Exploratory Data Analysis

The intentionality of reviews written by bloggers or customers could be purely descriptive, telling readers what must they expect visiting one particular coffee shop or argumentative, sharing their own experience to incentive people to go or not at the place. Every message contains one or both of those intentionalities and for extracting the maximum of possible information, polarity and subjectivity patterns are searched on every sentence into reviews.

Global analysis of reviews takes each one as a whole entity or **paragraph**. Then, they are splitting into **sentences** for discovering the intentionality of the parts of the message. Polarity on reviews according to the scores is displayed below.

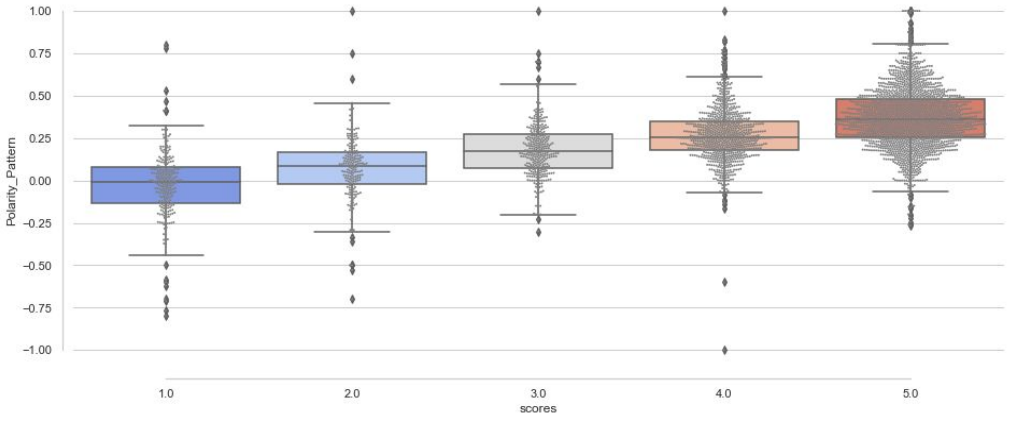

Figure 1: Polarity pattern through the scores

The polarity mean is ascending through the rise of the score, as it can be expected. How much change the positivity and negativity of data after to split paragraphs into sentences? Must we expect higher or lower mean in polarity and subjectivity distributions? The following figure displays a histogram of reviews as a whole paragraph and a histogram with reviews split into sentences. The result is that the mean shifted toward the left when we split reviews. It makes sense because we disintegrate the paragraph and determine the intentionality of every sentence. In this way, **we can find less positive messages that we could ignore in a global analysis** of reviews.

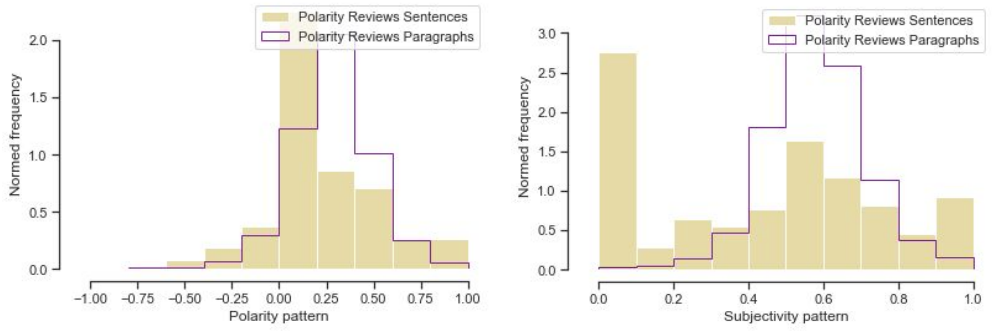

Figure 2: Histograms of Polarity and Subjectivity to reviews on paragraphs and sentences

Subjectivity analysis is interesting because the shape of the distributions is completely different: In paragraphs presence of data on the extremes is almost null and almost all data is concentrates in regular levels of subjectivity. In sentences, there is considerable information in the extremes, showing that we can find a strong presence of descriptive sentences, and argumentative sentences.

Initially, we had a dataset with almost three thousand and half of reviews. After the split, we have two thousand or rows on the dataset. All the following analysis use subsets of data choosing randomly just a thousand of rows and assuring they are representative of the entire data.

1. **Are people putting more emphasis and strong intention in positive or negative messages (polarity vs subjectivity)?**

Pearson coefficient between polarity and subjectivity patterns for reviews is 0.467. Considering that this is a random sample of our population of data, we are wondering if this statistical metric is just a chance or **exists a strong correlation between both variables** (null hypothesis). A paired bootstrapping was applied to calculate the Pearson coefficient of every sample. After one thousand trials, p_value is 0.813, **concluding that we fail to reject the null hypothesis** and the relation founded is not a chance. A correlation of almost 0.5 suggests that most subjectivity data tend to be more positive too, that means, more emphasis on positive messages, as Figure 3 shows on the scatter plot below:

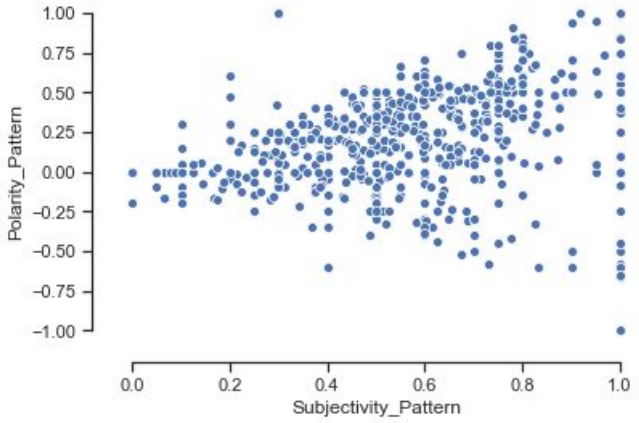

Figure 3: Polarity and Subjectivity patterns

2. **Correlation between polarity patterns and scores**

Taking a random sample of n = 1000, Pearson correlation between the polarity of sentences and the score of the review that they belong to is 0.3. To know if this is a chance, a paired bootstrapping is applied to calculate the Pearson coefficient of one thousand trials. p_value is 0.136, concluding that we fail to reject the null hypothesis and the **correlation between polarity and scores founded is not a chance**.

3. Law of primacy in persuasion

About the location of the information on every paragraph, could be tested the law of primacy in persuasion? This law, with detractors and praises, holds that information locates first has more effectiveness in persuasion.

It was analyzed the polarity and location of sentences inside reviews with the lowest and highest possible score (1 and 5 stars). The first and second half of the message are two categories observed separately. Figure 4 displays the histograms of both categories together to contrast the shape of the distributions.

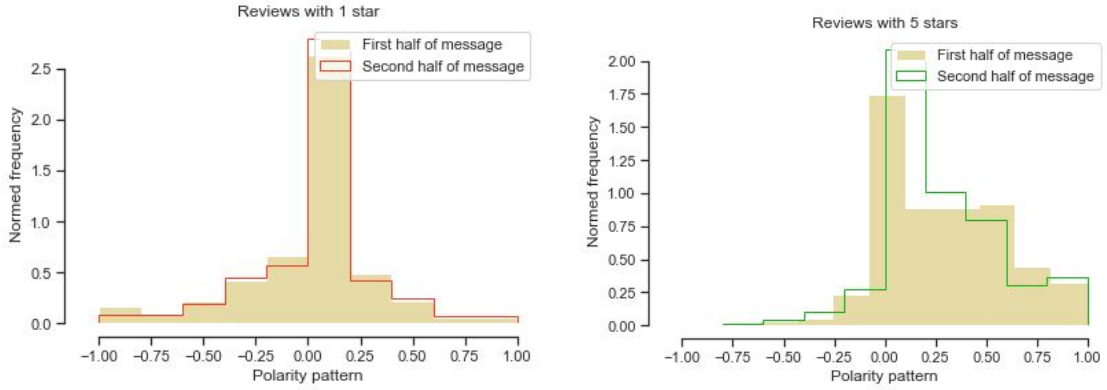

Figure 4: First and second half of messages for best and worst reviews

The lowest rating indicates similar distributions to both parts of the message. Using a t-test, we measure whether the average value differs significantly between these two independent samples. To do that, we choose the sample with less data (size n) and select randomly n rows of the second sample m times (in this case m = 1000) and recalculate p-value for the t-test. Finally, we want to know how often p-values were more than the significance level of 0.05. For 1000 tests, we got at least 75% times p-value is more than 0.05. Then, we fail to reject the null hypothesis of equals averages. However, using the same random samples, we measured how often the first half of the message has a polarity mean less than the second half of the message (more negative intentionality) and it got 97.4%. According to the test, we can't observe a significant difference between both distributions, but **there is a slightly more negative intention in first half respect to the second**.

Doing the same t-test to the highest rating, we got a p-value << 0.05 (around 0.005) then we reject the null hypothesis. **Distributions don't have equals expected mean values** and in 100% of the trials, the first half of the message is more positive respect to the second part.

Conforming to results, we don't reject the law of primacy in persuasion.

4. Polarity distributions of customer and blog reviews

Considering that we have two sources of information, can we find differences in polarity distributions of bloggers and customers reviews? Inspecting the boxplots of

customer and blog reviews, outliers were deleted using z score criteria. Figure 5 displays the cumulative distribution for both sample data. A t-test was applied between the two independent samples to measure whether the expected averages are equals. The p-value founded was 0.9483 and for subjectivity distributions, we got a p-value of 0.4441, suggesting that we **can't reject the null hypothesis of identical average score between customers reviews and blog reviews for polarity and subjectivity**. Then, we conclude that there is no significant difference between both distributions.

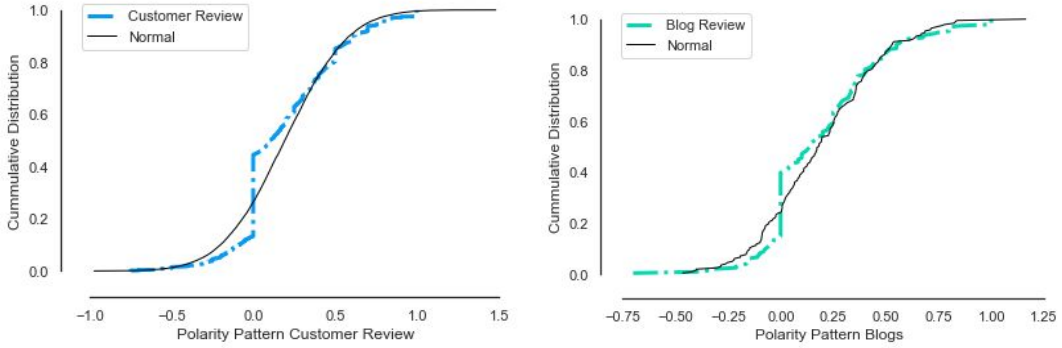

Figure 5: Cumulative distribution of Polarity for customer and blog reviews

5. Periods between positive and negative reviews

Finally, could we find any negative impact of bad scores/reviews on how long take to write the next review for a specific coffee shop?

The difference (in score and time) between consecutive reviews was calculated and we tabulated the difference between scores as *delta score* and the difference in days as the *period between reviews*. Following graph (Figure 6) displays the distributions of the most interesting categories. We are curious about the extremes *delta score* because they present the most significant differences. A *delta score* of *+4* represents a first user that put 1 star in his review followed by another that put 5 stars. Delta score of *-3* or *-4* represents the opposite. We are wondering if the period of positive *delta score* is longer than the inverse.

The mean for positive and negative *delta score* is 5.6 days and 4.92 days respectively. Observing distributions in Figure 6, we can see that a *positive delta* score of *+4* has longer periods between reviews that a negative *delta score* of *-4.*

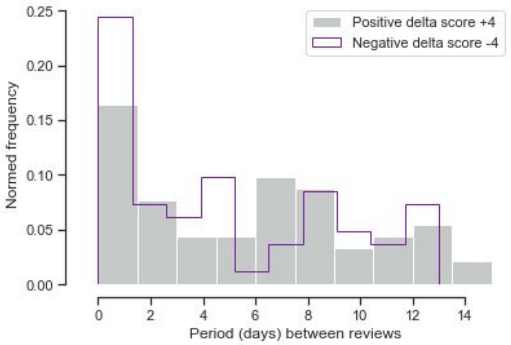

Figure 6: Histograms of positive and negative delta score of ± 4

After to compare all the possibles *delta scores* (Figure 7), we conclude that the positive *delta score* of *+4* takes more time than the rest of *delta scores* and it suggests that the hypothesis of the negative impact can't be rejected. We don't have evidence to know if the bad scores affect the fluency of people in coffee shops, because we only know, through the dates, when people are writing reviews, but the result at least says that the frequency of reviews could be affected.

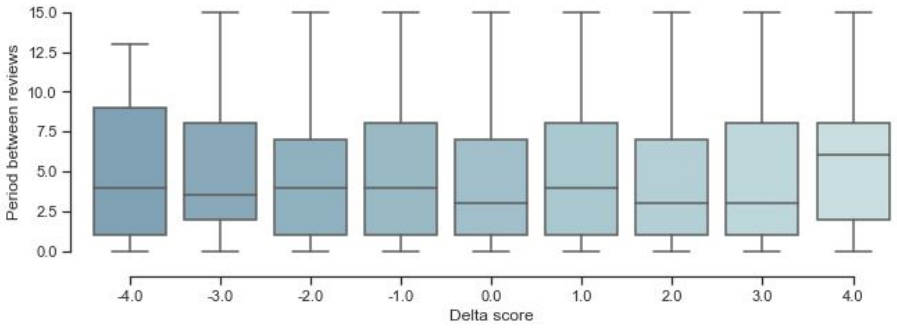

Figure 7: Boxplots of delta score in all the possible range