ASSIGNMENT – TEXT MINING AND SENTIMENT ANALYSIS

CASE STUDY REPORT

1. INTRODUCTION

The assignment ask us to choose a product sold on Amazon UK with at least 200 reviews, to scrape the main information about it and all the reviews.
We also implemented a sentiment analysis and a topic modeling using the reviews obtained from the website.
The sentiment analysis is the process of extracting the emotional content from a text. For this purpose we applied the dictionary-based approach, which considers the text as a combination of its individual words and the sentiment feature of the whole text as the sum of the sentiment content of the individual words. During this report we will exploit the procedures needed to run this analysis and also provide an evaluation of the results.
Regarding topic modeling, it tries to respond to the need of organizing a large collection of documents into natural groups. The goal is to discover the underlying topics through the analysis of the words in the original texts. In our research we will rely on a specific method to fit a topic model: the LDA (Latent Dirichlet Allocation). We will offer the insights about the methodology and its findings.

2. DATA

2.1 - PRODUCT DESCRIPTION

We chose the product "Apple 2020 MacBook Air Laptop M1 Chip, 13" Retina Display, 8GB RAM, 256GB SSD Storage, Backlit Keyboard, FaceTime HD Camera, Touch ID; Silver".
The web scraping techniques allowed us to get all the relevant information about this item, including the number of ratings (3.855 ratings) and the fastest delivery date (9^{th} June, as of today June 5^{th}).
In Figure 1 we provide a table summarizing the main technical details about the Macbook Air.

	Specifics	Technical_Detail
1	Display	Retina display 13.3-inch (diagonal) LED-backlit displa...
2	Processor	Apple M1 chip; 8-core CPU with 4 performance cores ...
3	Graphics and Video Support	Apple 7-core or 8-core GPU
4	Charging and Expansion	Two Thunderbolt / USB 4 ports with support for: Char...
5	Wireless	Wi-Fi: 802.11ax Wi-Fi 6 wireless networking, IEEE 80...
6	In the Box	MacBook Air; 30W USB-C Power Adapter; USB-C Char...
7	Height	0.41-1.61 cm (0.16-0.63 inches)
8	Width	30.41 cm (11.97 inches)
9	Depth	21.24 cm (8.36 inches)
10	Weight	1.29 kg (2.8 pounds)
11	Release Date	11/17/2020

Figure 1 – Technical details Apple 2020 MacBook Air

2.2 DATA DESCRIPTION

As we said in the introduction, it was necessary to scrape all the reviews available on the product webpage.
The reviews come from both the UK and other non-UK countries: we were interested only in the ones written in English. Hence, we had to detect which language was used to write them and after filtering we were able to obtain 459 reviews.

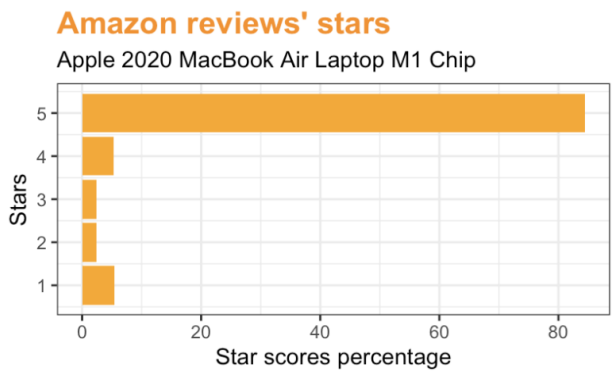

Figure 2 – Percentage of star scores assigned to reviews.

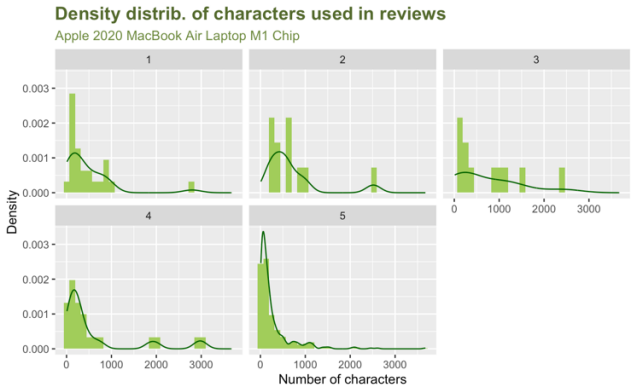

Figure 3 – Distribution of number of characters used to write reviews grouped by star score.

The output from our scraping activity is a data frame containing:

- Review title (character variable)
- Text (character variable)
- Number of stars (character variable), which we turned into a quantitative variable containing the number of stars for each review.
- Page number (numerical variable), specifying the page in which reviews are contained on the website.

Furthermore, a numerical ID was assigned to each review with the scope to univocally identify each of them.
After deriving the data, we can do some preliminary analysis through some graphical visualization provided in this report.
As represented in Figure 2, more than 80% of the users that have written a review assigned 5 stars to the product, suggesting an apparent satisfaction about it.
Figure 3 provides the density distributions of number of characters used to write reviews. Each distribution is relative to a single star score: all of them seem to exhibit right skewness independently on the number of stars assigned, suggesting that users tend to use lower numbers of letters (generally lower than 1000) to write their opinions.

3. RESULTS AND DISCUSSION

3.1 SENTIMENT ANALYSIS

Our sentiment analysis has been conducted using the dictionary-based approach.
As lexicon to be used in the research, we have chosen the "Bing" lexicon that contains 6,786 words: 2,005 with positive sentiment and 4,781 with negative sentiment.
Note that we could have used other lexicons to run our research, like "afinn" and "nrc".
The "afinn" lexicon assigns to each word a sentiment consisting in a numerical value that ranges from -5 to 5. However this dictionary contains 2476 words: "bing" could suit better the analysis due to the higher number of words that considers compared to "afinn".
The "nrc" lexicon assigns to each of the 13901 words contained an emotion coming from a large set (fear, anger, joy…). Therefore we think that such a large set of categories does not fit our analysis, since our aim is to discover the negative or positive sentiment of reviews that are written to evaluate a product.
Running our analysis, we have to keep in mind that the lexicons do not consider the grammatical structure, the word order or the context in which words are used.
The sentiment analysis was performed by using two different methods, both based on the chosen lexicon: the tidy approach and the udpipe approach.

3.1.1 TIDY APPROACH

To start off the analysis we have transformed the data in the tidy format, which implies splitting the text into a single word per row. Contextually, we removed the stop words and all the digits.

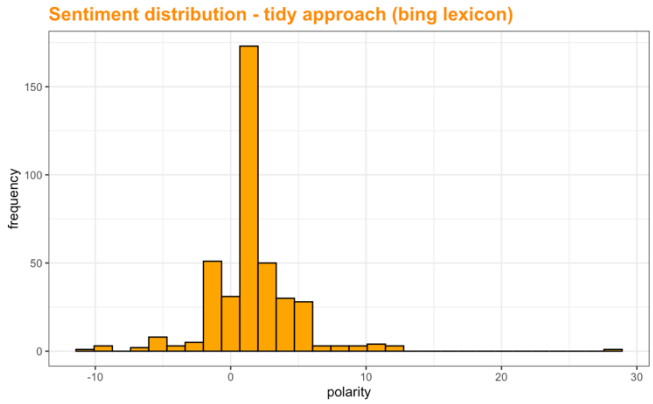

Figure 4 – Sentiment distribution from the tidy approach.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
-11.000   0.000   1.000   1.639   3.000  28.000      57
```

Figure 5 – Summary statistics for polarities from the tidy approach.

pol.sentiment <chr>	n <int>	percentage <dbl>
negative	73	15.904139
neutral	31	6.753813
positive	298	64.923747
NA	57	12.418301

Figure 6 – Frequency table (tidy approach): how many reviews report negative, neutral or positive sentiment.

We proceeded by treating the data in the tidy format: we matched the words contained in the "Bing" lexicon with the ones contained in our data, performing an inner join in order to obtain the values present in both datasets.
We counted the number of words associated with negative and positive sentiment for each review and we computed the difference between the two values to get the sentiment polarities.
A visualization of the sentiment distribution is provided above (Figure 4).
Looking at the summary statistics (Figure 5), note that our analysis produced 57 NA values out of 459 reviews: these values are linked to documents that have not been classified.
Moreover, the frequency table reported in Figure 6 tells us that the majority of the reviews (298 reviews --> 64.92 %) reported a positive sentiment polarity.

Figure 7 – Wordcloud (tidy approach): words contribution to the sentiment (green -> positive; red -> negative).
The bigger the size of each word, the higher the

3.1.2 UDPIPE APPROACH

The udpipe approach requires a specific udpipe format to be applied to the data: therefore we began our analysis by converting the original dataset.

For the purpose of the study, we also transformed the sentiment values “negative” and “positive” assigned by the lexicon into into -1 and +1 respectively.

This kind of method allows us to consider qualifiers (i.e. negators and amplifiers) that can contribute to the emotional content of a text but are not normally considered while implementing the tidy approach.

Therefore, we decided to use the following specifications:

- Negators: "not", "no", "neither", "none";
- Amplifiers: "like", "really", "very", "many";
- Weight: 0.8;
- Number of words both before and after: 2.

We calculated the sentiment polarity of each review using the udpipe model and obtained the overall sentiment distribution (Figure 8).

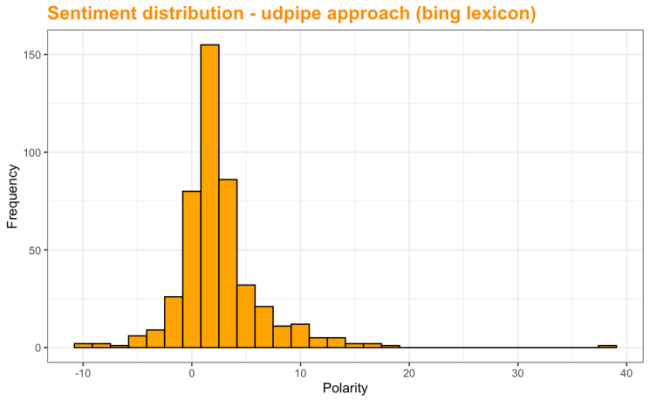

Figure 8 - Sentiment distribution from the udpipe approach.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-10.000   0.000   2.000   2.444   3.800  38.200
```

Figure 9 - Summary statistics for polarities from the udpipe approach.

pol.sentiment <chr>	n <int>	percentage <dbl>
negative	48	10.45752
neutral	73	15.90414
positive	338	73.63834

Figure 10 – Frequency table (udpipe approach): how many reviews report negative, neutral or positive sentiment.

We can now do some comparative evaluations between the results produced by the tidy approach and the udpipe one.

First, the udpipe frequency table provided in Figure 10 tells that 73.63% of reviews exhibit a positive sentiment polarity, while in the tidy approach the reviews with positive polarity represent 64.92% (Figure 6).

Second, the sentiment analysis conducted with udpipe has produced no NA values: all documents have been classified, unlike the tidy approach which produced 57 NA values.

In the picture showed in next page (Figure 11), the sentiment polarity distribution resulting from the two methods are represented: one difference we can highlight is that the udpipe approach has assigned a polarity value equal to zero (neutral reviews) to a higher number of reviews. This fact can also be observed looking at the frequency tables provided before (Figure 6 and Figure 10).

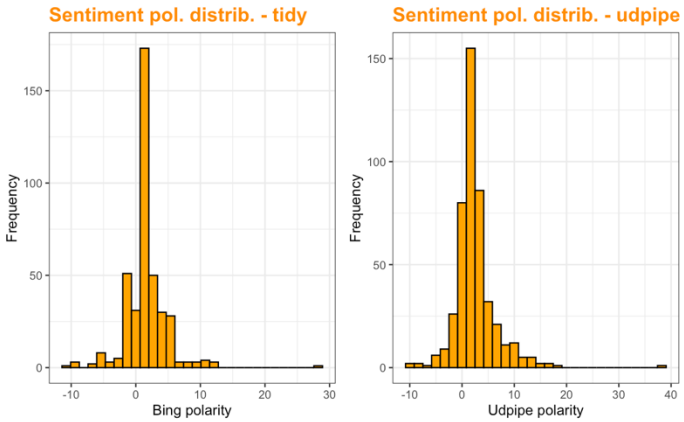

Figure 11 – Histograms representing the distributions of sentiment polarities for the tidy and udpipe approach.

One general dissimilarity between the two approaches examined until now concerns the stop words removal.
When implementing a sentiment analysis using the tidy approach, we usually remove the so called "stop-words" with the objective to delete the low-level information from our text and give more focus to the important information.
Instead, the udpipe approach does not require the stop-words removal during the data processing. If some words classified as stop-words are contained in the chosen lexicon, this could lead to differences in the results compared to other approaches.

3.2 TOPIC MODELLING - LDA (Latent Dirichlet Allocation)

As the next step of our research we implemented the LDA model, which is an unsupervised machine learning method that simultaneously estimates the mixture of words associated with each topic and the mixture of topics that describes each document. It uncovers latent topics that are not explicitly defined by the text miner, seeking to identify the hidden group of words representing each topic. Dirichlet distributions are employed to model the word allocations in LDA.
The implementation requires the dtm (document-term matrix) format for our data, so the first step was about converting our dataset.
The LDA model demands to specify the number of topics before running it: we chose to solve this question through the computation of perplexity values. Perplexity provides a measure of how well the topic model predicts new data.

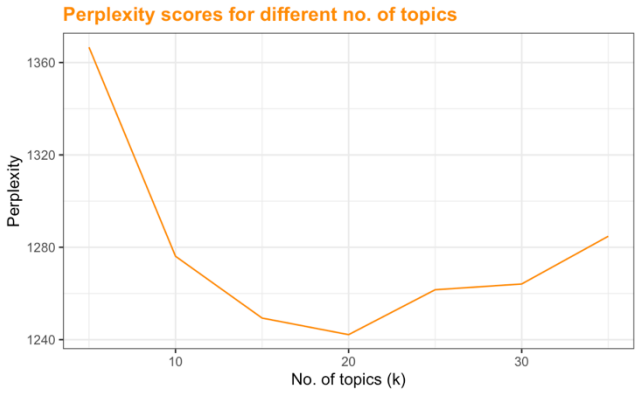

Figure 12 – Visualization of perplexity scores associated to number of topics from 5 to 35.

We split the data into a training set and a test set containing respectively the 75% and the 25% of the data.
After the splitting, we carried out multiple computations of perplexity based on different numbers of topics ranging from 5 to 35 with an increment of 5. Recall that generally the lower the perplexity, the better the fit.
Looking at the results (Figure 12), an ideal number of topics could be one between 10 and 20, where the values start to stabilize and the difference from one number to the other starts to look irrelevant. For simplicity we decided to use a value of topics equal to 10.
Following the choice of our parameter, the next step consisted in running the model.
We proceeded by looking at the beta distribution, which contains the probability of a term being generated by a specific topic.
A visualization of the top 10 words per topic based on the beta values is provided in Figure 13.

Considering the top terms, theoretically a first classification for the topics could be made:

1. value perceived based on price and design;
2. user satisfaction;
3. appreciation for apple products;
4. recommendation to buy the product;
5. quality design;
6. performances (battery, apps and software);
7. hardware specifications;
8. comparison with other apple products;
9. delivery and product warranty;
10. comparison with other products.

We can see that some topics we tried to explicit seem similar to each other, such as:

- 2, 3 and 4 regarding user opinions;
- 8 and 10 regarding comparison between products.

Figure 13 – Visualization of top words per topic in terms of beta values.

In fact, in this specific case study there are some issues that can be detected: the model has assigned similar words (e.g. buy/bought) or synonyms (e.g. computer/laptop/machine/macbook) to multiple topics.

After having defined the underlying topics, it could be useful to assign one topic to each document.

About this, considering the gamma matrix is ideal: it contains the proportions of words from each document that are generated from a specific topic.

By looking at the values of gamma for each review, we noted that some topics occur in equal proportion in some documents.

Therefore, it is not possible to assign a single topic to each document.

4. CONCLUSIONS

This present study provides an insight on how sentiment analysis and topic modeling can be conducted on Amazon product reviews.

The sentiment analysis, performed with two different methods, offers an evaluation of the emotional content of the reviews: normally the only way to measure the satisfaction about a product is represented by how high the average star rating is.

In this case scenario, looking at star ratings one could think that users are generally satisfied about the Apple 2020 Macbook Air, since more than 80% of reviews reported the maximum star score.

The results of our sentiment analysis are coherent in this sense, given that the majority of the reviews were classified as positive in both approaches. A graphical representation of what has been said is provided in Figure 14.

The dictionary-based approach adapts well to short texts, just like a review can be. However, note that it has some drawbacks: this methodology relies on a set of words predefined by the chosen lexicon, leading to the exclusion of words that could contribute to the text sentiment.

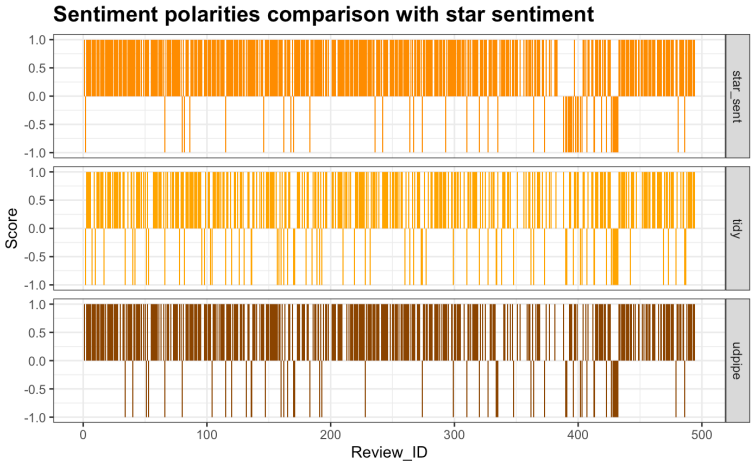

Figure 14 – Comparison between sentiment polarities scores and star sentiment scores:

- *+1 for positive reviews and star scores > 3;*
- *-1 for negative reviews and star scores ≤ 3.*

Lastly, the topic modeling has been conducted with the LDA model.

The model enables to identify the underlying topics that may drive customer sentiments and opinions. Note that the number of topics to be identified has to be specified before running the model: this could represent a concern since “the best” number of topics K does not exist.

Relatively to this research, K equal to 10 was chosen to run the analysis. However, trying to identify a subject for each topic was not so easy because LDA assigned similar words or synonyms to multiple topics, leading them to overlap.

Furthermore, explicit a single topic for each review has not been possible.

Due to these matters, it seems that LDA does not perform so well in the considered scenario. Remind that one limitation of this model is that topic coherence and meaning are evaluated through human analysis as there is not a definitive measure to assess them.