Karlsruher Institut für Technologie

Problemset 3 - Marketing Analytics

Institute of Information Systems and Marketing (IISM)

Julius Korch, Marco Schneider, Stefan Stumpf, Zhaotai Liu

Last compiled on December 14, 2023

Contents

Task 1: Important products features

Text Pre-processing

Before we analyse the reviews, we perform the preprocessing steps: stopwords removal, simplification, stemming and tokenization. In the following we explain the reasoning behind our preprocessing decisions.

Stopwords removal

Our first step of the preprocessing is to remove the stopwords. Stopwords are words that are very common in a language and do not add value to the analysis. Because in some contexts some stopwords may be relevant, we looked up the stopwords list and carefully reasessed which words to remove and which words to keep. Especially negative words and negotations might, compeltly change the meanonf of a sentence. For example "I love that these aren't carbonated I don't drink coffee or soda." Removing the negative word would result in a sentence with the opposite meaning. We also kept some connection words like "while" (e.g. While this is my favorite Monster Drink and I do love them. The price for 24 pack is now insane).

Furthermore we also added some words for removal that refer to the names of the brand. For example "red", "bull", "monster", "celsius". Since the brand information is alreadsy stored in another column, including them in the review text could reduce the interpretability of the lasso regression results.

Simplification

The next step we applied is to simplify the text data. We first convert the text to lowercase, remove all punctuation and numbers. Interestingly we found out in our data exploration, that most of the time a number was present in the review text, it was a number referring to the star rating the reviewer has given. The reviewers opinion regarding the price is mostly expressed in words (e.g. expensive). Thus we decided after our conduction of the analysis, that we delete all numbers from the review text. Additionally, we removes whitespaces, special punctuations like bullets (•) and different special characters like hearts to remove the noise in the data.

Stemming

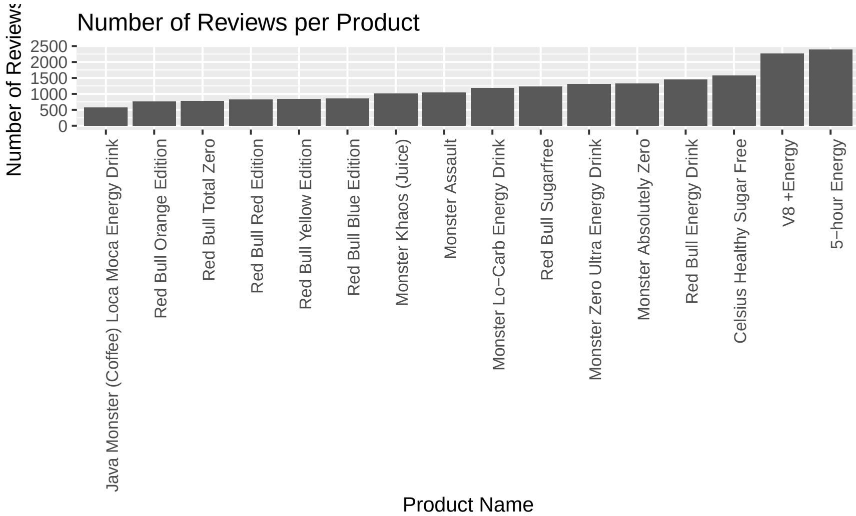
Stemming reduces the number of vocabulary variants by removing word suffixes, such as removing all "ing" (from "packaging" to "packag") and "ed" ("ordered" to "order") forms, thus reducing the dimension of the vocabulary. Fortunately, some suffixes with negative connotations, such as less ("useless" stays "useless"), were retained. In this large text sample, stemming can speed up processing and make the algorithm run more efficiently.

Tokenization

Tokenization is the process of splitting a string into a list of substrings called tokens. It is essential in our case because it helps to prepare the raw text data (reviews) for further analysis, such as lasso regression and topic modelling. Here, we split the reviews into unigrams. Unigrams are single words that can reveal the importance of certain keywords, such as specific ingredients or flavours. Bigrams are two words that often occur together in the document and can capture the meaning of certain contexts, such as 'good quality' or 'good value', which provide a better understanding of whether terms are positively or negatively associated. However, we tried our lasso regressions with both, unigrams and bigrams to compare the interpret ability of the results. We saw that bigrams do not make the interpretation of the output easier. In both cases we would need to look at the reviews themselves to place the uni/bigrams into context. Additionally we checked the wordclouds for both types which confirmed our assumption that bigrams do not provide much more information. Therefore, we decided to use unigrams because using them is more computational efficient and it is easier to look them up in the reviews (and thus put them in context and interpret them).

1.2 Choose appropriate products

Our main criteria for selecting the products to analyse are number of reviews for the products. To perform the analysis we want products with a high number of reviews. If we choose products with products with few reviews, we may not have enough data to do the analysis. Since the number of reviews for products is not in the same order of magnitude as the number of the same order of magnitude as the number of reviews for other products, comparisons between products lack statistical significance. Therefore, we first excluded products products with less than 500 reviews.

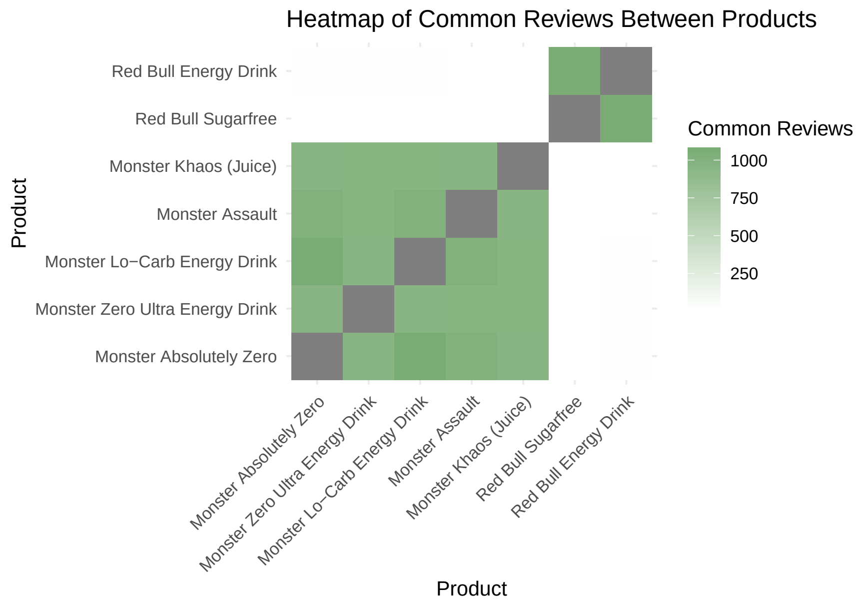


It would also be interesting for us to see the proportion of verified reviews. A review is verified if the customer bought the product directly from Amazon. Therefore, it would be preferable to perform the analysis only on products with products with only verified reviews, as this would guarantee a higher quality of the quality of the reviews, since people have actually bought the product and subsequently. In the table below, we can see that, with the exception of "V8 +Energy", the proportion of verified reviews is around 80%. If we excluded 20% or more of all reviews from the analysis, we felt that our dataset would be too small for a proper analysis. Therefore, we leave the unverified reviews in the dataset.

Explanation for Choosing the 5 Products with most Reviews:

Within the 10 most reviewed products, there are 5 different kinds of Monster and 2 different Red Bull energy drinks. Since on Amazon one can write one review for example for a bundle of different Monster energy drinks which occurs for all the energy drinks of the bundle, we decided to analyse how many reviews are the same in those 7 products most reviewed products. If a lot of reviews are the same within those products, we should not take them all for the analysis.

By looking at the heatmap above, we can see that there are a lot of common reviews between the Monster energy drinks as well as between the Red Bull energy drinks. Therefore, we decided to only take one of each for the analysis. We chose to take the "Monster Absolutely Zero" and the "Red Bull Energy Drink" since they are the most reviewed products of their brand. In conclusion, we chose to analyse the 5 products with the in total most reviews, namely "Monster Absolutely Zero", "Red Bull Energy Drink", "5-hour Energy", "Celsius Healthy Sugar Free" and "V8 +Energy".



This leaves us with a dataset with the following amount of observations:

```
## Number of observations: 8999
```

Lasso Regression for the 5 Products

Previous Information about the Lasso Regression:

1. In general, we want to find out what aspects of the reviews have a relatively large effect on the Amazon star ratings. This is challenging since there is a very high dimensionality of the word count data. For finding influential aspects, we apply the lasso regression because through conducting it, some coefficients of terms are exactly zero and thus, can be excluded from the analysis. Here, only a few variables/terms have an effect and small effects on the ratings are constrained to 0. In our analysis we chose the lasso regression which identifies the best model on regularization path with information criterion (AICc) and not with cross-validation. This simply has the reason that we compared the results of both versions and the results were very similar. Thus, we chose the AICc-regularization path because it has a higher computational efficiency. With this method, the optimal lambda for the model is identified based on the AICc.
2. We considered the possibility of an adjustment of word counts for document length but decided to not implement it to our analysis. The reason for this is that the document size is not varying too much. 75 % of all reviews are smaller or equal to 338 characters. Even though the standard deviation of the review length is 335 characters, we do not adjust the word counts for document length because we have heavy outliers (some have over 4000 characters) which disproportionately contribute to the variation.
3. In the text, we do not always want to refer to the coefficients, this is why we here want to once draw attention to the fact that not all presented terms have the same influence on the Amazon star ratings. Terms with higher coefficients drive the ratings more than terms with lower coefficients. One can see the exact coefficients in the printed outputs and can also see which terms are more or less important.
4. Not all the words outputted by the lasso regression can be used for the task. We have therefore selected the most appropriate words for the task, placed their meaning in the context of the dataset and interpreted them. Exemplary, by analyzing the output of the lasso regression for "Red Bull Energy Drink" we concluded that the term "anyon" which stands for anyone is not helpful in this case. This has the reason that by looking through the reviews we recognized that anyone is used in a lot different meanings. In "Would recommend to anyone!" the term is used to show that the customer would always recommend the product, but in "Why would anyone in their right mind pay $10 more?" the term is used to signalize that the price is too high. Since the meanings of the sentences in which this term occurs differs a lot, we decided to do not include such

terms. A second example would be “carri”. In reviews this term is used to signalize that when buying online, people do not have to carry it home from the supermarket or that a store does not carry it (is not in the sortiment) or that though the packaging of the cans they are easy to carry. We cannot conclude one clear meaning for this term and thus excluded it from our further analysis for “Red Bull Energy Drink”.

First we ran the lasso regression for **“Red Bull Energy Drink”**:

The Amazon star rating here was positively driven by the compared to other drinks little chemical taste and chemical composition (“chem”). Interesting is that often mothers are highly satisfied customers of this energy drink (“mom”). Customers appreciated the secure and good packaging (“secur” & “wrap”). People who were satisfied with the product said it was packed in a way that the cans were secure and properly wrapped by additional material. Furthermore it is interesting that satisfied customers mentioned that they bought it in a count, e.g. as a 4-pack or 6-pack (“count”). This is a good indicator that satisfied customers buy buy more than a single can to have several cans at home. Interestingly, satisfied customers mentioned that it is a good mixer for alcohol (“mixer”). That makes us think that this product is maybe used a lot at parties. Positive ratings sometimes compared this product from Amazon with the product from the supermarket (“retail”). Since the ratings were positively driven by this term, we expect that buying Red Bull on Amazon has some advantages for the customers, e.g. more convenience (no carrying of heavy cans) or a cheaper price/bigger discounts. Also, a lot of customers seemed to like the taste of this energy drink (“nectar”). They compared it to the taste of nectar which usually stands for a sweet and tasty liquid.

Reviews which rated the Red Bull Energy Drink low were driven by the fact that people called it a waste of money (“wast”). Fitting to this, reviewers mentioned that the price of the product is too high (“expens”) and that similar products (other energy drinks) are rated higher by those people (“similar”). We conclude this from the fact that the coefficient for this term is negative and thus, when comparing Red Bull to other drinks, it influences the rating negatively. Interesting is that also negative reviews were driven by the taste. People called the taste horrible, terrible, flat and nasty (“horrible”, “terrible”, “flat” & “nasti”). Negative reviews also contained that the cans or package arrived damaged, empty and sticky because cans were bust open (“sticki”, empti”, “bust” & “punctur”). As last point, people mentioned that the consume of this product is unhealthy (“healthi”) and the contained aspartame (artificial sweetener) is bad for peoples’ health (“aspartam”).

Lasso regression for **Monster Absolutely Zero**:

Here, the Amazon star rating is positively driven the fact that customers report that they were generally impressed (“impress”) and especially about the delicious taste (“delici”, “decent”) which is not too sweet compared to other energy drinks (“sweet”). By comparing the product on Amazon to the product in supermarkets (“retail”) we conclude that people prefer to buy it online than offline. Although, we do not know the specific reason for this, it might be because the package of cans arrived undamaged (“undamag”, “none”) and the products were in good shape after deliver or that it is hard to find in normal supermarkets (“hard”), but could also be due to convenience or a cheaper price. Interestingly, people who attend to classes seemed to like the product a lot (“class”). This counts for both teachers and pupils/students. This fits to the fact that people report to consume it on a daily basis (“daili”). Positive reviews also say that this drink energizes a lot after consuming it (“kick”) and customers were positively influenced by the good can sizes Monster offers (“ounc”).

The star rating of “Monster Absolutely Zero” is mainly negatively driven by people warning potential customers of this product (“beware”). Customers report that it is wasting time trying the drink and also a waste of money (“wast”). This matches the complaints about the price increase (“increas”) and the call for lower prices (“cheap”). The rating is also negatively influenced by the packaging. Customers report that the cans arrived wet, sticky, broken, exploded, shaken (“wet”, “sticki”, “broken”, “explod”, “shaken”) and poorly packaged/shipped (“poor”). Furthermore the flat taste of the drink motivated users to rate the product low (“flat”). Some customers even summarized their experience with this product as horrible (“horribl”).

Lasso regression for **5-hour Energy**:

The star rating of “5-hour Energy” is positively driven by people who consume it when they are on holidays (“holiday”). Also the effect of this drink seems to satisfy the customers. It is reported that consumers do not experience spikes (“spike”) which means that there are no extraordinary high peaks and following crashes in the energy level. Customers report that this is a life saving energy supply for them (“lifesav”). Furthermore, it seems to have a good effect on the metabolism (“metabol”) and people report that it works well for diabetics (“diabet”), since it contains no sugar (“sugar”). Positive influence on the star rating also has the fact that people like to consume it ice cold (“ice”) as a snack (“snack”). The size of the bottle is also a positive factor (“pocket”). Furthermore, when comparing to previous experiences, reviewers are glad that they discovered this product (“glad”) and switched to “5-hour Energy” (“switch”).

The star rating of “5-hour Energy” is negatively driven by people reporting that they dislike the taste (“yuck”, “terribl” & “gross”). The rating is also driven by the reported after effects of this product (“terribl”). Customers mentioned that they had different kinds of attacks, e.g. heart or panic attack, (“attack”), felt dizzy (“dizzi”, had unpleasant skin feelings (“skin”) and needed to go to the hospital because of the effects of “5-hour Energy” (“hospit”). Negative ratings are also driven by the fact that people report that the cans leaked into the box (“leak”) and that the product is a waste of money and time (“wast”).

Lasso regression for **Celsius Healthy Sugar Free**:

The star rating of “Celsius Healthy Sugar Free” is positively driven by people reporting that this drink did not make them feel ill (“ill”). Also comparing the drink to similar Rockstar energy drinks has a positive influence on the Amazon star rating (“rock”). We conclude that customers of this drink like it more than Rockstar drinks. Customers also like that the product is low carb (“carb”) and has only ten calories per serving (“ten”). This matches the fact that the fact that the drink contains niacian which is a vitamin that helps to convert food into energy (“niacin”) and that it has a high amount of minerals (“miner”). Furthermore, customers appreciate the power this drink gives them (“power”) and that it lights up their mood (“mood”). “Celsius Healthy Sugar Free” also seems to be a good drink for training (“train”).

The star rating of “Celsius Healthy Sugar Free” is negatively driven by people reporting that this is the worst energy drink they tried (“worst”). The rating is also negatively driven by the fact that people report that the drink causes breathlessness (“breath”), makes them feel nauseous (“nauseous”) and that they dislike the taste (“disgust”, “flat”, “yuck”, “nasti”, “horribl”, “terribl”). Furthermore, customers report that the cans or the order arrived damaged (“damag”) and that they do not like that the drink contains guarana (“guarana”). Additionally, the changed formula of the drink drives bad ratings (“formula”).

Lasso regression for **V8 +Energy**:

The star rating of “V8 +Energy” is positively driven by the fact that customers had concerns about consuming this product but this concerns dissolver after trying it (“concern”). This matches with the point that customers reported a surprisingly good experience (“surpis”). Customer who commute to work seem to like this drink a lot since this drives positive ratings (“commut”). Also on roadtrips this product seems to be a good choice (“trip”). Furthermore, customers like the taste (“excel”,“amaz”,“delici”, “great”, “best” & “awesom”) and the fact that the drink does not cause crash in the energy level when the effect wears off (“wear”). This goes well with the observation that customers appreciate that “V8 +Energy” does not have an unpleasantly heavy effect (“jump”). A last point which positively drives the Amazon star rating is that customer like this online-offer because the product is often not on stock in supermarkets and that the customers would like to have a stock of the product at home (“stock”).

Interestingly, the star rating of “V8 +Energy” is especially negatively driven by the taste of the product. This is contrary to the observations on drivers for positive ratings. Customers report a bland, funny, unnatural, nasty, fake, disgusting and terrible (after)taste (“bland”, “funni”, “unnatur”, “nasti”, “fake”, “disgust” & “terribl”). This matches with the reports that the drink leaves a bad feeling/taste in the mouth (“mouth”). Furthermore, customers report that the description of the product on Amazon is misleading since it does not say that the drink is sweetened with sucralose (“descript”) what fits with the negative influence on the rating that the drink is sweetened with sucralose (“sucralos”).

DIFFERENTIATE THE PRODUCTS FROM EACH OTHER - GRAPH FOR COMPETITIVE

Differentiation Factors of the 5 Products:

To clearly differentiate the products from each other, we compared the terms of our individual lasso regressions. Here, we focused on terms that are only present in one of the products. When the different regressions contained terms which were used as synonyms, we also excluded them.

The “Red Bull Energy Drink” can clearly be differentiated from the other analyzed products by the little chemical taste in comparison to other energy drinks (“chemic”). The lasso regression found that this is the biggest factor which drives the rating positively and does not occur for other products. Another good differentiation factor is that Red Bull is the only product here which contains the artificial sweetener aspartame (“aspartam”). A small research on google confirmed this fact. Customer do not like that Red Bull contains this and for all other products, aspartame did not influence the rating a lot.

One product feature which differentiates “Monster Absolutely Zero” from the other energy drinks is that customers appreciate the can sizes (“ounc”). It seems like the different offered sizes perfectly fit the needs of the customers. Since this was not a factor for the other products, we conclude that Monster offers their drinks in more suitable sizes. Also, people report to drink it daily (“daili”) and in class (“class”) which drives the ratings. These habits had not such an influence on the other products. We conclude from this that the differentiation here is that “Monster Absolutely Zero” is more consumed on a daily basis and by teacher or pupils than the other products. Although, these are not concrete product feature, we think they differentiate the use of the products and thus also count as differentiation factor.

“5-hour Energy” seems to not have such strong differentiation factors except that the can fits perfectly in pockets (“pocket”). This criterion was not mentioned for other products but also may not be very important. Although, the rating is positively driven by the fact that the product works great for diabetics (“diabet”), by looking into the reviews, we saw that this is also true for other products like “Monster Absolutely Zero” or “Celsius Healthy Sugar Free”. These products are also sugar free, but the ratings are not influenced that much by that fact. Even though, “5-hour Energy” cannot be differentiated properly by product features, we can do so by the effect. It is the only energy drink where people reported so heavy effects that a lot of them had to go to the hospital due to strokes, skin rashes and dizziness.

“Celsius Healthy Sugar Free” is the energy drink that best convinces customers of its good nutritional values. One serving has a low amount of carbonhydrates (“carb”), only ten calories (“ten”) and a high amount of minerals (“miner”). Some reviews of other drinks also include similar arguments but not in such an extent like for “Celsius Healthy Sugar Free”. We therefore interpret that this product stands out from the other energy drinks, especially in the area of health.

Lastly, “V8 +Energy” differs from the other products the energy boost by this drink is not unpleasantly heavy and that customers do not experience a crash in the energy level when the effect wears off (“wear”). Thus, compared to the other products the effects seems to be smoother and does not have such heavy impacts like e.g. “5-hour Energy”. Additionally, there seems to be a difference in the customer group since compared to the other products a lot of reviewers reported that they drink it while commuting to work or being on long road trips (“commut” & “trip”). This is not the case for the other products. We therefore conclude that “V8 +Energy” is more suitable for people who are on the road a lot and need a smooth energy boost. As a negative differentiation factor can the aftertaste be seen. Many customers reported the bad aftertaste. These reviews are significantly larger than for other products.

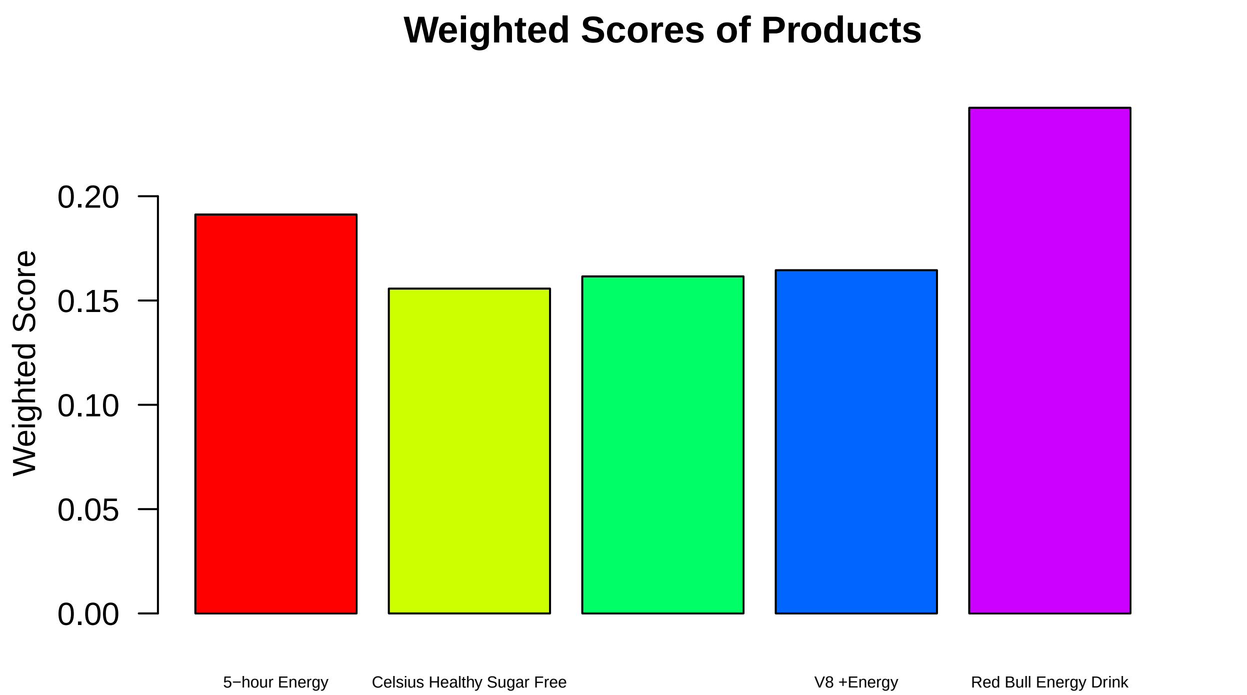
In conclusion, we can see that our products can be differentiated by the factors taste (only little chemical: Red Bull; bad aftertaste: V8 +Energy), ingredients/nutritional values (aspartame: Red Bull, low carb & calories & many minerals: Celsius Healthy Sugar Free), effect (heavy effect: 5-hour Energy, no crash: V8 +Energy), can size (good: Monster Absolutely Zero, pocket size: 5-hour Energy) and use (daily & in class: Monster Absolutely Zero; on trips/commuting: V8 +Energy). We think that these factors are the most important ones to differentiate the products from each other.

1.5 Competitive positioning

Firstly we run the lasso regression for unigrams for all products, in order to obtainthe 500 terms with highest positive influence and the 500 terms with highest negative influence.

For each product, we calculate the amount of times a term appears in the products’ review averaged over the amount of total reviews.

Then we could calculate the weighted score for each product by multiplying the term frequency with the coefficient from the lasso regression. And come up with a graph to show the competitive positioning of the products.



Red Bull Energy Drink: This product has the highest weighted score, significantly outperforming the others. This suggests that Red Bull is the leading product in terms of the factors measured. This could be due to various reasons such as brand strength, flavour preference.As analysed above for the individual products is, the Amazon star rating of 'Red Bull Energy Drink' was positively influenced by the low chemical taste and chemical composition ("chem") compared to other drinks. Many customers also seemed to like the taste of this energy drink ("nectar").

5-hour Energy: This product has the second highest score and is about half the score of Red Bull. This indicates that 5-Hour Energy is a strong competitor in the market, but still significantly behind the market leader.The score may indicate that 5-Hour Energy segments the market by targeting a specific group of people. For example, they targeted students and others who need energy to get through the day/night. When they feel tired, all they need to do is drink a 5-hour Energy drink to stay energised and focused so they can get things done. For instance, they chose students and others who need energy to get through the day/night. When they feel tired, all they need to do is drink a 5 Hour Energy drink to stay energized and focused so they can get things done. As a student said,"Awesome Tried 5-hour energy for the first time tonight in my desperation to finish a paper without falling asleep. I thought this would have the same effect as a caffeine pill and leave me jittery and anxious, but I feel fine! Drank about half the bottle and was able to concentrate and get a huge chunk of work done. Lifesaver. Will definitely be using in the future as needed.

Monster Absolutely Zero: With a score just over half that of 5-hour Energy, Monster Absolutely Zero has a noticeable presence in the market but does not significantly challenge the top two products. When customers talk about "Monster Absolutely Zero", they indicate that they are generally impressed ("impress") and particularly impressed by the delicious taste ("delici", "decent"), highlighting the influence of taste or brand image. On the other hand, it is also very common for people to complain about the price("cheap") and "health". Therefore, the positioning of products in market segments needs to be redesigned.

Celsius Healthy Sugar Free: This product's score is very close to Monster Absolutely Zero, indicating that it's a competitive alternative to Monster on the market. Its position suggests that it appeals to a similar customer base, possibly with a focus on health-conscious consumers given its Healthy Sugar Free descriptor. There is also some evidence to support this point. For example, the term "niacin" has a positive effect on the star rating of "Celsius Healthy Sugar Free", and niacin plays a role in maintaining skin health, supporting nervous system function and improving cholesterol levels.

V8 +Energy: This product has a significantly lower score compared to the others, suggesting that it is a minor player in this competitive set. This could indicate that it has a niche audience or that it does not perform well on the measured factors compared to the others.Negative comments such as "bland taste", "unnatural taste" and "disgusting" have a major impact on the appeal of their products. Improving the taste of a product is therefore extremely important if it is to gain consumer confidence and be successful.

The competitive positioning on this graph suggests that Red Bull Energy Drink has a dominant position in the market. The others fall into a hierarchy with 5-Hour Energy a distant second, followed by Monster Absolutely Zero and Celsius Healthy Sugar Free vying for third, and V8 +Energy well behind the others.

Task 2: Automated responses to complaints or concerns

2.1 Topic modeling

For task 2 we decided to use unigrams. One reason for that is that the product with the most negative reviews has about 200 of them, which is relatively small, so it could be difficult to extract meaningful bigrams. Another reason is that when exploring the data we did a topic modeling for both bigrams and unigrams and the results for the unigrams were much better interpretable than the result for the bigrams.

Since we have a lot of common reviews between some products we wanted to do the topic modeling on a subset of the data. For choosing the products we want to include in the topic modeling, we took a look at the amount of negative reviews (1 or 2 starts) for every product:

As we can see 5-hour energy has the highest amount of negative reviews, which is around 200. Since we do not think that 200 reviews are sufficient to do the topic modeling we started to include the products with the next highest amount of negative reviews. Since we already found out that there are a lot of common reviews when looking at product from one brand, we decided to only use one product per brand.

So in the end we came up with the same list of the products we chose for Task 1, which are:

"5-hour Energy" , "Celsius Healthy Sugar Free" , "Monster Absolutely Zero" , "V8 +Energy" & "Red Bull Energy Drink".

When filtering the dataset for this products, we have 745 reviews among the 5 different products, which should be distinct from each other.

After choosing the products we wanted to determine how many topics are the optimal amount for this dataset. Therefore we calculated the Bayes Factor for different amounts of topics (2-6):

The results we obtained are in the following table:

number of topics	log BF
2	6377.07
3	7645.57
4	7950.61
5	7915.1
6	7843.64

Since the model with 4 different topics obtained the highest result, we decided to use 4 topics for the further analysis. So firstly we run the topic modeling with 4 different topics:

```
##
## Top 5 phrases by topic-over-null term lift (and usage %):
##
## [1] 'whenev', 'ice', 'pineappl', 'gotitfre', 'potato' (30.2)
## [2] 'punctur', 'wrap', 'poor', 'mini', 'broken' (27)
## [3] 'scare', 'face', 'race', 'hospit', 'shake' (24)
## [4] 'formula', 'switch', 'secret', 'cheap', 'insan' (18.7)
##
## Dispersion = 1.63
```

Here we obtain the model summary, as we can see the Dispersion is 1,63. Also it shows the 5 most important words per topic, but for interpretation we want to look at more information since we have problems identifying the topics only based on those words.

So we calculated the 10 most probable words for every topic. The results can be found below:

```
## Topic 1 :
## tast     drink    flavor   like     energi   tri      juic     good     sucralos just     realli   free     sugar    bad      artifici
## 0.08418  0.04232  0.04055  0.03379  0.02867  0.02133  0.01460  0.01330  0.01305  0.01293  0.01037  0.00994  0.00981  0.00923  0.00901
##
## Topic 2 :
## can        one        star       product    order      box        two        packag     damag      receiv     open       flat       disappoint time       ship
## 0.08064    0.04566    0.04351    0.03467    0.03159    0.01907    0.01647    0.01609    0.01341    0.01290    0.01269    0.01252    0.01081    0.01020    0.01013
##
## Topic 3 :
## hour    energi  feel    work    tri     like    stuff   day     effect  time    just    dont    didnt   felt    made
## 0.02672 0.02611 0.02120 0.01915 0.01723 0.01685 0.01331 0.01158 0.01093 0.01056 0.01046 0.01036 0.01013 0.01009 0.00906
##
## Topic 4 :
## drink   energi  price   buy     product now     caffein use     much    pack    get     think   good    vitamin store
## 0.03763 0.03592 0.03569 0.02797 0.02547 0.02169 0.01855 0.01755 0.01411 0.01339 0.01315 0.01291 0.01139 0.01075 0.00992
```

After taking a look at the most probable words we came up with a first idea for every topic.

Topic 1: This topic seems to be about the taste of the energy drink. Our reason for that assumption are that both words "taste" and "flavor" seem to appear a lot within that topic. Also noticeable is that like is really high in this analysis so it seems that reviewers are rating how they liked the taste of the energy drink.

Topic 2: By looking at the most probable words of topic 2 it seems really clear that this topic is about shipping and packaging issues, since we see words like "box", "package", "ship" and "order" in this list. Another indicator supporting this interpretation are the words "damage" and "open" which could refer to the point that people received damaged packages or even a box where the contents of the package were broken (energy drink cans).

Topic 3: Topic 3 is not that clear when looking only at the 10 most probable words. Our idea was that this topic is about people reporting how they felt after drinking the energy drinks and which side effects they experienced. We come to this assumption because the words "feel" and "felt" seem to be important for that topic. Also that "effect" appears in this list is another indicator that topic 3 is about the influence of the energy drink on the people.

Topic 4: The idea we had is that this topic could be about price complains of customers. One reason for that assumption is that "buy" and "price" are the most probable words in this topic (and "money" is also part of the most probable words) and we would think that customers who use the word price in their review (which only had 1 or 2 stars) were not satisfied with the offer they got. Also maybe the presence of the word "store" could support this, because people may indicate that they could get a better price at a physical store than on amazon.

Since we had issues interpreting all the topics only from the list of the most probable words, especially topic 4, we also conducted a analysis which delivers the words with the "most pronounced difference to average per topic", where we compare the probability of a word to be found inside a topic to the average probability across all topics.

```
## Topic1 :
## flavor    sucralos  artifici  sweeten   mango     peach     sweet     green     blueberri pomegran  fruit     bzzagent  vfusion   aftertast veget
## 1.38628   1.38627   1.38626   1.38625   1.38625   1.38625   1.38624   1.38623   1.38622   1.38622   1.38622   1.38622   1.38620   1.38620   1.38619
##
## Topic2 :
## star     box      damag    flat     order    leak     ship     packag   dent     shipment expir    arriv    bust     explod   broken
## 1.38629  1.38628  1.38628  1.38627  1.38627  1.38627  1.38626  1.38626  1.38626  1.38625  1.38625  1.38625  1.38624  1.38624  1.38624
##
## Topic3 :
## felt      feel      work      heart     jitteri   tire      help      awak      minut     crash     sick      headach   sleep     experienc eat
## 1.38626   1.38625   1.38625   1.38624   1.38621   1.38621   1.38621   1.38620   1.38620   1.38619   1.38619   1.38618   1.38617   1.38616   1.38616
```

```
##
## Topic4 :
## price      complex    now        liquid     taurin     ounc       increas    formula    guarana    advertis   ridicul    cheaper    overpr     decaf      supplement
## 1.38628    1.38626    1.38625    1.38624    1.38624    1.38622    1.38622    1.38622    1.38622    1.38620    1.38619    1.38619    1.38618    1.38618    1.38618
```

Topic 1: This results seem to support our assumption that this topic is about the taste of the product, since we obtain words which are used to define the taste ("mango", "peach", "citrus"). Also the words "artificial" and "aftertaste" probably refer to the taste experience the consumer had. Therefore we concluded that topic 1 is **overall taste**.

Topic 2: Now we find even more words which could suggest that people received a damaged package ("damage", "leak", "explode", "broken" and "wet"). Here also the results support the hypothesis that topic 2 is about shipping and packaging issues. So our our definition for this topic is **shipment & packaging**.

Topic 3: Here also the results support the idea that topic 3 is about the unwanted effects of the energy drink on the consumer, because we find a lot of words which probably describe the feeling of the consumer after drinking an energy drink ("tired", "jittery", "headach", "sick"). So we defined this topic as **negative effects**.

Topic 4: Now we find even more words which could refer to the price the customers paid . Even "overpriced" is now part of that analysis. But also we find a lot of ingredients for energy drinks in this list such as taurin and guarana. But this could be because the reviewers may think that they paid to much for what is really inside the drink. So we would conclude that topic 4 refers to the **price-performance ratio**.

2.2 Predicting the dominating topic

Now we want to create a function that predicts the dominating topic of a review. So even if the customer had more than one issue with the product and mentioned different topics in his review we want to derive the most important topic of the review, so later the Chatbot can create an automated answer referring to that topic.

Therefore we created a function which takes the topic modeling and the preprocessed review we want to test as an input. Here it is important that the preprocessing of the review followed the same steps as the preprocessing of the topic modeling, otherwise the function will not find a match between the review and the topic modeling. Then we create a Document Term Matrix for this single review. After that we compare this Matrix with our topic modeling and come up with 5 probabilities (one for each topic) that indicate if the topics are part of the review. In the end we just compare the 5 different probabilities and search for the highest one. Since each probability refers to one topic we than just choose the topic with the highest probability.

After building the function we want to test based on a few examples if the topic prediction function actually returns the most dominating topic and if thats in line with what would a human think when reading the review.

1. Example

```
## [1] "The example review is:"
```

```
## [1] "One Star When they arrived they looked like they had been dropped, cans were damaged. Not happy
```

```
## [1] "The most dominant topic for this review is: 2"
```

For this review our algorithm choose topic 2 (shipment & packaging) which seems to be accurate since it seems like that the package and the contents were damaged, when it arrived.

2. Example

```
## [1] "The example review is:"
```

```
## [1] "Extremely Hazardaous Please if you love yourself or your kids-STAY AWAY.Long term effects of ex
```

```
## [1] "The most dominant topic for this review is: 3"
```

In this case our prediction also seems to work since the reviewer is referring about long-term unwanted effects of the energy drink, which could be seen as negative effects which is our definition topic 3.

3. Example

```
## [1] "The example review is:"

## [1] "Love RedBull, Hate New Amazon Price I have absolutely no problems with red bull and love the st

## [1] "The most dominant topic for this review is: 4"
```

This prediction also does seem to be in line with what a human would conclude, since the main concern of the customer is the price on amazon.

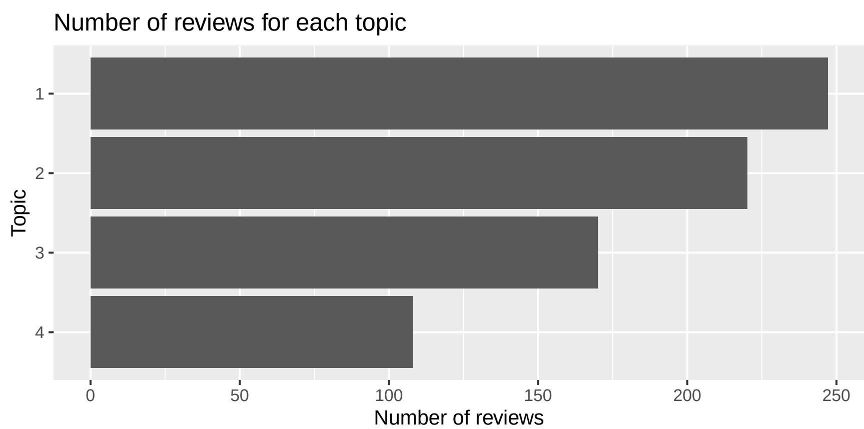
4. Example

```
## [1] "The example review is:"

## [1] "Not the best tasting, energy wise works decent but its hard ... Not the best tasting, energy wi

## [1] "The most dominant topic for this review is: 1"
```

Also in this case the prediction seem to work very well, because the reviewer has issues with the taste of the energy drink, which refers to our first topic.

After testing some samples we could see that our prediction model is able to predict the right dominant topic for the reviews since the results are in line with what we conclude when looking at the reviews.

Now we want to examine which topic is dominating the most reviews and how they are distributed. Therefore we calculated the most dominating topic for every negative review of the 5 products and plotted the distribution of the results:



As we can see the most often dominating topic is "overall taste" with more than 220 reviews. So the biggest amount of people is concerned with the taste of the products. This is followed by "shipment & packaging", which seems to be the dominating topic of around 180 reviews. The least often dominating topic when doing our prediction is the price-performance ratio.

2.3 Automated answers based on the prediction function

Now we set up GPT 4 to write automated responses to the reviews. Since we should use our dominant topic prediction for this task, we calculated the dominating topic for each review according to the function we built in 2.2.

Testing the answers

Since GPT could deliver different answers based on the instructions we set up 5 different prompts to test which one works best. After that we conducted a small survey where we showed participants answers of every prompt on 5 different reviews (index numbers: 3, 16, 23, 561, 654) and asked them to rate every answer on a scale of 1-5. Then we can derive the score for every prompt and chose the one with the highest results for our final prompt.

Since we want to show you an example that the different prompts lead to different answers of GPT we chose the following review for doing that:

"it seems like every other month I have busted cans I have a subscription that I will be cancelling, it seems like every other month I have busted cans. This month I had 3 broken cans and the mail man would not even deliver it because the box was falling apart, so I had to leave work early to pick it up from the post office. It then takes me another 20 minutes to clean all the good cans up. I love the idea of having this product come right to my doorstep but not if it creates more work for me."

1. Basic prompt

First we set up a prompt which takes the review, the topic we found out through our topic prediction approach and a short description of the topic. Also we gave it the following instruction to start with: "You are an helpful assistant that understands the customers needs to answer them. You are given a negative review of a energy drink product from consumers which are unhappy with the product. Also you are given the main topic of the review and the explanation of the topic as topic description. Given the content of review and the topic of the review formulate an answer. You have the following restrictions which you must adhere to in any case: - Do not make anything up you don't know"

Using this prompt the answer for our example review was good but way to long for a short answer to a review. Often times we couldn't even retrive an answer because the answer was to long.

2. Modifixing the basic prompt

One problem we discovered when we did the first prompt was that GPT answered from the perspective of the Red Bull customer service (or other energy drink brands). So we clarified in the instructions that this answer should be from the perspective of the Amazon customer support. Also we added a limit for the characters (500), because when we did not do that in the basic prompt we got really long answers (as seen in the example) which we feel like are not suited well for the review section on amazon.

Using this modified prompt the answer for our example review is:

"I'm sorry to hear about the issues with your shipment and packaging. We aim for your orders to arrive in perfect condition. Please contact us through your Amazon account so we can address this promptly, including arranging for a replacement or refund and providing feedback to our shipping partners to prevent future occurrences. Your satisfaction is our priority."

3. Adjustments

Then we noticed that GPT tended to start and end their answer differently. Therefore we gave it the restriction that the answer should always start with "Dear Customer" and end with " "Best regards from the customer service team". Also we saw that it include the term "price-performance ratio" in the review which is one of our topics. But we think that it is probably better to use the term price instead, since some people might not feel addressed when speaking about the ratio.

By using this adjusted prompt we are obtaining the following answer:

"Dear customer,

I'm truly sorry to hear about the recurring issues with the shipment and packaging of your energy drink order. We take such feedback seriously and aim to ensure that all products arrive in perfect condition. Please contact us directly so we can assist with a resolution, including a replacement or refund for the damaged items.

Best regards from the customer service team"

4. Specifications

For the last prompt we became even more specific in the instructions, for example explicitly explaining the meaning of the topic description. Also we added in the instructions that it should answer "like a professional customer service employee" would do to see if that maybe increases the quality of the answer. Also in some answers we saw that it recommended to look for other retailers. Since we dont feel like that this would be a suggestion of the own customer service, we added to the list of the restrictions that it should not recommend other retailers. Also in cases where the main topic is "negative effects" or "shipment & packaging" we wanted that GPT is offering the possibility of contacting the customer service for resolution, so we also added that part in the instructions.

When we use this more specified version of the prompt, the answer for the example review is: "Dear customer,

We're sorry to hear about the issues with the shipment and packaging of your energy drink order. It's important to us that your products arrive in perfect condition. Please reach out to our customer service team so we can address this matter promptly and to your satisfaction.

Best regards from the customer service team"

5. Reformulating the instruction

Since we only added things to the instruction in the last step we decided to completely reformulate the instruction to see if different wording could make a difference. But we will keep the list of restrictions that we built up through the different prompts since it includes basic things like how to address the reviewer and how long the answer should be. We changed the instructions in a way that we tell GPT to base the answer more on the topic and the topic description of the review to be able to determine whether that makes a difference in the quality of the answer.

Now the answer for our example review is:

"Dear customer, We're truly sorry to hear about the recurring issues with the shipment and packaging of your energy drink subscription. It's important to us that your products arrive in perfect condition. Please reach out to our customer service team so we can address this matter promptly and ensure your satisfaction with future deliveries.

Best regards from the customer service team"

Evaluation of the different prompts

Based on our example review we could see that the answer of GPT differs between the different prompts we are giving GPT. To be able to rate the quality of the answer and after that choose a final prompt we conducted a survey where we displayed all alternative answers for a review and asked the participants (mainly our group, but also 2 people who are not part of the course) to rate the suitability of the answer to the specific review on a scale of 1 to 5, where 5 means that the answer is perfectly suited for the concern the customer has and 1 that the answer is not appropriate at all. The participants were told that they should imagine that this answer was posted under the respective review and

Then we calculated the sum of the ratings for every answer and divided it by the amount of participants to get the mean score of every answer.

After that we summed up the score for every answer that was generated by using the same prompt and divided the result by 5 (since we use 5 different reviews in this survey). You can find the results of this calculation in the table below:

Prompt	Average rating over all 5 reviews
Prompt 1	1,93
Prompt 2	3,03
Prompt 3	3,6
Prompt 4	4,6
Prompt 5	4,53

Our results show that Prompt 1 delivered the worst answers, probably because we did not include a character limit and therefore the answers were way to long. We also can see that our modification and adjustments seemed to improve the answers of GPT. The last Prompt, where we changed the instruction did not change a lot in the quality of the answers, since the average rating of our voters is only 0,07 smaller than for prompt 4. But since prompt 4 obtained the highest score, we decided to use that prompt as our final template for asking GPT to write a response to the review. (If you want to test this prompt with your own reviews, you can find the corresponding code where you are able to insert a review at the end of this document)

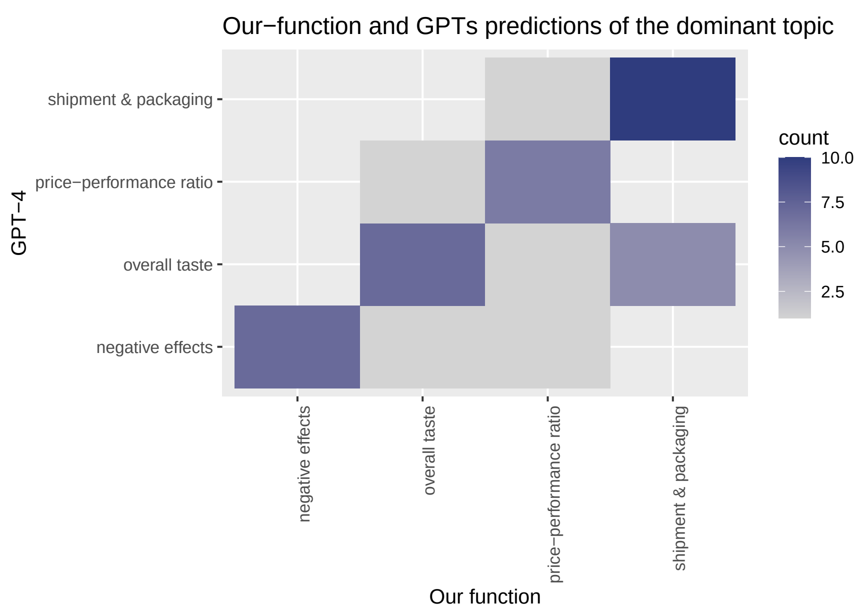
2.4 Automated answers without prediction function

For testing whether GPT generates better answers when we do not give it the topic we derived using our prediction model we developed 2 different approaches:

Predicting the dominant topic

First we want to take a random sample of 20 reviews and compare the prediction of our prediction model with a prediction of GPT. Therefore we give GPT the review and ask to return the most dominant topic of the review. Therefore we also gave GPT all the topic names and the respective description, so that GPT uses the same categories as we did. By doing that we are able to analyze whether there are differences between our prediction and the one of GPT.

The following plot shows the comparison of our function and GPTs prediction. The darker the color the more often the topic was predicted. Assuming our prediction and the one of GPT does not make a difference, than the boxes would appear only on the diagonal of the heatmap.



As we can see in most of the cases GPT predicts the same most dominant topic as we did with our own prediction function, since the boxes on the diagonal are the most darkest. For example in the cases where we predicted the “negative effects” as most dominant topic of the review, GPT did the same prediction for every of that reviews. But in some cases GPT actually predicts a different topic than we did. For example when we predicted the “price-perfomance ratio” as the most dominant topic for some reviews GPT instead predicted the “negative effects” or the “shipment & packaging” as main concern of the customer. Also when we predicted the “overall taste” as the most dominant topic in one case GPT instead assigned the negative effects topic to that review instead of the taste. In this sample the main difference between our prediction and the one of GPT occurs when we predicted “shipment & packaging” as the most dominant topic. In 5 cases GPT actually thinks that instead the main concern of the reviewer is the “overall taste”. Therefore we took a look at two examples where this occurred:

Review 1: “One Star Taste Awful!”

In this case GPT seems to be more accurate since the review is not about the shipment & packaging in any ways but only refers to the taste as the reason for the bad rating. One reason why our prediction failed in this case could be that the review is very short (4 words) which seems to be a problem for our prediction algorithm since it has to make the decision on much less information.

Review 2: “Two Stars This is the worst monster flavor.... But I still drink it”

Also in this case GPT seems to be more accurate because again the review has nothing to do with the shipment & packaging and again only refers to the taste. But as the other example also this review is relatively short. One conspicuousness is that in both reviews the term “star” appears which could maybe be a driver for our prediction to put it in the shipment & packaging category.

All in all we could see that in most cases of this random sample our prediction is in line with the one from GPT, but in for some reviews our prediction seems to fail and the answer of GPT is more accurate. Therefore we would expect that also the answers of GPT without giving them the topic of the review could be more accurate and therefore better suited.

Evaluation of the answer quality

Secondly we want to compare answers of the model of task 2.3 to answers of GPT without the input of the topic and the topic description. Therefore we selected 10 different reviews (we used the same reviews that we tested the 5 different prompts from Task 2.3 with and added the following index numbers: 333, 45, 299, 483, 101) and generated answers for both of them. Then we created a poll were we showed 6 people (including us 4 and 2 outside raters) the review and both answers and asked which one they liked better as answer for the respective review or if they are indifferent between those. Then we can derive what the majority of our poll participants prefer to get an impression which answers are better suited.

Therefore we set up a GPT prompt which has the review as the input and the following instructions: “You are an helpful assistant that understands the customers needs and gives the customer an answer to address his concerns. You are given a negative review of a energy drink product. Given the review you should address the concern of the customer by finding the main issue they have.”You have the following restrictions which you must adhere to in any case: - Do no make anything up you don’t know” - You keep your responses to no more than 500 characters long (including whitespace)” - You sign off every message with ‘Best regards from the customer service team”’ - Do not address the reviewer by name, just call him “customer”.” - Do not make anything up outside of the context that is given you!”

To show that the different prompts lead to different answers we chose one example review and generated one answer with giving the input of the topic and one without the topic as input.

Example review: “Extremely Hazardaous Please if you love yourself or your kids-STAY AWAY.Long term effects of excessive caffeine are well known Hypertension,Cardiac arrhythmia,Stroke,Aggressive Behavior and DEATH.Only yourself to blame, remember you have read this warning.”

Answer with giving the topic:
“Dear customer,We’re truly sorry to hear about your experience with the energy drink. Your health and safety are our top priorities. We take concerns about negative effects very seriously. Please reach out to our

customer service team to discuss this matter further so we can address your concerns appropriately.Best regards from the customer service team”

Answer without giving the topic: “Dear customer,We’re concerned to hear about your experience. Our product is designed to be enjoyed responsibly, and we always advise consumers to follow the recommended serving size. We take health issues seriously and will pass on your feedback to our product safety team for review.Best regards from the customer service team”

As you can see the answer without giving the topic is more about the product design and the intake recommendation as the prompt with giving the topic delivered an answer which is more concerned about the health of the reviewer.

After we got the results of the survey we calculated the score for both options. We did this by assign 1 point every time a participants chose 1 of the answers to the prompt with which the answer was generated. In case participants stated that they were indifferent between the 2 answers we assigned a point to both options. So the highest possible score for an option is amount of the participants (6) multiplied with 10 (since we asked every participant for his opinion on 10 different reviews) which equals 60. Then we summed up the points for both options and compared the score.

In 27 of the 60 cases our participants preferred the answer generated with the model without telling it the most dominant topic we found using our prediction function. In 18 cases people preferred the answer where we gave the model the main topic of the review and in the last 15 cases our participants were indifferent between the two answers.

So when we sum up the score for both options the answers without the topic as input obtained a result of 43/60 points and the answers with telling GPT the topic had a final score of 33/60 points. So we would conclude that in most cases the model without giving GPT the topic on average generates the better answers. Possible reasons for that may be that some reviews have more than issue with the product, but when we tell GPT the main topic it only refers to that whether when we do not have the topic as input it can reply to more than one issue. Another reason could be that the answers when we do not give the topic are more specific to the review, because GPT does not have more information to base the answer on. And a third reason could be that GPT is better in identifying the main concern of the customer, which we already found evidence for above.

But also in some cases the model which has the topic & the topic description as inputs produces answers which were more appealing to our raters than the other model. For example for one review regarding the price of the product GPT suggested to “consider bulk purchases”. Maybe this was a too specific suggestions for our raters or they did not have the financial requirements to do big bulk purchases.

Review Inputs

We created a small application that allows inputting a custom review. The review is then preprocessed and the outputs of the GPT functions from 2.3 and 2.4 are returned. To start the application just run the two code chunks below. The application will then automatically start in a new window.