Andrea Marelli 1075139, Luciano Rota 1066942, Chiara Torri 1066761

ASSIGNMENT OF TEXT MINING AND SENTIMENT ANALYSIS COURSE

INTRODUCTION

The aim of this assignment is to make sentiment analysis and topic modelling on scraped reviews from a product available on Amazon. We chose as product Apple iPhone 11, 64GB, Black (Renewed), and scraped all the titles, stars and comments, in order to capture the polarity of the product reviews and try to infer how many and which are the topics discussed by the users.

DATA

We produced a code that scraped all the data needed for the analysis:

1. We first scraped the "Product Details", without the rank or the number of customers reviews, and saved it in *prod_det_split*;
2. Then, we scraped the number of customers ratings (*prod_nrat*), finding 2167 ratings;
3. The fastest delivery date was obtained by firstly scraping the entire "id" html tag containing, among other details, the date; secondly creating a regex; thirdly vectorizing it and finally accessing to the vector element of interest, the single date ("Monday, 26 June").
4. Finally, we proceeded with reviews scraping, creating a tibble (*data*) with variables "*title*", "*text*", "*star*", "*page*", "*doc_id*".

Before diving into the analysis, we had to face the problem of having the reviews written in different languages. Hence, we performed language detection, with the help of the library *cld2*, to identify the reviews in English, German, Spanish, French and Italian (see *Figure 1*), and we decided to drop the other languages because they were only a small percentage. After this, using the library *deeplr*, we translated in English all the reviews, creating the tibble *transl_data*. However, we noticed that for some documents the translations were not so accurate, so we considered also this problem when performing the analysis.

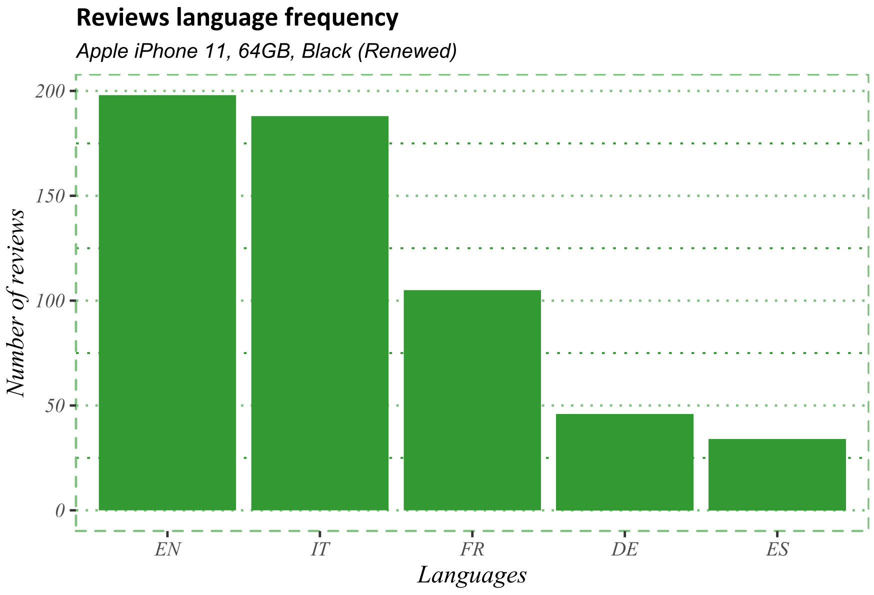

Figure 1

In order to apply a preliminary analysis of the sentiment, we studied the distribution of the stars. We considered as "*positive*" a rating of 4 or 5 stars, and as "*negative*" all the numbers below: 37.5% reviews are associated with "*negative*" score, 62.5% with "*positive*". It is also interesting to take a look at the distribution of the reviews length among different stars and sentiments (*Figure 2*): negative comments seem to be longer than positive ones.

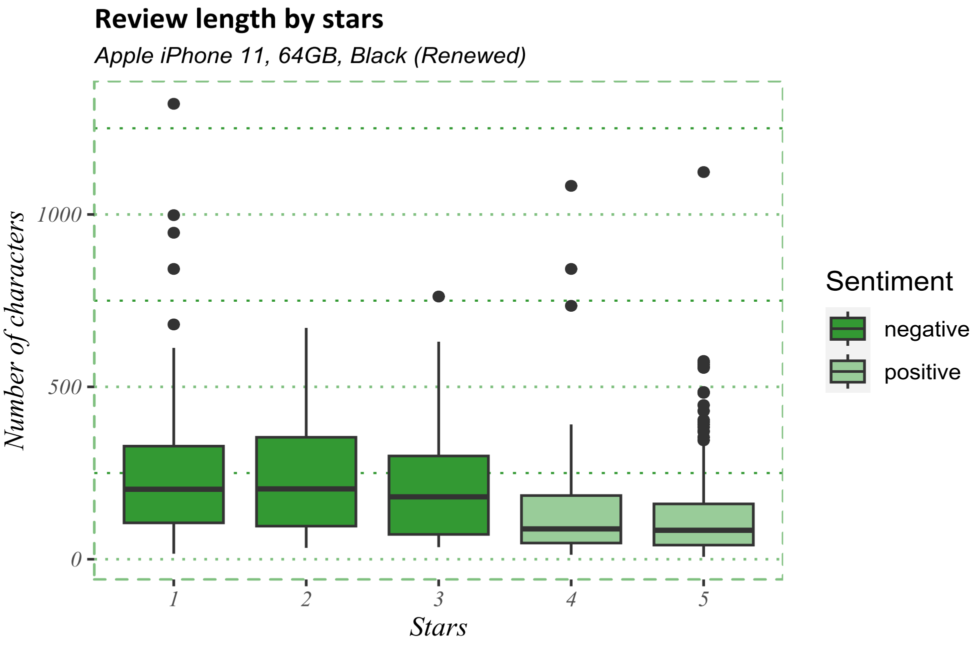

Figure 2

RESULTS AND DISCUSSION

Sentiment analysis

In order to perform sentiment analysis on the reviews of the product, we decided to apply the tidy and the udpipe approaches on the tokens and using the *afinn* lexicon. We applied the udpipe sentiment analysis adding the negators, amplifiers and deamplifiers provided by the library *qdapDictionaries*, with a weight of 0.8, and considering 3 words before the token.

For what concern the stopwords, we removed them for the tidy approach, but not for the udpipe. We chose this approach because many of the negators, amplifiers and deamplifiers are included in the stopwords list, leading to an almost complete loss of the benefits that derived from the inclusion of the valence shifters in the analysis.

Regarding the lexicon composition, *afinn* is a general purpose lexicon, so it may not be the most suitable to analyse comments related to technological goods. Nevertheless, the reviews were mostly not technical, hence this lexicon may still be appropriate.

It is also noteworthy that due to the imperfect translation of some comments and our decision about the stopwords, the tidy approach was able to compute the polarity scores for 403 documents only. Because of that, we performed the comparison only for the documents that were assigned a score by both methods, to obtain more consistent results.

Looking at the distributions of the scores of the two approaches, we noticed that the udpipe has a higher mean with respect to the tidy approach (respectively, 2.80 and 1.47). The same applies to the median (respectively, 3 and 2).

Moreover, the udpipe distribution is more dispersed (the standard deviations are respectively 5.22 and 3.78) and the minimum and maximum values are higher in absolute values. This confirms our expectation about the effect of the inclusion of valence shifters in the analysis.

Generally speaking, the distributions of the scores for the two methods have a correlation of 0.81; they are displayed in *Figure 3*.

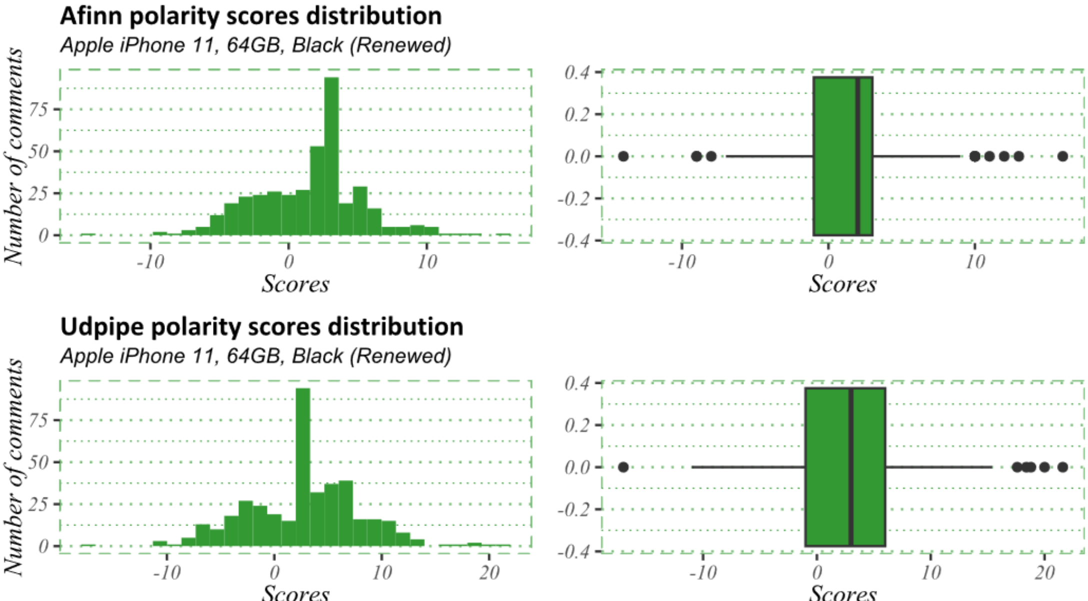

Figure 3

We further compared the two distributions by reclassifying the polarity on a negative-neutral-positive scale and producing a confusion matrix, shown in *Table 1*.

		udpipe		
		negative	neutral	positive
afinn	negative	86	7	23
	neutral	7	4	13
	positive	9	3	251

Table 1

The documents that are misclassified by the tidy approach are especially represented by false negative. This misclassification is probably caused by the negators that, if neglected, cause a shift in the polarity of the documents. Concerning the misclassifications related to the neutral category, it is instead likely that they have been generated by the absence of amplifiers and deamplifiers.

Given the benefits of the udpipe approach, we decided to deepen the sentiment analysis only using this method. In particular, we investigated the word contribution to each sentiment.

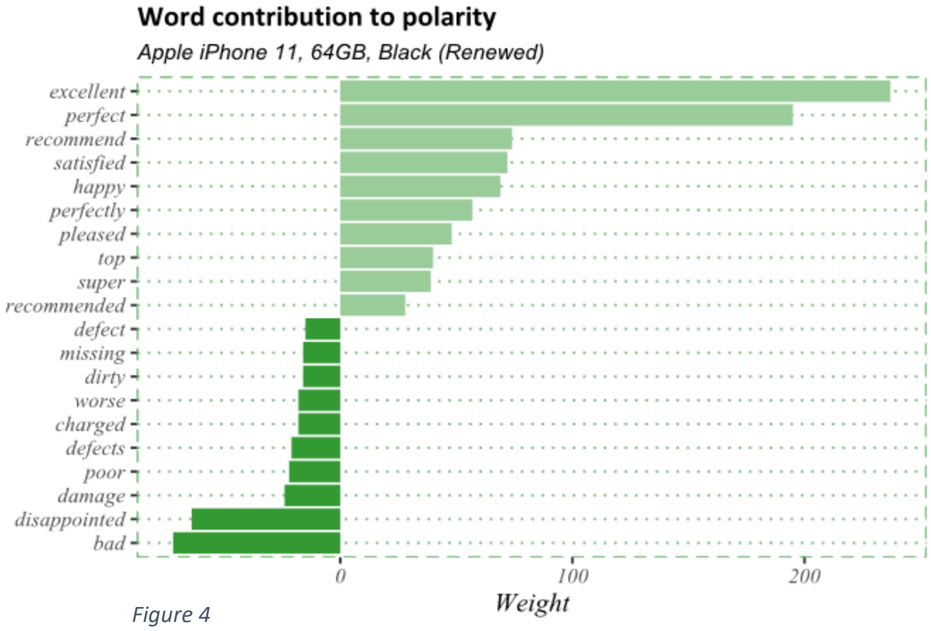

Figure 4

Figure 4 shows the words that contributed the most to both "positive" and "negative" sentiment. The contribution weight is measured as the product of the value assigned to the word by *afinn* and its frequency.

Finally, we also made a comparison between the sentiment obtained from udpipe and the sentiment obtained from the stars of each comment. As shown in *Figure 5*, udpipe and stars agree on almost all the positive comments, but not for the negative ones.

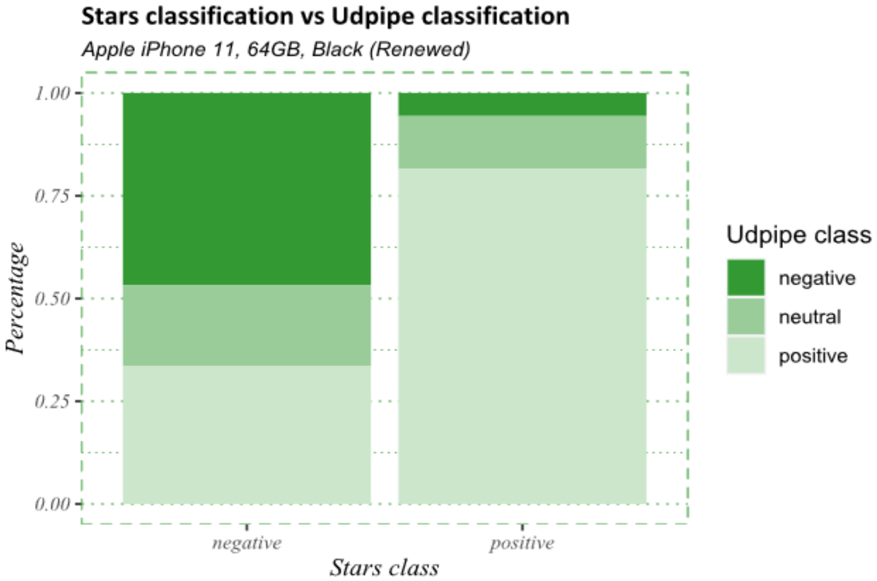

Figure 5

Topic modeling

In this section we performed LDA to understand which are the reviews topics.
We started from *transl_data*, and after doing tokenization and removing stop words we transformed it into the document-term matrix *dtm*, which is the object we need for LDA.
In order to find the best number of topics *k*, we first divided *dtm* into a training and a test set (75%-25%). The training set was used to run LDA for different values of *k*, while the test set was used to compute the relative perplexity score; the value of *k* that minimizes the perplexity score is (theoretically) the suggested number of topics to use.
We considered values for *k* ranging from 2 to 30, and we set the value of the hyperparameter α equal to 0.01 because a-priori it is reasonable thinking that each review covers just one topic.
Figure 6 shows the perplexity score for each value of *k*.

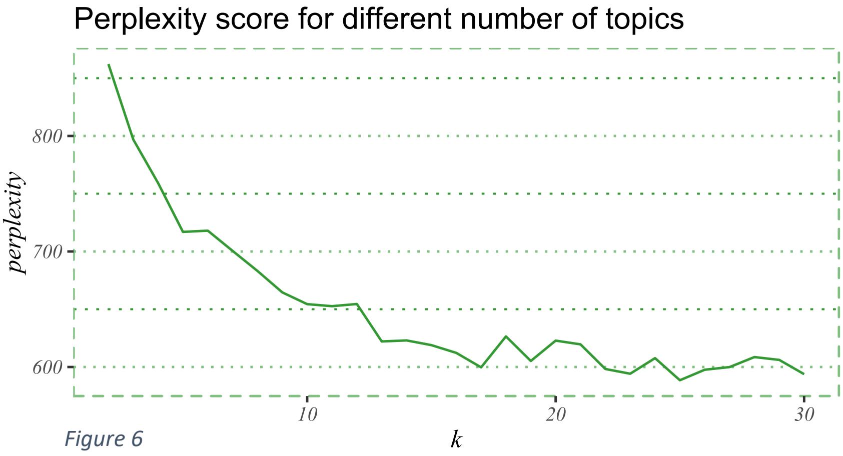

Figure 6

The score is minimized in correspondence with *k*=25; however, if we follow the elbow rule, we can be satisfied with a number of topics equal to 13 since the variation in the perplexity from there onwards is not so significant.
Even though this approach is statistically valid, it is unlikely our reviews talk about 13 different topics; furthermore, when we looked at the most representative words, a large part of the topics was basically characterized by the same

terms, making it difficult to disentangle what the topics were about. For this reason, we decided to abandon this approach and to rely on our human judgement.
Since we are dealing with Amazon reviews about a mobile phone, we thought that perhaps it is more proper to set a low number of topics, also for a matter of interpretability; after some attempts, we decided to set *k*=3, that gave us the best results. *Figure 7* shows the ten most representative words for each topic.

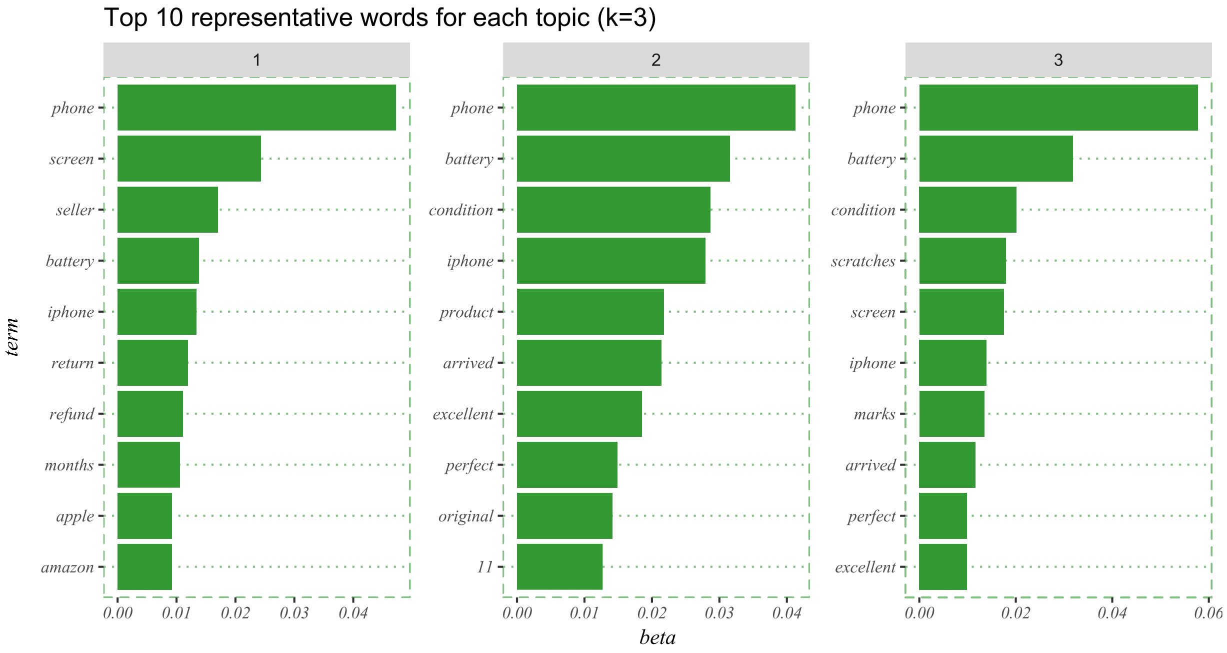

Figure 7

Topic 1 has as peculiar words "seller", "return", "refund" and "months", so it is reasonable to think that this topic is about the extent to which it has been easy for the customer to return the product and to obtain the refund. We labelled this topic as "*return*".

Topic 2 and 3, instead, are more difficult to disentangle since there are no particular words that distinguish each other. Thus, we focused on those words in Topic 2 that are different with respect to Topic 3. Given that in Topic 2 there are words like "delivery", "packaging", "fast", it becomes clearer that this topic is about the delivery and the product opening; Topic 3 is instead more about the product itself, with words like "life" (referring to product and battery life), "display" and "scratch". Thus, we labelled these topics respectively "*delivery and product opening*" and "*product*". Overall, given the context, we think that the topics we isolated are quite legit.

As a last resort, we assigned each review to the topic with the highest gamma value.
We found that most of the reviews, 257, are about the topic "*delivery and product opening*", while 136 are about the topic "*return*" and the remaining 167 talk about the "*product*" itself. According to these results, it is to be highlighted that many customers were not satisfied with the product and returned it; looking at the words that characterize the topic "*return*", this could be due to problems related to the screen or bad battery performance, and especially with respect to the latter, this may be the case since we are talking about a refurbished mobile phone. Furthermore, as we find "screen" and especially "battery" also in the other topics, we could argue that those are the main problems customers face with this product.

CONCLUSION

In conclusion, after performing sentiment analysis using tidy and udpipe approach, we found that in general the contribution of positive words is larger than that of negative ones, and this result is confirmed also by the comparison with the stars distribution. As regards topic modeling, we identified three main topics: "*return*", which results clearly from our analysis, "*delivery and product opening*" and "*product*", which on the contrary were harder to disentangle.