain the sentiment labels.

3.4.2. Prompt refinement for customer sentiment analysis

A csv file containing a column of reviews was given and the prompt used was: *Pretend you are a data scientist manually labeling data points for sentiment analysis based on text reviews. For each review in the attached file, determine if the sentiment of the review is positive, negative or neutral. Return the output file with the sentiment column filled in.*

3.4.3. Result and performance metrics (Precision, Recall, F1 Score)

From the sentiments generated by GPT4, we compared it back with the ground truth labels. From the results obtained, we are immediately able to notice that the performance of this technique does not compare to that of logistic regression.

	Test Data
Precision	0.63
Recall	0.84
F1 Score	0.72

Table 6. Performance for GPT4's sentiment analysis

4. Predictions and Further Analysis on 2012 Unseen Data

The Multi-head Self-Attention Model was identified as the top-performing model among all candidates. Subsequently, we leveraged this model to evaluate a fresh dataset. Specifically, we extracted reviews for the top two most popular products from the 2012 data: Quaker Oatmeal Cookie and Energy Drink.

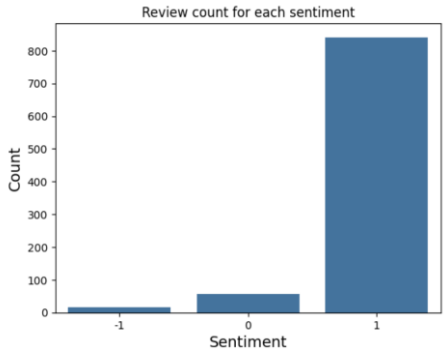


Fig. 2 Class Distribution for Oatmeal Cookies

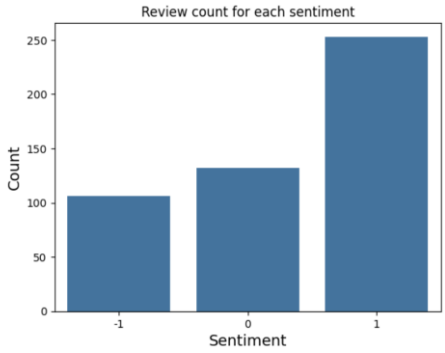


Fig. 3 Class Distribution for Energy Drinks

Our model gave the following performance:

	Oatmeal Cookies	**Energy Drinks**
Precision	0.92	0.67
Recall	0.93	0.66
F1 Score	0.92	0.66

Table 7: Performance for the 2 Datasets

It was observed that the model's performance decreased notably in scenarios where there was a higher prevalence of negative (-1) and neutral (0) classes. This decline could stem from the limited representation of these two classes within the training dataset, resulting in the model's reduced effectiveness in identifying them.

4.1. Aspect Based Sentiment Analysis (ABSA)

In addition to predicting the sentiment classes, aspect based sentiment analysis, or fine-grained opinion mining, was used to provide further analysis on these 2 most popular products. It is a technique used to identify sentiments towards specific topics within a string of natural language. This is useful when text contains multiple topics, and the sentiment towards each individual topic has to be identified. An example of this is “The food was good but the price was so high and was not worth it”. ABSA aims to identify a positive sentiment toward the aspect food but negative sentiment toward the aspect price. For this project, the aspects to be identified will be relevant purchasing factors that customers take into account when deciding to continue patronizing a particular store or re-purchasing a product.

4.2. SpaCy Aspect Classifier

The SpaCy package was utilized to extract aspect terms from the reviews, followed by the application of a classifier to determine the aspect class, corresponding to our predefined Key Performance Criteria (KPC) for each review. This methodology offers valuable insights into the specific KPCs mentioned within the reviews, particularly across different sentiment classes. For example, following the sentiment prediction using our multi-head attention model, we leverage this information to identify areas for improvement within the negative sentiment class for the top 2 products purchased in 2012. One advantage of employing this method for aspect identification is the ability to predefine the aspect class based on the relevant KPCs for a particular product. However, a limitation is the inability to predict the sentiment for each KPC, particularly when multiple KPCs are identified within the same review. When this happens, we are able to identify that the review does in fact reference multiple KPCs but would not be able to accurately attribute the contributions of each KPC to an overall sentiment predicted.

aspect	-1	0	1
delivery	4	5	148
health	1	1	29
packaging	8	13	274
portion	7	14	150
price	1	3	69
quality	10	10	193
taste	18	18	589
texture	11	20	335

Fig. 4 Number 1 Product - Oatmeal Cookie

aspect	-1	0	1
delivery	3	16	23
health	13	13	24
packaging	37	98	147
price	1	6	14
quality	41	83	128
taste	57	123	182

Fig. 5 Number 2 Product - Energy Drink

4.2.1. PyABSA

To offer more granular insights into the sentiment related to specific aspects of a product for businesses, PyABSA, as a modularized framework built on PyTorch for reproducible ABSA, was employed to extract aspect terms from reviews and classify the sentiment associated with each aspect. Since it incorporates a range of models including attention-based, graph-based, and BERT-based models, along with a diverse collection of datasets such as Yelp, MMAs, Restaurant14, and MOOCs data, PyABSA offers a robust solution for handling various ABSA tasks effectively (Yang et al., 2023). This enables us to apply our data for ABSA effectively, without concerns about data scarcity, particularly regarding absent ground-truth labels for aspects in our dataset.

Taking the Top 2 Product as an example, PyABSA was able to automatically tokenize the input sentence and identify the relevant aspect terms according to these tokens extracted. Disregarding the overall sentiment predicted previously (Predictions column), sentiment predictions were made correspondingly for each aspect with probabilities and confidence levels computed for developers' evaluation of ABSA performance due to the nature of unsupervised learning. For businesses using this model, it would not be possible to specify the KPC for the model to identify, and as such KPCs may not be identified in order to classify sentiment.

	sentence	Predictions	tokens	aspect	sentiment	probs	confidence
86	not a big fan . . . flavor for kids , but too ...	0	['not', 'a', 'big', 'fan', ' ', ' ', ' ', ' ', ' ', ' ', 'fla...	[mix, water, mix, caffeine]	['Negative', 'Negative', 'Negative', 'Negative']	[[0.9807977080345154, 0.011564690619707108, 0....	[0.9808, 0.9938, 0.9808, 0.9945]
428	too confusing " a squeeze " ? how do i know ho...	-1	['too', 'confusing', '"', 'a', 'squeeze', '",...	[bottle, cup, taste, flavor]	['Negative', 'Negative', 'Negative', 'Negative']	[[0.9970003962516785, 0.002317150356248021, 0....	[0.997, 0.9953, 0.9925, 0.7394]
51	surprisingly wonderful drink additive . . . as...	1	['surprisingly', 'wonderful', 'drink', 'additi...	[drink]	['Positive']	[[0.00529313273278938, 0.0003364187723491341,...	[0.9944]
120	average taste and pep vitamin squeeze isn ' t ...	0	['average', 'taste', 'and', 'pep', 'vitamin', ...	[water, caffeine]	['Negative', 'Positive']	[[0.9904863238334656, 0.008949346840381622, 0....	[0.9905, 0.9698]
404	waste of time this didn ' t taste of fruit pun...	-1	['waste', 'of', 'time', 'this', 'didn', "'", '...	[fruit punch, taste]	['Negative', 'Negative']	[[0.9991748929023743, 0.0006168010295368731, 0...	[0.9992, 0.9988]

Fig. 6 Screenshot of the output

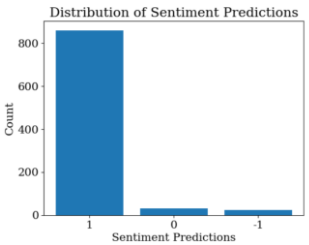
5. Discussion of customer sentiment insights

5.1. General insights obtained from the sentiment analysis using different models

Through the deployment of our models, ranging from the baseline models Naive Bayes and logistic regression to the more complex transformer models DistilBERT, RoBERTa, and multi-head attention we were able to analyze and compare customer sentiment insights depending on the strengths of each model. The use of transformer models and the performance metrics associated with them gave a deep insight into the importance of context when predicting the sentiment of a customer review. A word that may be associated with a positive or negative sentiment in isolation may not necessarily have that same sentiment when evaluating the context it was used in, which was supported by the clear performance in transformer models relative to the baseline models. Although predicting customer sentiment can have a great value to the company, the ability to analyze it further into KPCs could further help companies improve by being able to align their brand and products in the future with customers' expectations.

To demonstrate the insight generation process, we sampled the top 2 most popular products in our dataset, namely product IDs B007JFMH8M and B006MONQMC, as examples. This can assist businesses in finding room for improvement from the negative aspects in the positive reviews and prioritizing efforts to improve customer satisfaction by pinpointing the lacking areas highlighted in the negative reviews.

5.1.2 Insights for Product B007JFMH8M - Improvement suggestions from the negative aspects in the positive reviews



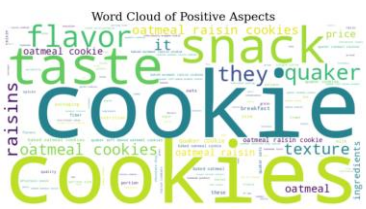


Fig. 7 Distribution of Predicted Class *Fig. 8 Word Cloud of Positive Aspects*

Given that the majority of reviews are predicted as positive by the overall sentiment model, this example demonstrates how businesses can enhance recommendations by identifying negative aspects within the

predominantly positive reviews. In addition, The word cloud illustrates the aspects highlighted in the collective reviews, with font size representing the frequency of positive mentions. From this, it's apparent that customers highly praised these cookies for their flavor, taste, and texture, indicating significant appreciation for these qualities.

Among reviews predicted as positive, aspects that had received more negative comments than positive and neutral were extracted for analysis. Some users highlight negative aspects like sugar content, nutritional panel, fat, and carbs, raising potential concerns about the product's nutritional value and healthiness. This indicates a need for companies to consider incorporating healthier ingredients or providing clearer nutrition labeling to address these concerns, especially for customers who prioritize health-conscious choices.

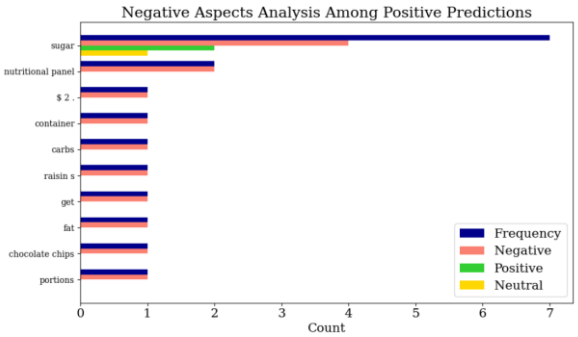
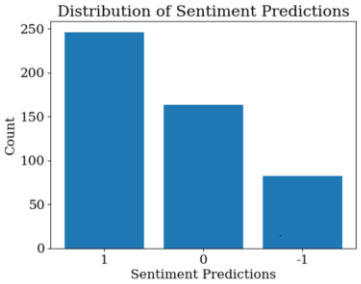


Fig. 9 Negative Aspect Analysis

5.1.3 Insights for Product B006MONQMC - Addressing criticisms from negative reviews



The distribution of sentiment predictions reveals that class 1 is the most prevalent, followed by class 0 and -1. While class 1 remains predominant, it's evident that classes 0 and -1 also hold significance and cannot be overlooked. This distribution underscores the diverse range of sentiments expressed in the dataset. The focus of the analysis would be on how companies can pinpoint areas of improvement to meet customer expectations.

Fig. 10 Distribution of Predicted Sentiments

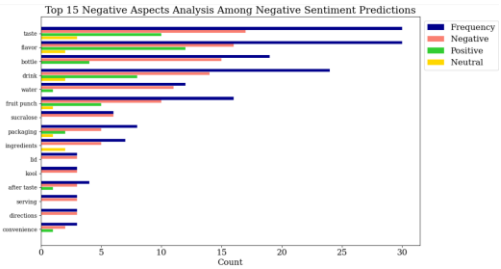


Fig. 11 Negative Aspect Analysis

Fig. 12 Word Cloud

Given the extensive list of over 70 aspects that garnered notable criticisms, the bar chart exclusively highlights the top 15 areas in need of improvement, consistently drawing more negative comments than

positive ones. However, all negative aspects are comprehensively depicted in the word cloud visualization.
Several key purchasing factors were highlighted in the bar chart, such as taste, favor, packaging, and ingredients, implying a significant discrepancy between customer expectations and the current product experience. To bridge this expectation gap, it is imperative for companies to prioritize enhancements in these areas. Strategies may include revamping the taste and flavor profiles, redesigning packaging for improved visual appeal, and providing clearer usage instructions. Additionally, focusing on convenience factors such as serving size and ease of use can enhance overall customer satisfaction. By actively addressing these pinpointed areas, companies can not only mitigate negative feedback but also foster deeper customer engagement and loyalty towards the brand.

5.2. Comparative analysis of the models' performances and selection of best model

Using the performance metrics precision, recall, and F1 Score can provide a general analysis of each model's performance and justify the selection of multi-head attention model as the best model used to predict sentiment analysis in the Amazon fine food dataset. However, it is important to value a key feature outside of the performance metrics when evaluating a model's performance, the model's training and computational demand. A company deciding on a model to use for estimating sentiment analysis may want to also consider how exhaustive each model is relative to its predictive performance when assessing the optimal classification model for their company. A product with not many KPCs that competes in the market mostly in one category (e.g. price) could result in sentiment analysis that could easily be processed with just a baseline model such as Naive Bayes and logistic regression that could simply analyze isolation words (e.g. cheap and expensive) and obtain high performance metrics with very little investment in data analysis and computational power. However, a more complex dataset with wider array of goods such as our Amazon fine food dataset would require a more complex model to understand the nuances and contextual relationships of words and sentences in the review texts, resulting in the multi-head attention model being chosen as the optimal model in sentiment analysis given our chosen dataset. The use of our different range of models can give insights to companies to understand the importance of analyzing their datasets and understand the strengths and weaknesses of each of the models to properly align their data collection with the model that is most suitable for them given their resources and budget.

5.3. Challenges, limitations and gaps in each model

Next we will assess the challenges and limitations of the different models tested when classifying sentiment analysis:
Baseline Models: The simplicity of these models resulted in the lowest performance metrics (precision, recall, F1 Score) due to the models being unable to capture the context or relationships between different words in the model. It simply trained the data by associating weights to each word depending on the ground truth sentiment of the training data using probability, and treating each word as an independent feature with no impact on the rest of the text.
DistilBERT, RoBERTa, DistilRoBERTa: Similar performance results due to being variants of the Google BERT pretrained model, both having extensive training runtimes due to the large number of parameters (67 and 124 million). Another limitation is the inability to run hyperparameter fine-tuning due to the GPU constraints.
Multi-head Self-Attention Model: Although offering the highest performance metrics with relatively faster runtimes than the other transformer models, the intensive computational demand resulted in the inability to gridsearch the model with too many different parameters, potentially missing out on some optimal hyperparameters in the model.
SpaCy: SpaCy excels because it enables the user to define the KPCs relevant to the usecase, ensuring that each KPC can be accounted for. SpaCy's main limitation however is the inability to isolate sentiment across KPCs for reviews that contains multiple topics. For reviews that only contain one KPC, we would be able to take the overall sentiment and fully attribute it to that KPC.

PyABSA: The main limitation from PyABSA is that this pretrained model does not enable the business user to define the KPCs. The model performs its own extraction, using its internal architecture of multiple transformer models. While this is good as it mitigates the need for training, it is not guaranteed to generate a collectively exhaustive list of KPCs relevant to the user.

6. Future directions

To enhance the business insights retrieved from our models, an improvement could be obtaining a dataset that contains ground truth for each of the KPCs that are product or industry specific to the business seller. On a business level, this either requires manual labeling or alternatively, to request for customers to label each of their reviews, for example grouping their review into food quality. This would allow our complex models to be trained on additional categories rather than an overall sentiment, which would unlock the full potential of these highly intensive models in classification of sentiment analysis based on different categories such as price, quality or taste.

The ability of the transformer models to understand contextual relationships and process sequential data could result in a multi label classification model that can predict how many KCPs are associated to a review as well as the sentiment of each of those KCPs while maintaining the unrelated ones neutral. The increased complexity of this model could then potentially require more hyperparameter tuning to improve the sentiment analysis transformer models to optimally analyze the contextual relationships of the texts and deliver better performance metrics.

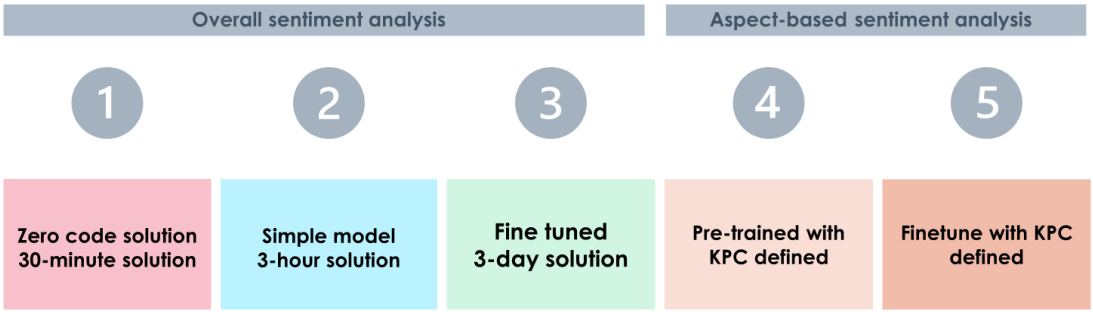
A company interested in these insights and analysis could be convinced to provide greater funding allocation in order to satisfy the computational requirements to deploy aspect based modeling in predicting KPC sentiments. The higher costs would be matched with higher meaning insights, giving companies enriching insights into which specific categories should be improved to align their products with the expectations and priorities of the consumers.

7. Conclusion

In conclusion, in this report we detail the methodologies across a range of methods with varying complexities, in order to determine suitability of various methods based on the business requirements. It is not recommended that the zero-code method be implemented (F1 = 0.72), as while it is a quick and easy method to obtain sentiment scores, there are better methods with low complexity to benefit from, as long as the business has data science capabilities. Assuming a business only requires overall sentiment, logistic regression performs unexpectedly well using a tokenizer (F1 = 0.88), even without the requirement of preprocessing the natural language or synthesizing other language features such as polarity. For businesses that can accommodate higher complexity, a multihead attention method (F1 = 0.90) is recommended in order to allow the model to properly represent contextual features. Lastly, if finer grain opinion mining is required, particularly in presenting sentiment on an aspect level, SpaCy can be used, however noting that there might be overlaps in KPCs. Finally, PyABSA is used as an unsupervised means to extract out aspects relevant to the reviews, and perform sentiment analysis on it. The drawback here however is that the KPCs cannot be defined by the business and are done automatically by the model. For a business to fine tune a ABSA transformer model based on individual products or categories, a large amount of resource and labeled data is required. Having all this in mind, we believe, based on the business requirement and amount of resource provided, a suitable scope of data science project can be made based on selecting an appropriate model.

8. Appendices

8.1. Recommendations for implementation



8.2. References

Li, J., Cui, Q., & Chen, Z. (2021). Exploring Transformers in Emotion Recognition: a comparison of BERT, DistillBERT, RoBERTa, XLNet and ELECTRA. Retrieved from https://arxiv.org/pdf/2104.02041.pdf

Yang, H., Zhang, C., & Li, K. (2023). PyABSA: A modularized framework for reproducible aspect-based sentiment analysis. Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. https://doi.org/10.1145/3583780.3614752