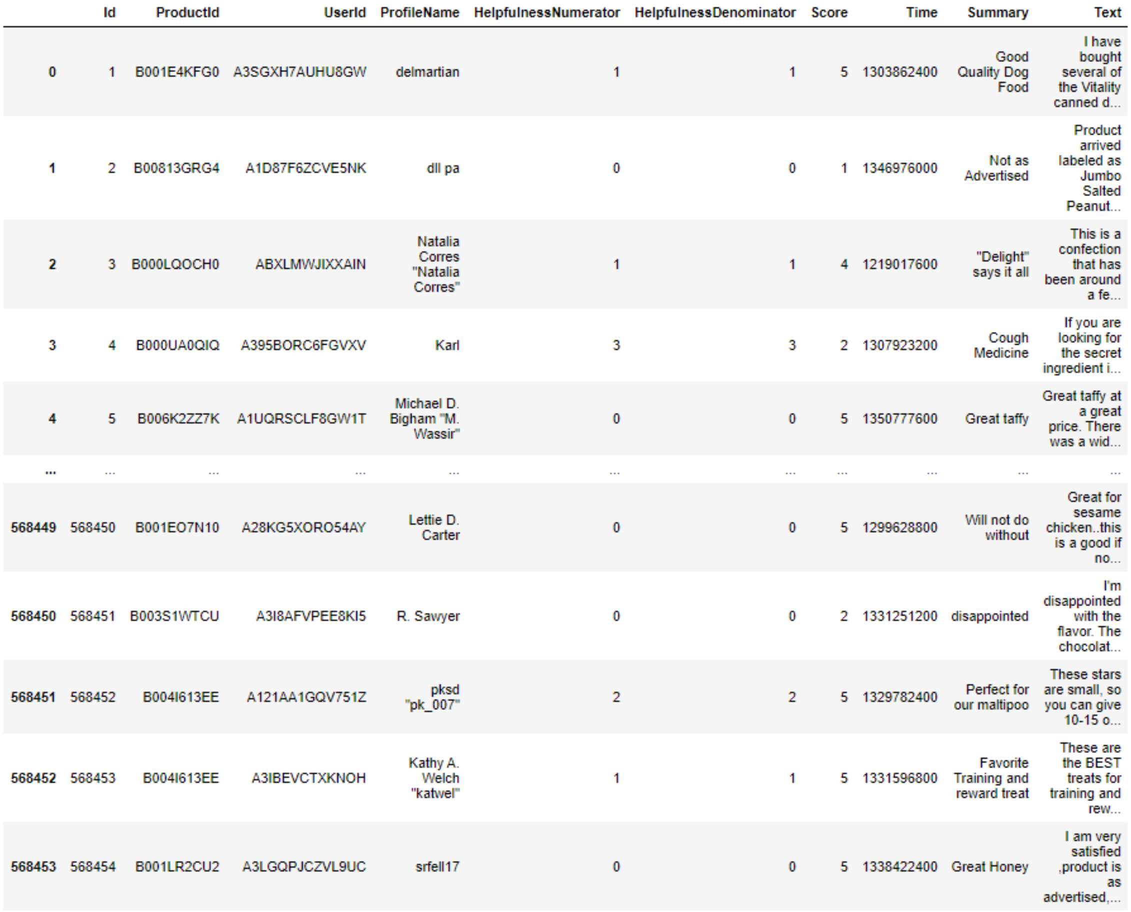
Amazon Fine Food Reviews dataset analysis

The following analysis refers to Amazon Fine Food Reviews, a *public dataset* that contains reviews of *fine foods* (non-day-to-day foods, for example caviar, wines, oysters, etc.) and other Amazon categories. Data were collected over 10 years through October 2012 and includes more than 500,000 reviews. Reviews are mainly composed of: product id, user id, score *(note* 1-5), summary of the review (Summary) and the full review (*Text*). Download made on the https://www.kaggle.com/snap/amazon-fine-food-reviews?select=Reviews.csv. The techniques of *natural language processing (NLP) – Text Blob and Latent Dirichlet Allocation (LDA) were used in* the python programming language for dataset *analysis.*

The goal of the study is to verify whether the reviews are positive or negative and, in the case of negative ones, to evaluate the probable topics of these evaluations and recommend what could be done to decrease the number of negative evaluations.



	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	5	1303862400	Good Quality Dog Food	I have bought several of the Vitality canned d...
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	1	1346976000	Not as Advertised	Product arrived labeled as Jumbo Salted Peanut...
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	1	4	1219017600	"Delight" says it all	This is a confection that has been around a fe...
3	4	B000UA0QIQ	A395BORC6FGVXV	Karl	3	3	2	1307923200	Cough Medicine	If you are looking for the secret ingredient i...
4	5	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. Wassir"	0	0	5	1350777600	Great taffy	Great taffy at a great price. There was a wid...
...	...	...	...	...	...	...	...	...	...	...
568449	568450	B001EO7N10	A28KG5XORO54AY	Lettie D. Carter	0	0	5	1299628800	Will not do without	Great for sesame chicken..this is a good if no...
568450	568451	B003S1WTCU	A3I8AFVPEE8KI5	R. Sawyer	0	0	2	1331251200	disappointed	I'm disappointed with the flavor. The chocolat...
568451	568452	B004I613EE	A121AA1GQV751Z	pksd "pk_007"	2	2	5	1329782400	Perfect for our maltipoo	These stars are small, so you can give 10-15 o...
568452	568453	B004I613EE	A3IBEVCTXKNOH	Kathy A. Welch "katwel"	1	1	5	1331596800	Favorite Training and reward treat	These are the BEST treats for training and rew...
568453	568454	B001LR2CU2	A3LGQPJCZVL9UC	srfell17	0	0	5	1338422400	Great Honey	I am very satisfied ,product is as advertised,...

Table 1 - Part of the *dataset*

In table 1, we can observe all the features described in the introduction, such as *Score* and Text*,* and the total number of observations that were recorded.

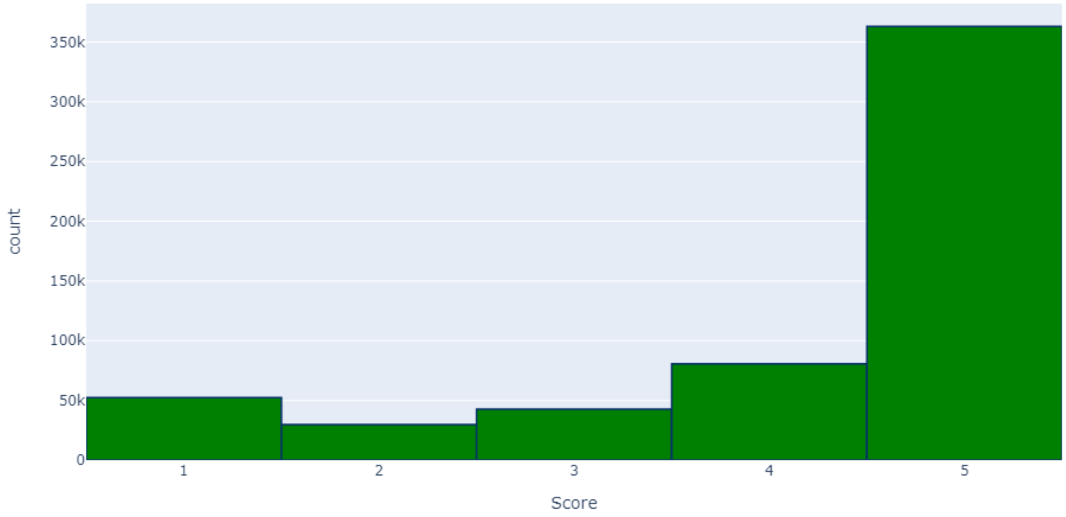


Figure 1 - Score Histogram

The first step was to analyze the *Score* feature and how it is distributed in the dataset. Its range is 1-5 and the higher the number, the better the product was evaluated. In Figure 1, we see that most reviews were rated at 5, so we can expect that most of the reviews written (Text) were positive as well.

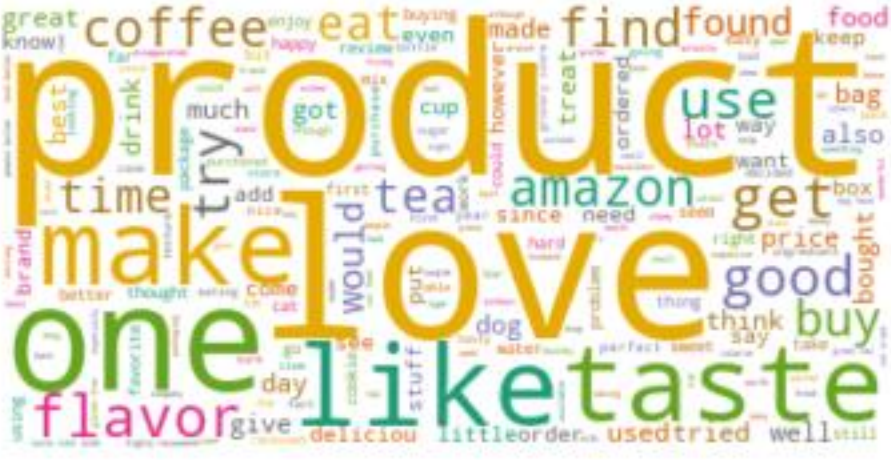


Figure 2 - Wordcloud of the entire dataset

To confirm that most reviews were positive, we have in Figure 2, featured words such as: *love*, *amazon*, *taste*, *delicious* etc. This confirms that most reviews are in general speaking well of the products purchased on the site and is positive for the company.

Then, to separate the reviews between positive and negative, a Sentiment Analysis was done with TextBlob of *each of the Text to* check its polarity, meaning -1 very negative and +1 very positive.

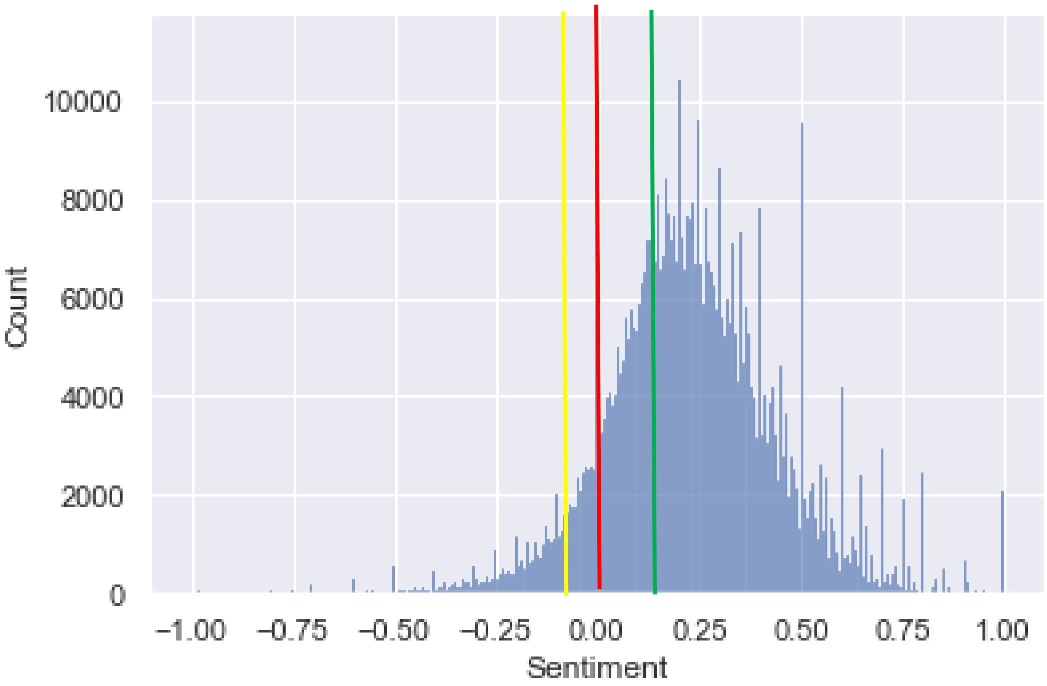


Figure 3 – Histogram do Sentiment

Figure 3 shows that most reviews are positive, i.e., they have a sentiment greater than zero and are to the right of the drawn red line, complementing what was obtained in Figure 1. However, we have 2 areas of the chart that we can classify as neutral, a neutral feeling in no review, they are: area between the red and yellow lines (1) and between the red and green lines (2). In them, the feeling is close to zero, carrying neutral.

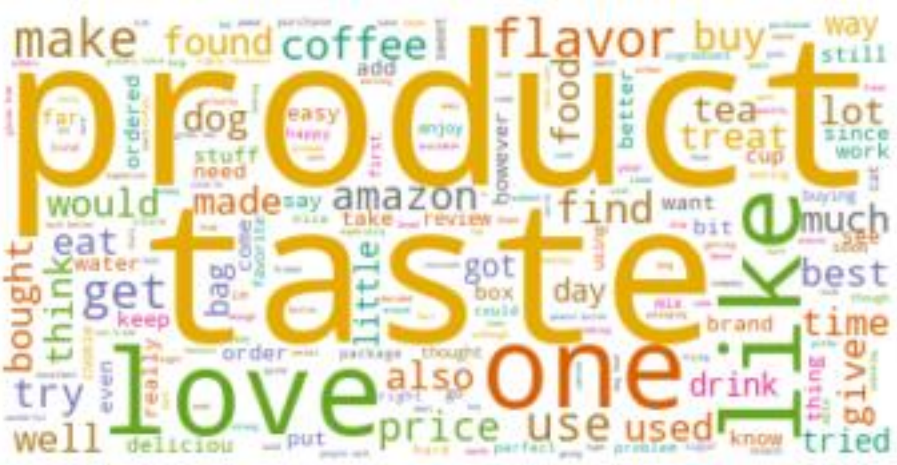


Figure 4 - Wordcloud with sentiment > 0

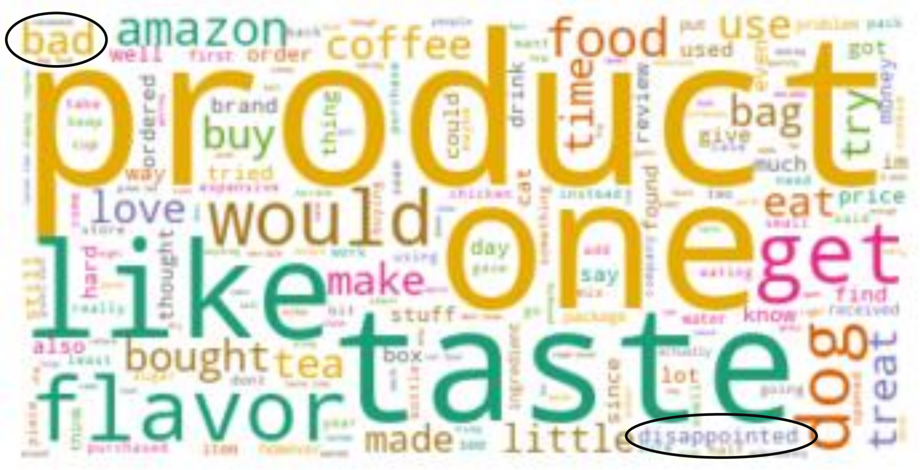


Figure 5 - Wordcloud with sentiment < 0

Looking at figure 4, we see that it is practically identical to Figure 2, that is, a large number of positive reviews. In Figure 5, negative words appear, for example, bad and disappointed, circulated in black, even if small, because part (1) was included in the process of *generating wordcloud,* indicating the negative feeling in the reviews.

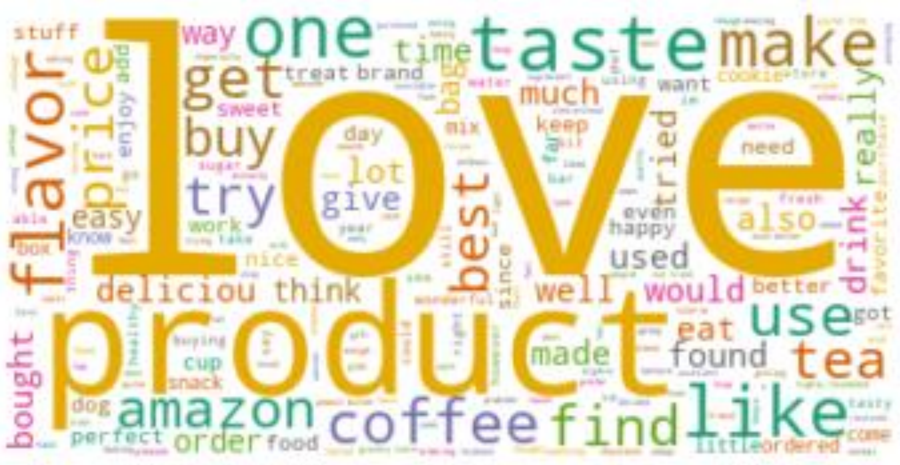


Figure 6 - Wordcloud with sentiment > 0.2

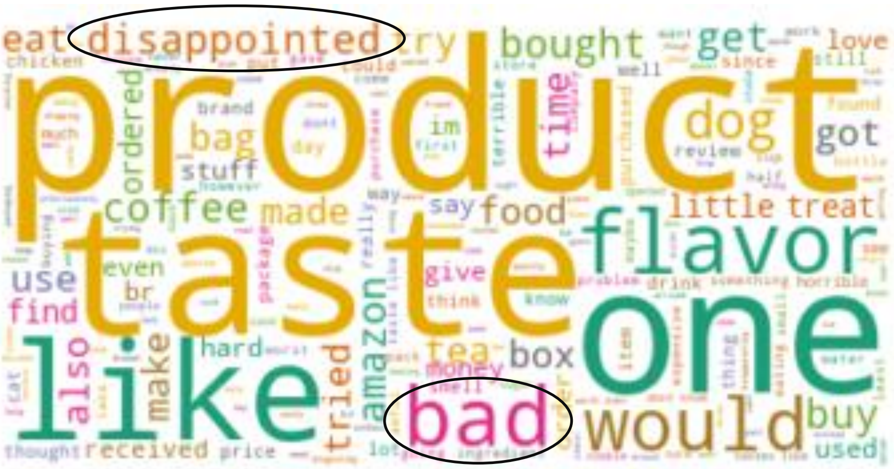


Figure 7 - Wordcloud with sentiment < -0.1

Performing the same analysis, only removing the areas (1) and (2), therefore, the neutral reviews, we obtain figure 6, and then we notice a change in the highlighted *word, love,* showing the positive feeling. In Figure 7, we see the opposite, the words *disappointed and bad* (circled in black), the same as figure 5, gained more prominence, revealing a greater negative feeling.

To obtain some information about negative reviews (area to the left of the yellow line in figure 3), we used the LDA technique for *Topic Modeling,* so it will be possible to identify the possible main topics of these negative evaluations. Configured the parameter to get 3 topics from the reviews, it was obtained:

```
[(0,
  '0.025*"like" + 0.025*"taste" + 0.016*"coffee" + 0.015*"flavor" + 0.013*"tea" + 0.011*"just"
+ 0.011*"bad" + 0.008*"tastes" + 0.008*"br" + 0.008*"product"'),
 (1,
  '0.017*"product" + 0.013*"amazon" + 0.011*"disappointed" + 0.011*"box" + 0.008*"just" + 0
.008*"order" + 0.008*"buy" + 0.008*"ordered" + 0.007*"bag" + 0.007*"received"'),
 (2,
  '0.018*"chicken" + 0.017*"food" + 0.016*"dog" + 0.011*"eat" + 0.011*"like" + 0.011*"br" + 0
.009*"treats" + 0.008*"dogs" + 0.007*"loves" + 0.007*"just"')]
```

Analyzing the topics, we have:

Topic 0: drinks, taste, taste and *bad* (with the sense of being bad)

Topic 1: disappointed, order, delivery

Topic 2: pets, pet food.

Starting with topic 2, it's not much help as it's talking about pets and there's no negative feeling present. Topic 0 addresses drinks and their taste and also features the *word bad,* which may mean that those who bought some drink did not like the taste. Topic 1 addresses order and delivery and a feeling of disappointment, with the customer not satisfied with what was delivered, whether the product or food did not meet the expectation, because in the ad was different or the taste was not so good, if it is a food, or was delivered the wrong order.

Finally, with the results obtained, it would be interesting for the company to verify that what is being delivered is what was announced, the quality of the foods and beverages that are being sold, because, as most of the *dataset* are fine *food observations,* the price of these items are high, generating high expectation, which if not met, generates bad reviews, hurting the company's sales and verifying that orders are being delivered correctly by the carrier/post office.