UNIVERSITY OF PISA

THE GLASGOW NORMS

Project work - Data Mining

Maria Grazia Antico
Eleonora Baracco
Davide Innocenti
Giorgia Cestaro

2021/2022

Index

1. Data Understanding

The project takes into consideration the dataset The Glasgow Norms, a collection of normative evaluations of 4500 English words, of which the psycholinguistic dimensionalities are highlighted. At the basis of this work there is a group of 1368 individuals of English birth, whose age varies from 16 to 73 years, who participated in the realization of the same, providing judgments on various words, focusing on evaluating this list of words according to 9 dimensions outline the characters, representing a substantial contribution to those academic studies, and not only, in the psycholinguistic sphere that until now had available nothing more than lexical collections particularly reduced and anything but multidimensional.

1.1 Data Semantic

The dataset is composed of 4682 records (rows) for a total of 13 independent variables (columns).
Of the thirteen variables, only one is of Categorical type, that is, the one containing the words themselves; the other 12 are numeric variables, two of which are Int64, while the remaining 10 are of Float64 type (Tab 1).
Only one of these quantitative variables is binary and that is "**Polysemy**" which indicates the possible polysemy of a word.

- **Length:** (Continuous Quantitative Variable) is the *length of* each word.

- **Arousal (AROU):** (Continuous Quantitative Variable) is a measure that establishes the level of activation of the defines levels of *arousal* and *calm*.

- **Valence (VAL):** (Continuous Quantitative Variable) is a measure that attaches importance to a word, which can be more or less *positive*.

- **Dominance (DOM):** (Continuous Quantitative Variable) is a value that indicates the feeling of control that a word provides, moving from *dominant* to more *controlled* levels.

Arousal, Valence and Dominance are used to denote the *Emotional character* associated with the words. Higher values of these three dimensions are associated with strong emotionality.

- **Concreteness (CNC):** (Continuous Quantitative Variable) represents the degree to which a word can be recognized by our senses, effectively measuring *concreteness* and *abstractness*.

- **Imageability (IMG):** (Continuous Quantitative Variable) measures the degree to which a word is difficult to imagine.

Concreteness and Imageability are two naturally related aspects that make a semantic contribution in their own right for words.

- **Familiarity (FAM):** (Continuous Quantitative Variable) represents, on a subjective level, the "familiarity" of a word. Words that are more familiar are recognized earlier than unfamiliar words.

- **Age of acquisition (AOA):** is an estimate derived from adult individuals regarding the first acquisition, written or spoken, of a word. The variable has a range of 7 points, subdivided in steps of 2, and a final period that goes from 13 years onwards. Characteristically, words acquired at an early age are recognized earlier and better than those acquired later.

- **Size:** measures the size of what the word represents in the common imagination. Larger words are more easily acquired and recognized by individuals.

- **Gender (GEND): a** value associated with the femininity (or masculinity) of a word.

- **Polysemy (POL): (**Binary Discrete Quantitative Variable) tells us if a word is polysemic or not.
- **Web_Corpus_Freq (WCF):** How much each word is present in Google's newspaper corpus.

To recap:

	LENGTH	AROU	VAL	DOM	CNC	IMAG	FAM	AOA	SIZE	GEND	POL
Variable	Quant. Cont.	Quant. Cont.	Quant. Cont.	Quant. Cont.	Quant. Cont.	Quant. Cont.	Quant. Cont.	Quant. Cont.	Quant. Cont.	Quant. Cont.	Quant. Bin.
Domain	[1,..,16]	[1,..,9]	[1,..,9]	[1,..,9]	[1,..,7]	[1,..,7]	[1,..,7]	[1,..,7]	[1,..,7]	[1,..,7]	[0,1]
Dtype	Int64	float64	float64	float64	float64	float64	float64	float64	float64	float64	Int64

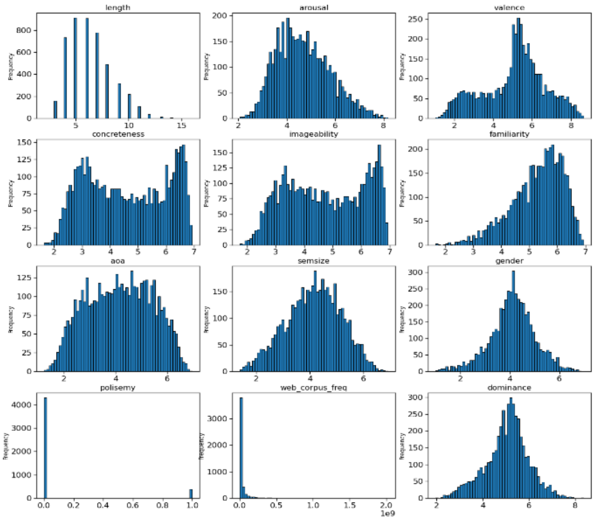
TABLE 1. DESCRIPTION OF VARIABLES

1.2 Distribution of variables and statistics

To better understand the dataset, we decided to visualize the distributions of all variables through a histogram (Fig. 1), and then focus the analysis on the classical statistical values (Fig. 2) and the skewness and kurtosis values (Fig. 3).

As you can see, in the histogram of the variables **length** and **polysemy** the rectangles are not adjacent, this is due to the fact that both are discrete variables with integer values, also **polysemy**, being binary, has only two columns, which allow us to observe with immediacy the clear predominance of non-polysemic words within the dataset.

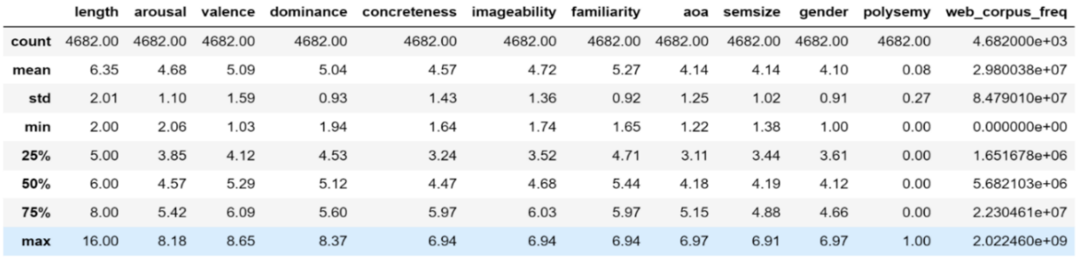
Below are statistics of the distributions, with a particular focus on measures of centrality (Tab.2).



FIGURES 1. HISTOGRAMS OF THE DISTRIBUTION OF VARIABLES

TABLE 2. CENTRALITY VALUES OF THE VARIABLES

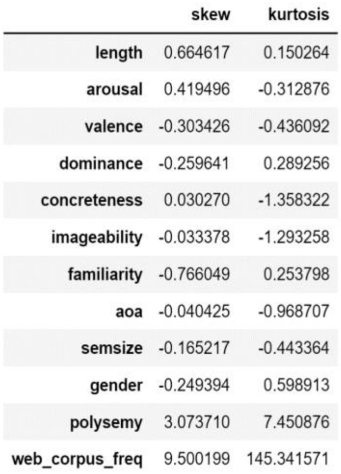
	MODE	MEDIAN	MEAN	VARIANCE
length	5	6	6.35	4.02
arousal	4.0	4.57	4.68	1.20
valence	5	5.29	5.09	2.54
dominance	5	5.12	5.04	0.86
concreteness	3	4.47	4.57	2.05
imageability	3	4.68	4.72	1.85
familiarity	6	5.44	5.27	0.85
aoa	5	4.18	4.14	1.56
semsize	4	4.18	4.14	1.04
gender	4	4.12	4.10	0.83
polysemy	0	0	0.08	0.087



	length	arousal	valence	dominance	concreteness	imageability	familiarity	aoa	semsize	gender	polysemy	web_corpus_freq
count	4682.00	4682.00	4682.00	4682.00	4682.00	4682.00	4682.00	4682.00	4682.00	4682.00	4682.00	4.682000e+03
mean	6.35	4.68	5.09	5.04	4.57	4.72	5.27	4.14	4.14	4.10	0.08	2.980038e+07
std	2.01	1.10	1.59	0.93	1.43	1.36	0.92	1.25	1.02	0.91	0.27	8.479010e+07
min	2.00	2.06	1.03	1.94	1.64	1.74	1.65	1.22	1.38	1.00	0.00	0.000000e+00
25%	5.00	3.85	4.12	4.53	3.24	3.52	4.71	3.11	3.44	3.61	0.00	1.651678e+06
50%	6.00	4.57	5.29	5.12	4.47	4.68	5.44	4.18	4.19	4.12	0.00	5.682103e+06
75%	8.00	5.42	6.09	5.60	5.97	6.03	5.97	5.15	4.88	4.66	0.00	2.230461e+07
max	16.00	8.18	8.65	8.37	6.94	6.94	6.94	6.97	6.91	6.97	1.00	2.022460e+09

FIGURE 2. ANALYSIS OF THE STATISTICS OF THE VARIABLES

The centrality values immediately confirm that none of the distributions is properly normal (mean, mode and median should coincide) but we can easily identify those distributions that are closest to the condition of normality: **valence**, **dominance**, **semsize** and **gender.** In Table 2 the variable polysemy is also presented to underline how these values are significant only for certain types of variables and distributions. This variable, in fact, has practically identical central values, despite not being a normal distribution. Looking at the graphical representations, however, these ambiguities are immediately removed by immediately noting the shape of the distribution through the histogram, and it is instead possible to easily grasp where the mean and median are positioned with respect to the distribution through the boxplot (presented next).



	skew	kurtosis
length	0.664617	0.150264
arousal	0.419496	-0.312876
valence	-0.303426	-0.436092
dominance	-0.259641	0.289256
concreteness	0.030270	-1.358322
imageability	-0.033378	-1.293258
familiarity	-0.766049	0.253798
aoa	-0.040425	-0.968707
semsize	-0.165217	-0.443364
gender	-0.249394	0.598913
polysemy	3.073710	7.450876
web_corpus_freq	9.500199	145.341571

FIGURE 3. SKEWNESS AND KURTOSIS VALUES

It was decided to consider the **skewness** and **kurtosis** values valid since all distributions except **web_corpus_freq** are single-mode. **Skewness** is a measure of symmetry, or more precisely, the lack of symmetry. A distribution is symmetric if it looks the same to the left and right of the center point. Negative values represent a left-shifted curve, positive values a right-shifted curve, and values close to zero a normal distribution.

The **kurtosis**, on the other hand, measures the departure of a curve from distributive normality, indicating whether the values are uniformly distributed or are concentrated around the mean value. Again, normality is represented by values close to zero. Negative values indicate a flatter curve and positive values a sharper one.

The lowest kurtosis values are found in the variables **concreteness**, **imageability** and **aoa.** Observing the statistics and the histograms of the first two variables we can state that these are the distributions that deviate the most from normality (obviously neglecting polysemy and web_corpus_freq). In fact, the mode of both deviates significantly from the respective mean and median. From the histogram we also note that both have local maxima around the mode. This happens because the histogram is a graphical representation of a distribution in frequency classes and the modal interval (class that with the same amplitude has the highest frequency density) of the distributions under consideration does not coincide with the frequency class in which is the mode. This consideration also leads to evaluate as non-significant the values of skewness for *concreteness* and *imageability* that are very close to zero but opposite, despite their graphical form is similar.

The gender variable, on the other hand, has the highest kurtosis value, and we note in fact a distribution quite concentrated around the mean values, with thin tails, furthermore the skewness value indicates that the left tail is longer (a tendency observed here but confirmed with an analysis of outliers later) and consequently we note a distribution slightly shifted to the right.

The **length** distribution is clearly shifted to the left, but it is the curve that presents the least deviation from normality according to the kurtosis value. It is hereafter represented with a histogram that besides showing the distribution (with adjacent rectangles) also shows how many words of a certain length are polysemic (Fig. 4).

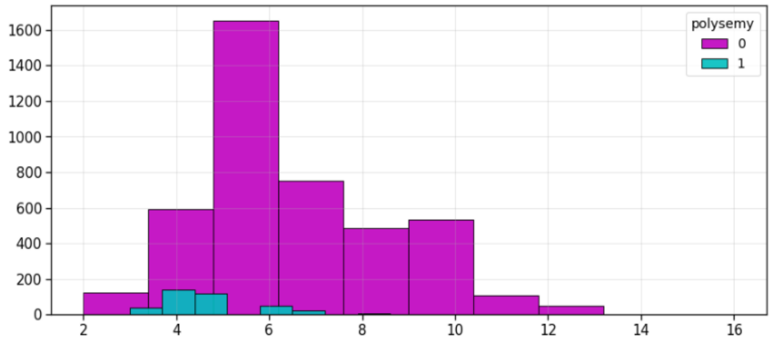


FIGURE 4. DISTRIBUTION OF 'LENGTH' WITH SUBDIVISION OF POLYSEMIC AND NON-POLYSEMIC WORDS.

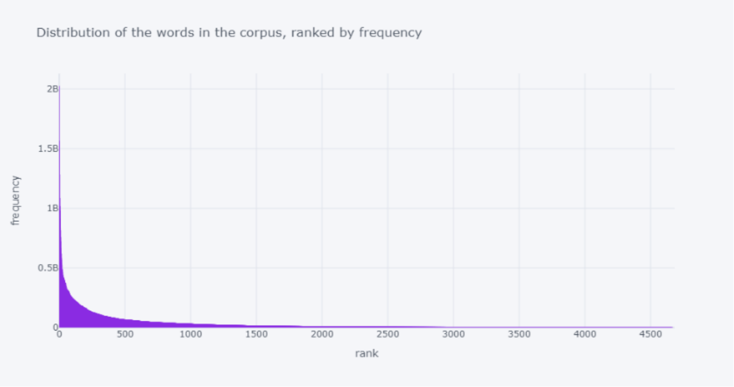


FIGURE 5. FREQUENCY DISTRIBUTION OF WEB_CORPUS_FREQ

The variable **web_corpus_freq**, instead, deserves a closer look. The curve shown in the histogram is certainly not a normal curve, but it shows an interesting feature of natural language, namely the existence of many words that have more or less the same frequency (not very high) and a few words that instead have a very high frequency (Fig. 5). It was decided to better analyze this tendency since the dataset is composed of chosen words and not actual text, so no grammatical, semantically empty words that are widely used to articulate speech appear. Despite this, the Zipf's behavior of the language still turns out to be respected. We present below, therefore, a bar graph (Fig. 6) showing how the frequency of a word is inversely proportional to its rank (position occupied by a word in a descending frequency order).

Zipf's law is often represented in double logarithmic scale, the equation is then transformed into the equation of a straight line. Through the logarithmic representation it is possible to notice that the law is not completely respected by the words of the dataset (as the rank increases the frequency should decrease more and more slowly). The dataset, moreover, being composed by chosen words, is not representative of the English language as a natural language; therefore, it is correct to notice discrepancies with the theoretical trend.

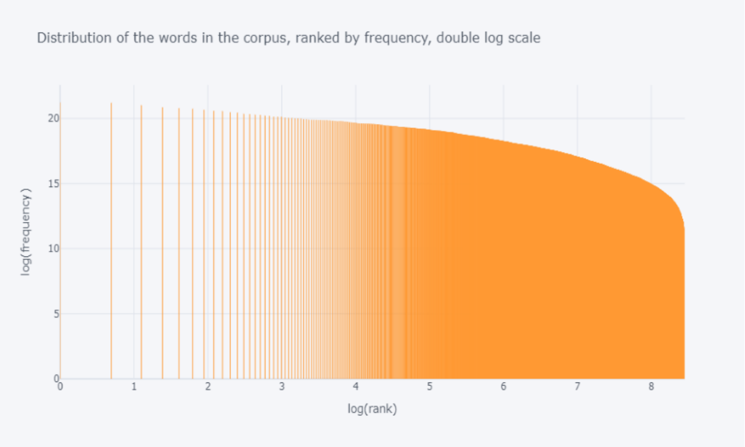


FIGURE 6. FREQUENCY DISTRIBUTION OF WEB_CORPUS_FREQ IN LOGARITHMIC SCALE

1.3 Transformation of Variables

In light of what has been said in the previous paragraph, we can immediately state that the variable *polysemy* should not be transformed into a logarithmic variable. Moreover, the variable *web_corpus_freq* is the most suitable for this type of transformation. However, it was decided to make an organic discourse on all variables.

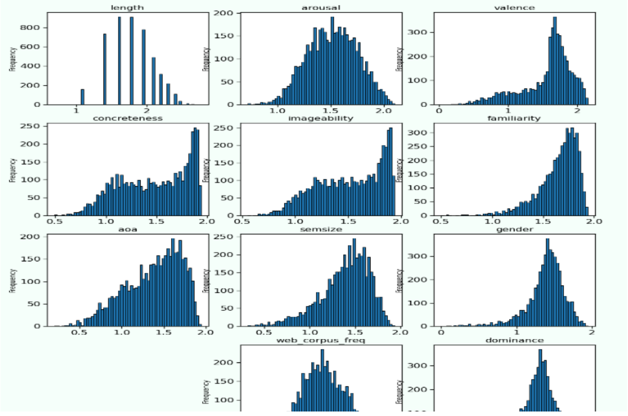
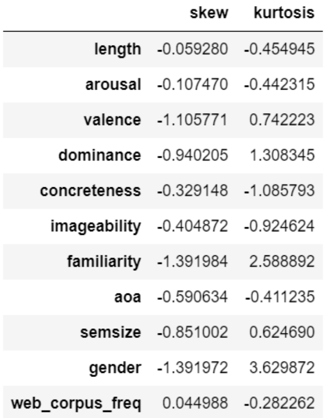


FIGURE 8. LOGARITHMIC DISTRIBUTIONS OF THE VARIABLES



	skew	kurtosis
length	-0.059280	-0.454945
arousal	-0.107470	-0.442315
valence	-1.105771	0.742223
dominance	-0.940205	1.308345
concreteness	-0.329148	-1.085793
imageability	-0.404872	-0.924624
familiarity	-1.391984	2.588892
aoa	-0.590634	-0.411235
semsize	-0.851002	0.624690
gender	-1.391972	3.629872
web_corpus_freq	0.044988	-0.282262

FIGURE 7. SKEWNESS AND KURTOSIS VALUES OF LOGARITHMIC VARIABLES

The graphical representation confirms that **web_corpus_freq** is the variable which undergoes the greatest transformation in terms of its distribution, which is closer to a normal distribution. It can be seen that the **imageability** and **concreteness** variables no longer have a local maximum. In addition, almost all the distributions are now more shifted to the right with long tails on the left, going to extremes a trend already noted previously. The skewness values calculated on the logarithmic variables confirm what has just been said since, from a comparison with the values of the non-transformed variables, we note a general negativisation.

The kurtosis values, on the other hand, become more extreme for those variables that, presented in a non-logarithmic form, already had a positive value, i.e. **gender**, **familiarity** and **dominance**. An opposite but less evident tendency (therefore in a direction of normalization) is identified in those variables that presented a negative kurtosis, in particular **aoa.** The measures of centrality, on the other hand, confirm that the variables **concreteness** and **imageability** lose their local maximum around the mode, as this value is now decidedly closer to the mean and median (Tab.3).

	MODE	MEDIAN	MEDIA
length	1.61	1.79	1.80
arousal	1.38	1.51	1.52
valence	1.61	1.66	1.57
dominance	1.61	1.63	1.60
concreteness	1.09	1.49	1.47
imageability	1.09	1.54	1.51
familiarity	1.79	1.69	1.64
aoa	1.61	1.42	1.37
semsize	1.38	1.43	1.39
gender	1.38	1.41	1.38
web_corpus	not unimodal	15.55	15.62

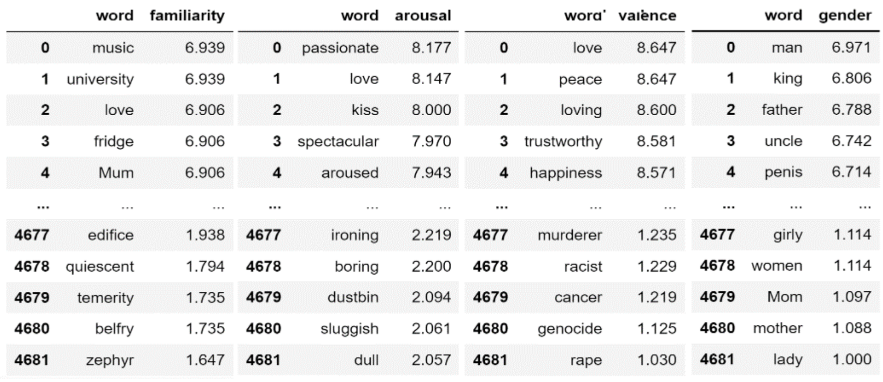
TABLE 3. CENTRALITY VALUES OF LOGARITHMIC VARIABLES.

1.4 Data Quality (missing values, outliers)

In the aforementioned analysis, we focused on two types of dimensions inherent in the quality of the data available: *Accuracy* and *Completeness*. With Accuracy we mean how much a given piece of information actually reflects the reality of the facts, and to measure how accurate the dataset really was, the analysis was directed towards the search for semantic inconsistencies or easily detectable errors.

In fact, we immediately verified that each word contained in the dataset really had the length reported by the attribute "**length**" by creating a list containing each value of word length and comparing it, through the function *.equals(),* with the values of the variable "**length**", obtaining a positive result.

After that, a sample of words was compared with each attribute in the dataset, looking for the presence of any errors, but finding that there were none. An example of what has just been discussed is shown in the table alongside (Tab. 4).



	word	familiarity		word	arousal		word	valence		word	gender
0	music	6.939	**0**	passionate	8.177	**0**	love	8.647	**0**	man	6.971
1	university	6.939	**1**	love	8.147	**1**	peace	8.647	**1**	king	6.806
2	love	6.906	**2**	kiss	8.000	**2**	loving	8.600	**2**	father	6.788
3	fridge	6.906	**3**	spectacular	7.970	**3**	trustworthy	8.581	**3**	uncle	6.742
4	Mum	6.906	**4**	aroused	7.943	**4**	happiness	8.571	**4**	penis	6.714
...	...	...	**...**	...	...	**...**	...	...	**...**	...	...
4677	edifice	1.938	**4677**	ironing	2.219	**4677**	murderer	1.235	**4677**	girly	1.114
4678	quiescent	1.794	**4678**	boring	2.200	**4678**	racist	1.229	**4678**	women	1.114
4679	temerity	1.735	**4679**	dustbin	2.094	**4679**	cancer	1.219	**4679**	Mom	1.097
4680	belfry	1.735	**4680**	sluggish	2.061	**4680**	genocide	1.125	**4680**	mother	1.088
4681	zephyr	1.647	**4681**	dull	2.057	**4681**	rape	1.030	**4681**	lady	1.000

TABLE 4. COMPARISON OF WORDS WITH ACTUAL PAID VALUES OF ATTRIBUTES.

Subsequently we checked how the dataset behaved in relation to any missing values within the attributes, recording as the only variable containing missing values "**Web_Corpus_Freq**" which, according to what is reported by the above analysis, shows a total of 14 missing values. It is therefore possible to state with certainty that this data collection appears to be complete, reporting as a percentage of missing values a meager 0.30%, in only one attribute. In order to manage these values, we thought that words with null values are simply not attested in the corpus, so we did not opt for any particular technique except to replace the "**Null**" values with **0**. This procedure is to be avoided in case we want to use the variable in a logarithmic version because when the variable is zero its logarithm will tend to - ∞.

With regard to the outliers, once again using the *.figure()* function, the BoxPlots of each attribute were graphically represented, in search of values highly different from the first and third quartiles (Fig.9).

It is shown, moreover, the BoxPlot of the variable "**Web_Corpus_Freq**" transformed in logarithmic scale, resuming the speech of the previous paragraph, and with missing values already set to 0 (Fig.10).

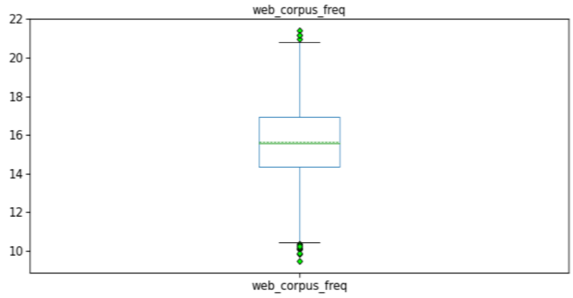


FIGURE 9. BOXPLOT OF LOGARITHMIC WEB_CORPUS_FREQ

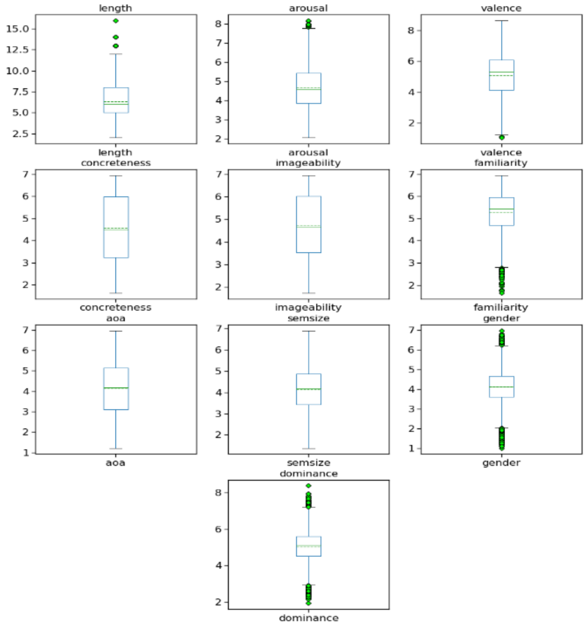


FIGURE 10. BOXPLOT OF LOGARITHMIC VARIABLES

As you can see from the images not all attributes contain Outliers. More specifically:

Aoa	0
Arousal	11
Concreteness	0
Dominance	133
Familiarity	55
Gender	164
Imageability	0
Length	16
Polysemy	379
Semsize	0
Valence	2
W_C_F	610
Word	0

Not considering "*Polysemy*", as dichotomous, and "*web_corpus_freq*" we decided to keep at least temporarily the outliers within the dataset because it is possible to note that there are hardly any single cases clearly distant from the Quartiles, and, even if there were, they would still have an important meaning in the context of this work representing in fact a word within the corpus, with all the values assigned to it. Nevertheless, it was decided to test what would happen by removing such values from the various attributes and one particular case that caught our attention concerns the variable "*Gender"*.In this case, in fact, as shown in the graph opposite (Fig.11), by eliminating the values far from the central ones, we would cut out **70%** of words considered feminine against a meager **30%** of words that are instead masculine.

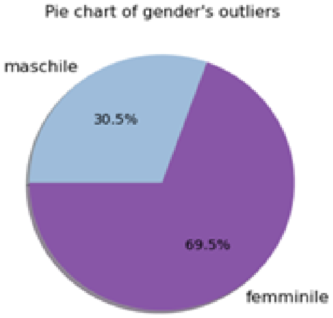


FIGURE 11. GENDER VARIABLE WITH POSSIBLE ELIMINATION OF OUTLIERS.

All this, in relation to the English language which has, as its own characteristic, that of remaining as neutral as possible, and which, at least unlike other vocabularies, presents however a discrete Bias in favor of masculine terms, cutting out the most feminine terms (<=2) from the frequencies, while also maintaining the most masculine (>= 6), as can be seen on the side in Figure 1.2

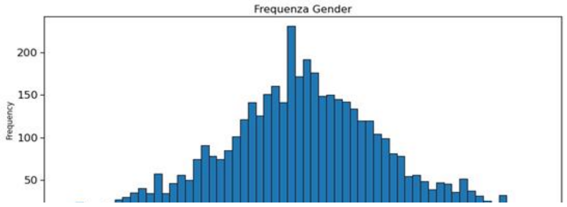


FIGURE 12. HISTOGRAM OF GENDER WITH ELIMINATION OF OUTLIERS.

1.5 Correlations between variables

Figure 13 shows a heatmap of the existing correlations between the quantitative variables of the Dataset using, as a method, the Pearson correlation. It is possible to note the presence of some highly correlated variables such as '*valence*' and '*dominance*', '*aoa*' and '*length*', '*arousal*' and '*semsize*' and others characterized by a low correlation, including '*imageability*' and '*aoa*', 'semsize' and '*concreteness*', '*valence*' and '*gender*'. In particular, as explained in the first paragraph, the variables *'arousal', 'dominance' and 'valence'* are common in that they describe the emotional impact of a word, in fact there is a strong correlation between them (Fig.14). On the other hand, in the case of the relationships between personal characteristics of a word (the subset of variables *'aoa', 'familiarity', 'imageability'* and '*concreteness*'), a positive correlation can also be noted here, particularly between *imageability* and *concreteness* with the difference, in this case, of the presence of some negative correlations, including *'familiarity'* and '*aoa*' (Fig.13).

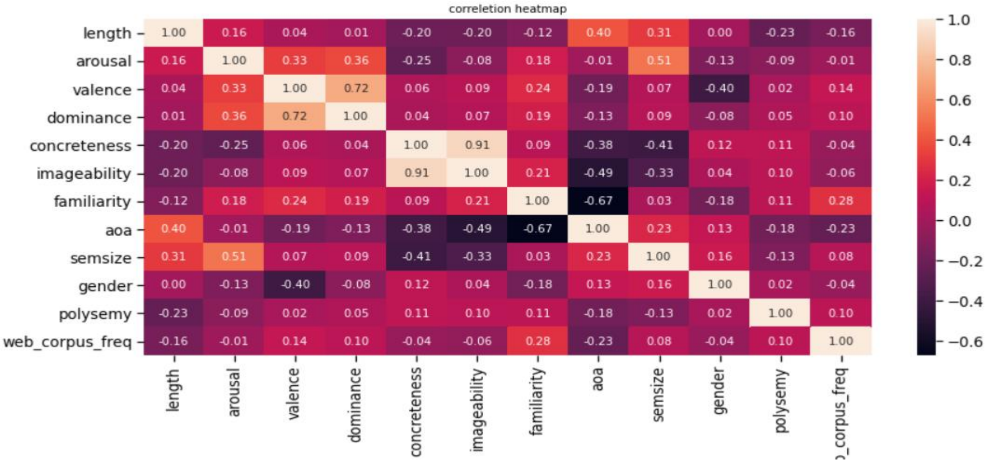


FIGURE 15. HEATMAP OF CORRELATIONS BETWEEN VARIABLES.

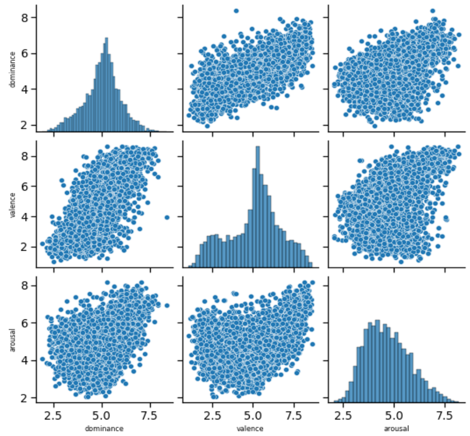


FIGURE 14. CORRELATIONS AMONG VARIABLES DESCRIBING EMOTIONAL IMPACT OF A WORD.

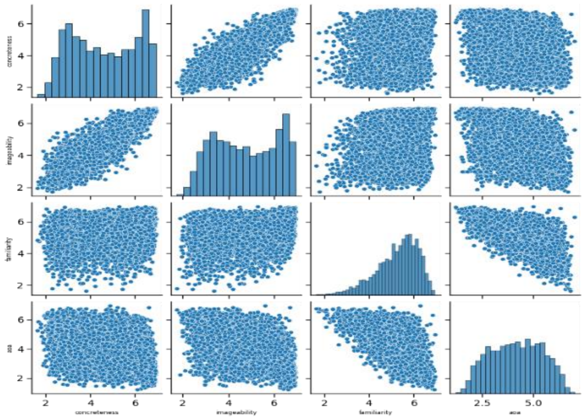


FIGURE 13. CORRELATIONS AMONG VARIABLES DESCRIBING PERSONAL CHARACTERISTICS OF A WORD.

In the following subsections we will analyze in depth some pairs of variables considered relevant.

1.5.1 Valence and Dominance

In Figure 16, in addition to presenting the scatter plot of the two variables, we wanted to analyze this plot in relation to the variable 'gender'.

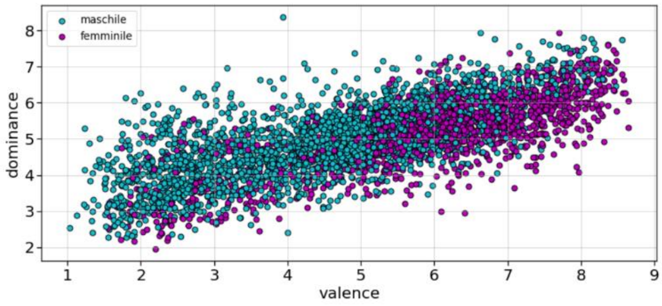


FIGURE 16. SCATTERPLOT OF VALENCE AND DOMINANCE IN RELATION TO 'GENDER' DIVIDED INTO 2 CATEGORIES.

The denotation of the legend was done taking into consideration the following values:

- Values of the 'gender' variable from 1 to 3 denote female behavior.
- Values of the 'gender' variable from 4 to 6 denote male behavior.

From Fig.16 it is possible to notice how, with a greater value assigned to a word and a greater perceived control of this word, the words denoting a female behavior are present in the majority, unlike, instead, the words denoted by a male behavior that are characterized by a lower value and a lower perceived control.

Since, however, the aforementioned division of values was considered 'inflexible', we tried to divide the values of 'gender' into 3 categories, introducing the category "neutral" so that we could include the intermediate values of 'gender' (3 and 4), leaving the extreme values denoted strictly by a female (1 and 2) and male (5,6) character.

From the graph it is possible to note how the intermediate values, i.e. those considered in the 'neutral' category, are more characterized by those denoting male behavior, insofar as the area characterized by female behavior remains almost unchanged, unlike the male area, which is "dominated" by the 'neutral' area (Fig 17).

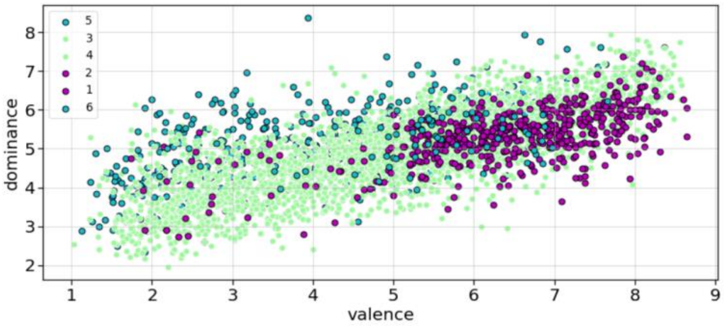
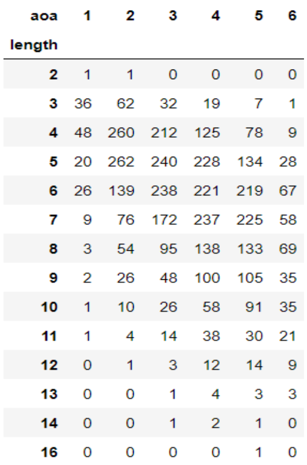


FIGURE 17. SCATTERPLOT OF VALENCE AND DOMINANCE IN RELATION TO 'GENDER' DIVIDED INTO 3 CATEGORIES.

1.5.2 Aoa and length

To represent the correlation between *'aoa'* and '*length*', a violinplot was chosen:



aoa length	1	2	3	4	5	6
2	1	1	0	0	0	0
3	36	62	32	19	7	1
4	48	260	212	125	78	9
5	20	262	240	228	134	28
6	26	139	238	221	219	67
7	9	76	172	237	225	58
8	3	54	95	138	133	69
9	2	26	48	100	105	35
10	1	10	26	58	91	35
11	1	4	14	38	30	21
12	0	1	3	12	14	9
13	0	0	1	4	3	3
14	0	0	1	2	1	0
16	0	0	0	0	1	0

TABLE 5. CROSSTAB OF AOA AND LENGTH

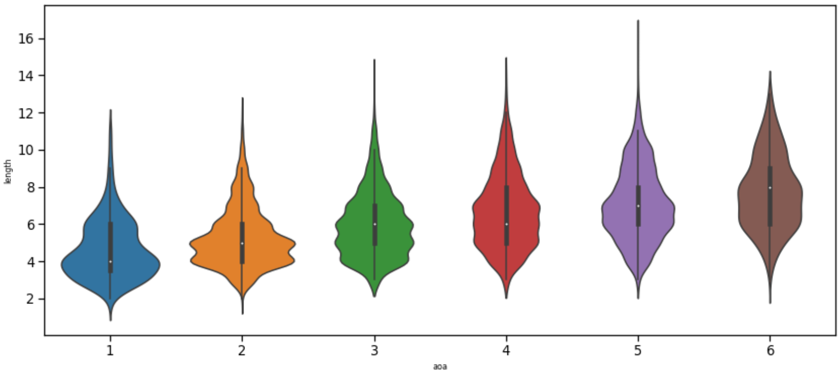


FIGURE 18. VIOLINPLOT OF AOA AND LENGTH.

From Fig. it is 18possible to highlight an increasing order of the distributions, which denotes that as the years in which a word is acquired increase, so does the length of that word. In addition, to highlight this relationship, a crosstab (Tab. 5) of the two variables was created from which an iplot of 'length' bars was subsequently derived, subdividing the color bars according to age (Fig. 19). Observing the values of the crosstab and those represented in the graph, it can be seen that the bars denoted by years 1,2 and 3 are increasing up to words with length 5 or so, and thus it is assumed that in the first years of age words with shorter lengths are mainly acquired, and are instead in turn decreasing as the number of letters in a word increases, i.e., longer words are acquired more at older ages, particularly after the age of 3.

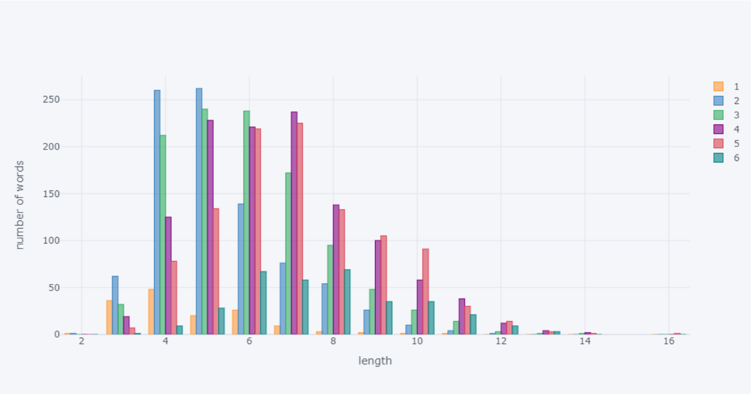


FIGURE 19. BAR GRAPH OF LENGTH IN RELATION TO AOA, CREATED WITH PLOTLY.

1.5.3 Imageability and aoa

There is a negative correlation between the above variables. Also here, in order to have a greater clarity of the relation, a crosstab was created (Tab.6) and represented a bar iplot of imageability subdividing, also here, the bars in different colors according to the age (Fig. 20).

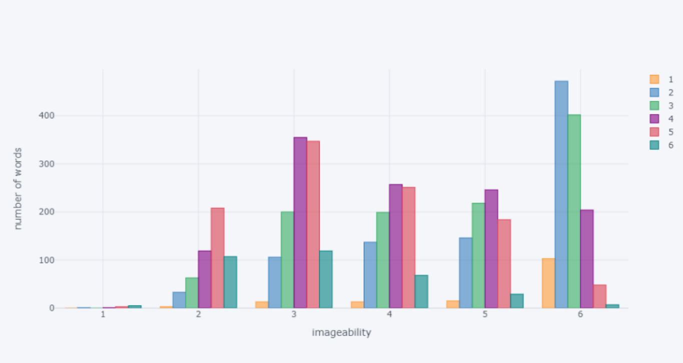
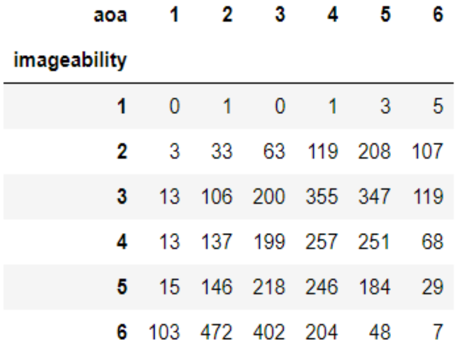


FIGURE 20. BAR GRAPH OF IMAGEABILITY IN RELATION TO AOA, CREATED WITH PLOTLY.



aoa	1	2	3	4	5	6
imageability						
1	0	1	0	1	3	5
2	3	33	63	119	208	107
3	13	106	200	355	347	119
4	13	137	199	257	251	68
5	15	146	218	246	184	29
6	103	472	402	204	48	7

TABLE 6. CROSSTAB BETWEEN IMAGEABILITY AND AOA

Words denoted by a high level of imaginability are those that are acquired in the early years of age and, on the contrary, words characterized by low imaginability are acquired mainly in an older age.

1.5.4 Aoa and familiarity

Aoa and *familiarity*, like the previous pair of variables, are also negatively correlated. To represent their relationship, the following scatter plot was used (Fig. 21):

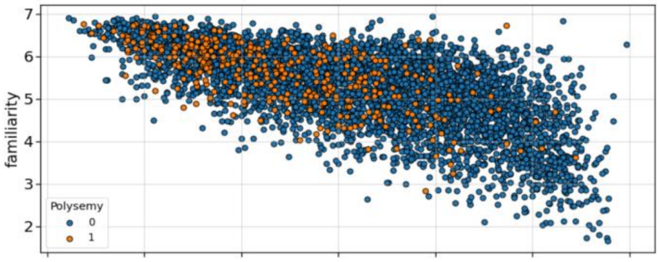


FIGURE 21. SCATTER PLOT OF AOA AND FAMILIARITY BASED ON POLYSEMY.

In Fig.21 words were classified according to '*polysemy*'. Words denoted by strong familiarity and low age of acquisition are mainly composed of polysemic words. As age increases and the relative decrease in familiarity caused by negative correlation, mainly non-polysemic words are found.

1.5.5 Imageability and Concreteness

This pair is the most correlated in our Dataset (correlation 0.91). As in the case of valence and dominance, we wanted to represent a scatter plot by classifying words according to gender:

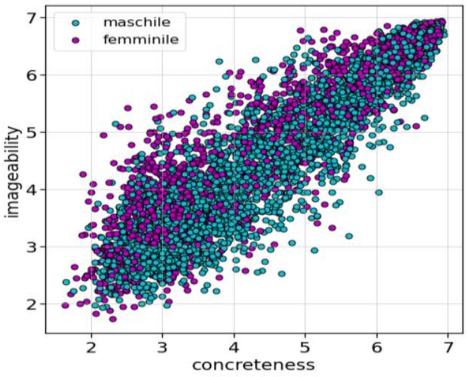


FIGURE 23. SCATTER PLOT BETWEEN IMAGEABILITY AND CONCRETENESS WITH GENDER DIVIDED INTO 2 CATEGORIES.

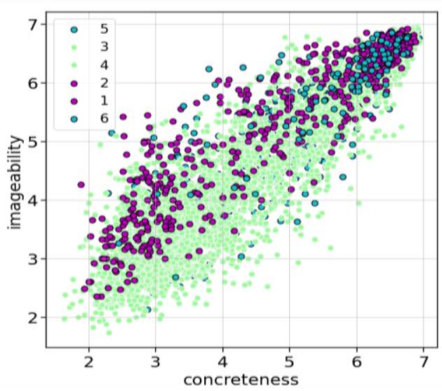


FIGURE 22. SCATTER PLOT BETWEEN IMAGEABILITY AND CONCRETENESS WITH GENDER DIVIDED INTO 3 CATEGORIES.

From Fig. 23, with the subdivision of the two genders, it can be seen that as the level of imaginability of a word increases, the concreteness of that image also increases and, in particular, those with less imaginability are associated with male behavior and those with slightly more imaginability are those associated with female behavior. In the case of dividing 'gender' into three categories (Fig. 22), as seen in the previous graphs, the 'neutral' category significantly overpowers the male category except for a group of words characterized by high concreteness and imaginability indices and having a male denotation.

2. Clustering

To perform clustering, it was decided to reduce the dataset. The variables *polysemy, web_corpus_freq, length, arousal* and *familiarity were* simply excluded from the analysis. The variables linked by a high linear correlation were merged through the average of their values, thus obtaining two new variables: the **visual** variable from imageability and concreteness, the **emotion** variable from dominance to valence. The dataset for clustering is therefore composed of these newly created variables and the original *gender* and *aoa* variables. In addition, it was decided to eliminate outliers to perform clustering with the kmeans and hierarchical algorithms. For the dbscan algorithm it was decided to evaluate the difference between a clean dataset and one with outliers.

2.1 K-Means

With the K-Means algorithm, using the minmax method as normalization, the possible alternatives of K values were compared by setting 'k-means++' as the method to find the initial values of centroids and, as values of n_init and max_iter those set by default, i.e. 10 and 100. For each K, the following SSE and Silhouette values were identified. It was considered that the "best" result was with K=8. This choice is derived from the observation of the following graph (Fig.24), which shows that the optimal number of K values is between 7 and 9, i.e. the area where there is the "elbow" of the graph.

K	SSE	SILHOUETTE
5	361	0.242
6	327	0.233
7	299	0.221
8	277	0.225
9	259	0.222
10	246	0.214

TABLE 7. VALUES OF SSE AND SILHOUETTE FOR EACH VALUE OF K

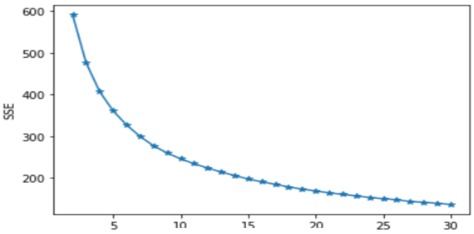


FIGURE 24. ELBOW METHOD

Analyzing, therefore, these values in Table 7 above, we see a slight increase in Silhouette with K=8 (0.225) compared to the adjacent K values (0.221 and 0.222). Assuming, therefore, as an optimal value K=8, it is possible to

it is wanted to represent the various relations of the present variable in our dataframe through one scatterplot.

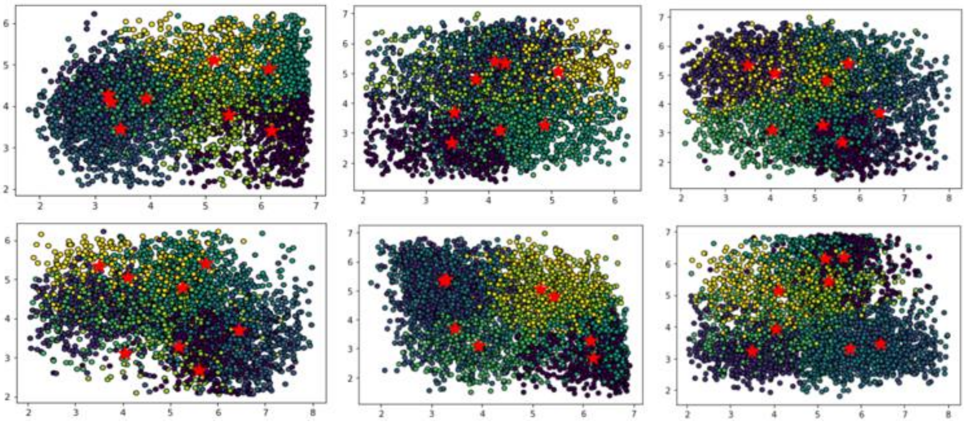


FIGURE 25. SCATTERPLOT OF VARIABLES WITH DIVISION INTO 8 CLUSTERS.

As can be seen, many centroids are close to each other and this can also be seen by observation of the following parallel coordinate graph (PCP). Given their proximity, we hypothesize the possibility of merging some clusters (Fig.26).

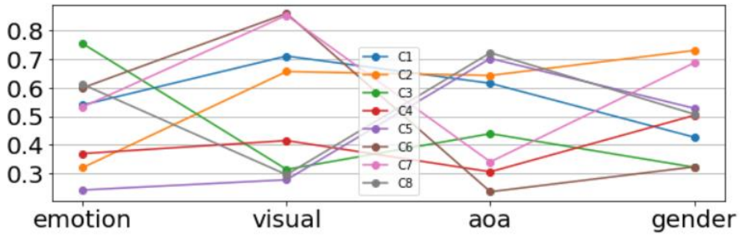
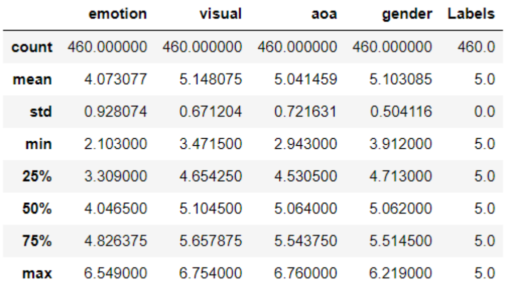


FIGURE 26. PLOT PARALLEL COORDINATES

Each cluster is, in addition, composed of the following number of words:

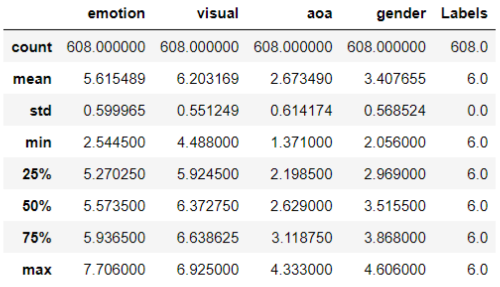
```
        (array([0,  1,  2,  3,  4,  5,  6,  7]),
array([675, 516, 461, 573, 642, 469, 607, 564]
```

In this way, each cluster contains a set of words with different characteristics. In cluster number 5, for example, we notice that the average of the variables of *emotion, visual* and *aoa* is quite high; deducing therefore that, in this cluster, there are words denoted by male behavior acquired at an average age of 5 years and characterized by a high visual possibility of that word (Fig.28). In cluster number 6, instead, there are words denoted by an almost female behavior acquired in a low age and characterized by very high emotional and visual levels (Fig.27).



	emotion	visual	aoa	gender	Labels
count	460.000000	460.000000	460.000000	460.000000	460.0
mean	4.073077	5.148075	5.041459	5.103085	5.0
std	0.928074	0.671204	0.721631	0.504116	0.0
min	2.103000	3.471500	2.943000	3.912000	5.0
25%	3.309000	4.654250	4.530500	4.713000	5.0
50%	4.046500	5.104500	5.064000	5.062000	5.0
75%	4.826375	5.657875	5.543750	5.514500	5.0
max	6.549000	6.754000	6.760000	6.219000	5.0

FIGURE 28. PRINCIPAL STATISTICS OF VARIABLES IN CLUSTER 5.



	emotion	visual	aoa	gender	Labels
count	608.000000	608.000000	608.000000	608.000000	608.0
mean	5.615489	6.203169	2.673490	3.407655	6.0
std	0.599965	0.551249	0.614174	0.568524	0.0
min	2.544500	4.488000	1.371000	2.056000	6.0
25%	5.270250	5.924500	2.198500	2.969000	6.0
50%	5.573500	6.372750	2.629000	3.515500	6.0
75%	5.936500	6.638625	3.118750	3.868000	6.0
max	7.706000	6.925000	4.333000	4.606000	6.0

FIGURE 27. PRINCIPAL STATISTICS OF VARIABLES IN CLUSTER 6.

2.2 Hierarchical clustering

Similar to previous analyses, the variables "*emotion*", "*visual*", "*aoa*" and "*gender*" were used, applying a Euclidean distance function for the points. The values are always normalized and the variables do not contain

Outliers. The most common approach for hierarchical clustering is one that groups clusters based on their similarities and it is called "Agglomerative", and has been used for all subsequent hierarchical graphs. Three methods were applied: Full, Single and Average and the following dendograms show the last p=30 clusters on the total 4658 elements of the dataset.

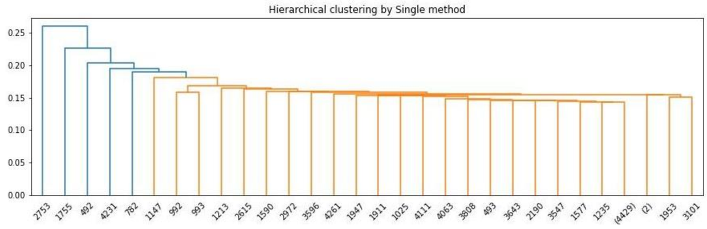


FIGURE 29. DENDOGRAM WITH SINGLE METHOD

The first method used is the Single method. As can be easily guessed, this is not a very appropriate method for the structure of the dataset. The number of colors within the representation was automatically generated by the algorithm (Fig.29).

The second method is of the Average type (Fig.30). Unlike the previous one it appears much more appropriate to our values and we decided to represent the same number of clusters previously set as optimal in Kmeans algorithms, equal to 8 by setting color_treshold = 0.524.

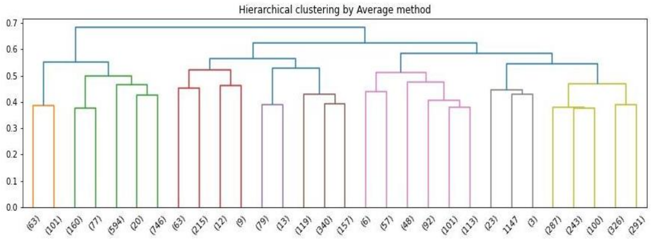


FIGURE 30. DENDOGRAM WITH METOOD AVERAGE

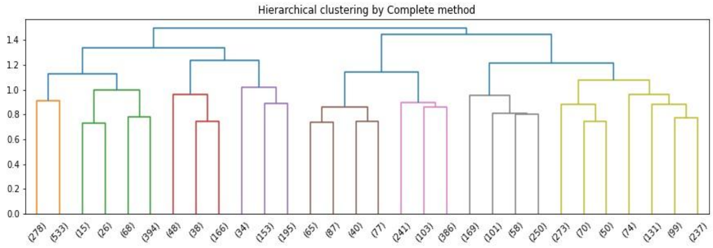


FIGURE 31. DENDOGRAM WITH COMPLETE METHOD.

The last of the three is the Complete method. As well as in the hierarchical Average 8 clusters were highlighted by setting color_treshold = 1.1 and the graph appears well delineated and agglomerated.

The following values are achieved with the above settings:

- Siluette Complete: 0   .14122441149270284
- Siluette Single: -0.10611988268766
- Siluette Average: 0   .12227175130419715

And, for a more complete analysis, the various methods were analyzed in more detail, resulting in the following counts within the various clusters:

- Complete array: ([934, 382, 503, 811, 252, 578, 269, 730])
- Single array: ([4451, 2, 1, 1, 1, 1, 1, 1])
- Average array: ([299, 417, 1597, 616, 1247, 164, 27, 92])

This analysis was useful to see the composition of each cluster for each method and in fact, as it immediately jumps to the eye, the "Single" method generates highly inhomogeneous groups, going to group all the data in a single cluster.

For the same reasoning, both Complete and Average appear well structured and finally between the two we preferred the first one that obtains a Silhouette value, although marginal, however higher, recording 0.141.

2.3 DBSCAN

DBSCAN is a density-based algorithm and splits the elements of the dataset into center points, border points, and noisy points. The noisy points identified by the algorithm can be seen as outliers; therefore, we decided to try clustering both with a dataset with outliers and with a previously cleaned one (for this operation we based ourselves on the definition for which an outlier is such if greater than the third quartile or less than the first plus 1.5*IQR). In order to better understand the behavior of outliers, normalization was performed with both Min Max Scaler and Robust Scaler.

2.3.1 Dataset with outliers

In order to identify the optimal values of minimum number of neighbors and EPS, the graphs obtained by ordering the points according to their distance from the k-th neighbor were evaluated, which informs on the best value of EPS for a predetermined minimum number of neighbors. Therefore, the silhouette values of the various combinations were observed and the one with the largest silhouette but which could detect some outliers was chosen. The optimal minimum number of neighbors was found to be 8, combined with an EPS of 0.19 for the minmax (Fig.32) and 0.55 for the robust scaler.

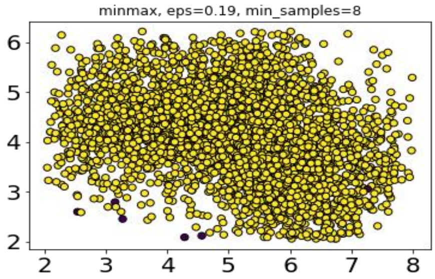


FIGURE 32. DBSCAN WITH MINMAX, EPS=0.19

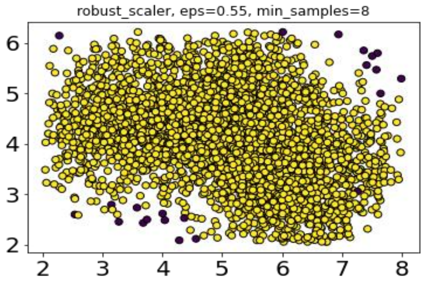


FIGURE 33. DBSCAN WITH ROBUST SCALER, EPS=0.55.

The silhouettes are 0.213 for the minmax and 0.295 for the robust scaler, respectively. It can be seen immediately that there are more outliers detected by the robust scaler. The minmax, on the other hand, tends to color everything the same color.

2.3.2 Clean Dataset

After cleaning the dataset using the classical definition of outliers, we can see that the boxplot still identifies some outliers.

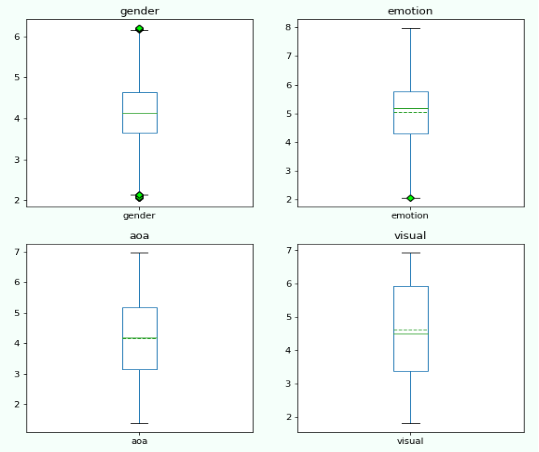


FIGURE 34. BOXPLOT OF VARIABLES AFTER ELIMINATION OF OUTLIERS.

We again wanted to observe the behavior of the algorithm used on data normalized with both minmax (Fig. 36) and robust scaler (Fig.35), precisely in order to evaluate the detection of anomalous points.

We observe that for the same EPS and minimum number of neighbors, the algorithm used on data normalized with minmax has a higher silhouette of 0.220, but with fewer noisy points. While on data normalized with robust scaler the silhouette decreases to 0.277.

Therefore, the best result is obtained by keeping the outliers and using the robust scaler to normalize the data.

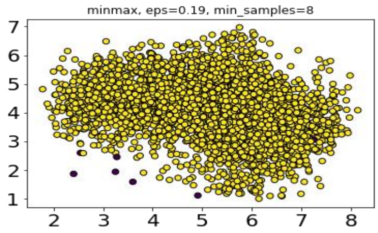


FIGURE 36. DBSCAN WITH MINMAX, AFTER ELIMINATION OF OUTLIERS.

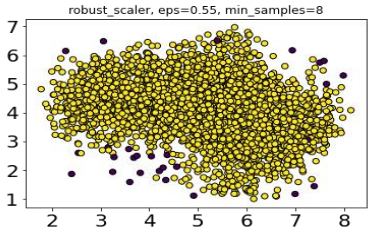


FIGURE 35. DBSCAN WITH ROBUST SCALER, AFTER OUTLIERS ELIMINATION.

2.3.3 Conclusions

Having eliminated, in the previous phase, the outliers present in the variables, and since the distribution appears to be quite dense and unified, DBSCAN seems not to work well on our Dataset. Comparing, moreover, the values of SSE and Silhouette of the three algorithms, it was decided to choose the K-means algorithm as the best, as it appears to have a higher Silhouette (0.226), although this is not a value considered 'optimal'. As mentioned in section 2.1, in the case of K-means, given the proximity of the centroids, a merger between the clusters is assumed. The best solution that could be created, even at a visual level, observing the following plot between *'visual'* and *'aoa'*, would be the division into two clusters.

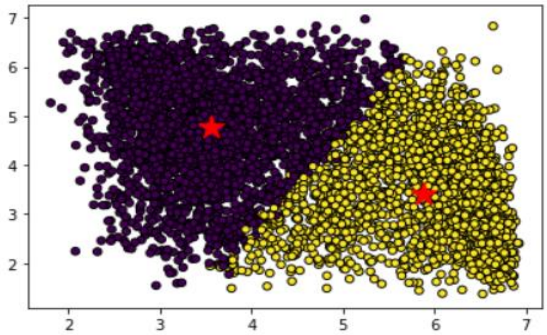
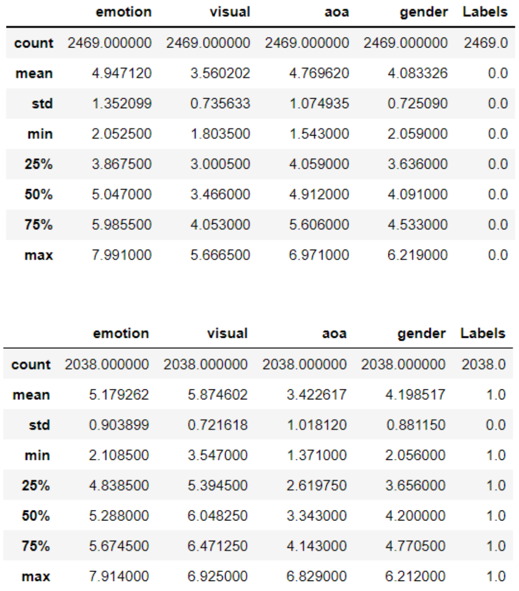


FIGURE 37. PLOT OF VISUAL AND AOA WITH K-MEANS METHOD AND K=2.

With this hypothetical clustering, the following clusters would be formed:

```
(labels {0: 2038, 1: 2469}
sse 590.8890855650861
silhouette 0.302004587161052
```

obtaining, although an increase in the Silhouette compared to K=8, an increase also notable in the SSE.



	emotion	visual	aoa	gender	Labels
count	2469.000000	2469.000000	2469.000000	2469.000000	2469.0
mean	4.947120	3.560202	4.769620	4.083326	0.0
std	1.352099	0.735633	1.074935	0.725090	0.0
min	2.052500	1.803500	1.543000	2.059000	0.0
25%	3.867500	3.000500	4.059000	3.636000	0.0
50%	5.047000	3.466000	4.912000	4.091000	0.0
75%	5.985500	4.053000	5.606000	4.533000	0.0
max	7.991000	5.666500	6.971000	6.219000	0.0

	emotion	visual	aoa	gender	Labels
count	2038.000000	2038.000000	2038.000000	2038.000000	2038.0
mean	5.179262	5.874602	3.422617	4.198517	1.0
std	0.903899	0.721618	1.018120	0.881150	0.0
min	2.108500	3.547000	1.371000	2.056000	1.0
25%	4.838500	5.394500	2.619750	3.656000	1.0
50%	5.288000	6.048250	3.343000	4.200000	1.0
75%	5.674500	6.471250	4.143000	4.770500	1.0
max	7.914000	6.925000	6.829000	6.212000	1.0

TABLE 8. PRINCIPAL STATISTICS OF VARIABLES WITH K=2.

Furthermore, looking at the average of the values of the variables present in the two clusters, it is possible to infer that the K=2 partition would lead to the composition of clusters of words that are almost similar to each other, with no net difference (Tab.8).

In conclusion, although a value of K=2 would seem, visually and in terms of Silhouette, to be one of the best options, in reality a value of K=8, although high and not characterized, provides optimal Silhouette values and allows for more detailed clustering, given the 4459 words that make up the Dataset and the wide scale of values of the variables chosen.

3. Classification

The purpose of the classification is to form a predictive analysis in order to define the various records of the dataset on the basis of predefined classes. It was initially thought to use as a class the variable "**polysemy**" as intuitively fitting for the purpose of classification, but later it was preferred to opt for another attribute that was not so strongly unbalanced as polysemy.

The choice, for the practical purposes of evaluating different algorithms, fell on "**gender**", transformed into binary values 0 and 1 to indicate female and male class membership respectively. As Attribute Set was used the dataset cleaned of variables "**word**" (as not suitable for our purpose), "**polysemy**" and "**web_corpus_freq**". The remaining variables are of continuous numeric type and do not present missing values or particularly controversial values for our models, and in any case Decision tree algorithms manage in an autonomous and optimized way this type of problems, and consequently we have not retouched any value of the variables in the Attribute Set.

The Training Set was finally populated by 70% of the values (3277) and the Test Set by the remaining 30% (1405).

3.1 Classification with Decision Tree

In the development of this algorithm "**Gini Index**" was used as a measure of impurity and the values *min_sample_split* and *min_sample_leaf* were set to 10 and 20 units respectively, since, as we will see later, these are the values that lead to better results. In Fig. 38 you can see the performance of the algorithm in the first four levels of nodes, which sets the main node according to the variable you consider most important, in this case "**Valence**", and displays with bluish colors the male nodes and colors tending to red the female ones. Also, the more a node has a faded or white color, the more the impurity value increases.

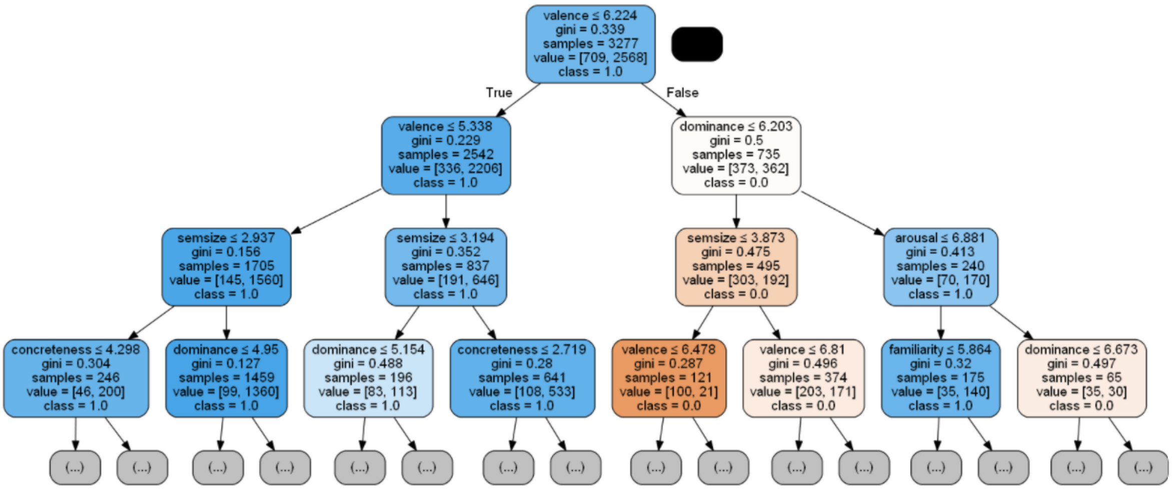
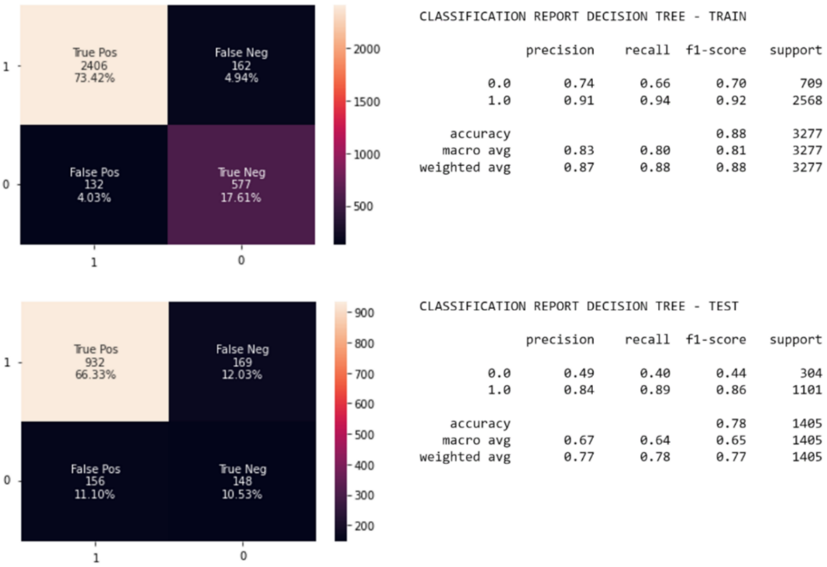

[redacted]

FIGURE 38. DECISION TREE WITH MAX_DEPTH=3

In order to quantitatively evaluate the algorithm, all the necessary indices were calculated regarding the use of the algorithm on both the Training Set and the Test Set. For a better understanding of the results it was useful to observe a *Confusion Matrix*. As can be seen in the Figure39, the Decision Tree tends to prefer True Positive in a substantial way and behaves fairly well with the training set, reaching an accuracy value of 88%. The performance decreases if applied to the test set as the accuracy does not exceed 78%. The trend of True Positive and True Negative can be seen in Fig. 40.

To analyze the results, Accuracy alone is not sufficient; therefore, **Precision**, **Recall** and **F1-score** values were also evaluated. It should be noted that F1-score represents the harmonic mean between the two previous values and can therefore be taken into consideration to get a general picture of performance. The results reported here are further confirmed by the *ROC curve* (Fig.40) which tends not to distance itself much from the bisector and to have a curve not particularly important implying a not too contained number of False Positives and False Negatives. Finally, to estimate the quality of the Decision Tree, **K-Fold Cross Validation** was calculated on the dataset. A single measurement generates data that are not very precise, which is why we apply a **k=10** that repeats the iteration k-times and gives us a single value that characterizes the average of the 10. We obtain 0.863 as a result, or **86.3%** accuracy according to *cross validation*, a fairly high value.



```
CLASSIFICATION REPORT DECISION TREE - TRAIN

              precision    recall  f1-score   support

         0.0       0.74      0.66      0.70       709
         1.0       0.91      0.94      0.92      2568

    accuracy                           0.88      3277
   macro avg       0.83      0.80      0.81      3277
weighted avg       0.87      0.88      0.88      3277
```

```
CLASSIFICATION REPORT DECISION TREE - TEST

              precision    recall  f1-score   support

         0.0       0.49      0.40      0.44       304
         1.0       0.84      0.89      0.86      1101

    accuracy                           0.78      1405
   macro avg       0.67      0.64      0.65      1405
weighted avg       0.77      0.78      0.77      1405
```

FIGURE 39. CONFUSION MATRIX ON TRAIN AND TEST SET

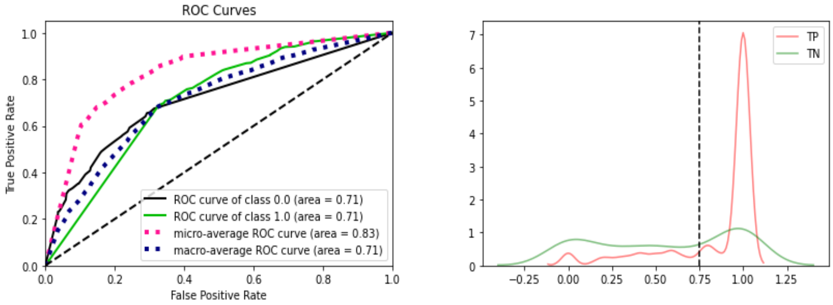


FIGURE 40. ROC CURVES AND REPRESENTATION OF TP AND TN

3.2 Classification with other Algorithms

To compare the performance of the Decision Tree, other classification models were chosen in order to verify which one could work better. The following Algorithms have been chosen:

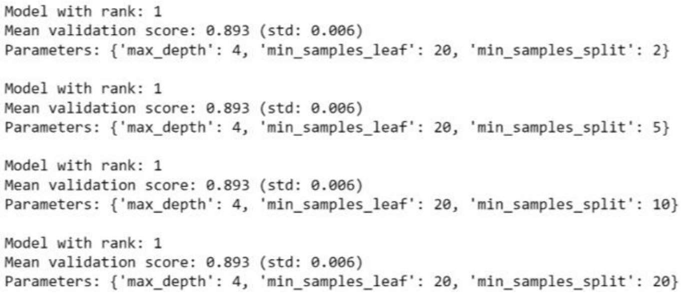
- *KNN*
- *Random Forest*
- *Gaussian Naive Bayes*
- *Support Vector Machine (RBF kernel)*
- *Support Vector Machine (Sigmoid kernel)*

The following table shows the results of the various measurements:

	Accuracy Test set	F1-Score Test set	Cross Validation
Decision Tree	0,759	0,846	0,853
KNN	0.802	0.881	0.884
Random Forest	0.837	0.902	0.900
Gaussian N. B.	0.781	0.862	0.864
SVM (rbf)	0.830	0.900	0.899
SVM (sigmoid)	0.784	0.879	0.879

TABLE 9. COMPARISON OF THE VARIOUS CLASSIFICATION ALGORITHMS

As mentioned earlier, the evaluation of Accuracy values alone is not sufficient. Therefore, the final evaluation will be the average of the three most important values: *Accuracy, f1-score and Cross Validation*. It is evident in this case that **Random Forest** and **Support Vector Machine** with rbf kernel obtain better results than the other algorithms.



```
Model with rank: 1
Mean validation score: 0.893 (std: 0.006)
Parameters: {'max_depth': 4, 'min_samples_leaf': 20, 'min_samples_split': 2}

Model with rank: 1
Mean validation score: 0.893 (std: 0.006)
Parameters: {'max_depth': 4, 'min_samples_leaf': 20, 'min_samples_split': 5}

Model with rank: 1
Mean validation score: 0.893 (std: 0.006)
Parameters: {'max_depth': 4, 'min_samples_leaf': 20, 'min_samples_split': 10}

Model with rank: 1
Mean validation score: 0.893 (std: 0.006)
Parameters: {'max_depth': 4, 'min_samples_leaf': 20, 'min_samples_split': 20}
```

FIGURE 41. PARAMETER TUNING

Regarding the tuning of the parameters of our basic algorithm, Fig. 41 shows the results. At parity we observe the models with **min_sample_leaf** parameter set to 20, intended as the minimum number of units needed to be a "leaf node"; raising this value, according to model ranking, would lead to a less accurate model. Subsequently there is **min_sample_split** that indicates the minimum number of records needed to make the algorithm proceed to the division of a node into two or more leaves, and that clearly the lower it is, the better results are obtained in relation to the lowering of the threshold set by us.

3.3 Considerations

The initial goal was to set up a Decision Tree that behaves in the best possible way with the dataset and therefore, for this purpose, it was preferred to use a class label as balanced as possible to avoid extreme cases. All the results reported concern only the test set, aware of the fact that the same models obtained far better results if applied on the training set. Despite this and given the uncertain nature of the dataset, it was still possible to obtain satisfactory results even reaching peaks of almost 85% accuracy.

The value of **Cross Validation was** similar among the algorithms we used, but the tree turns out to be the model with the lowest **Accuracy**, despite the parameters had been left deliberately low. Surprisingly instead, with the same settings, the **Random Forest** returns the best performance among all realizing a ROC curve substantially higher, allowing us to argue that it is definitely the algorithm that best suits the dataset.

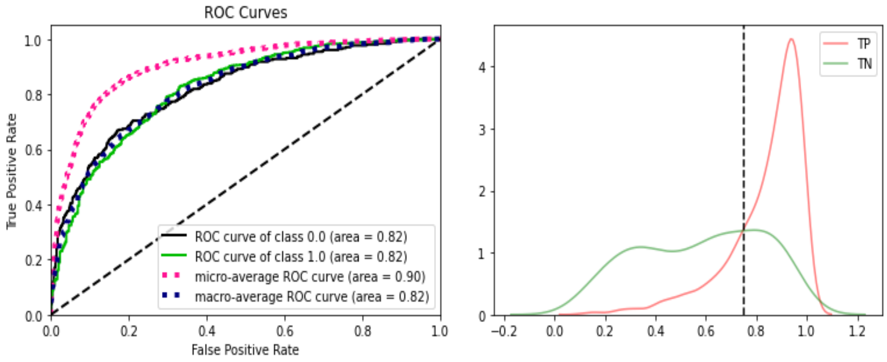


FIGURE 42. ROC CURVES WITH RANDOM FOREST

4. Pattern Mining

For this section it was decided to discretize the variables through the *pandas* **qcut** function, thus creating frequency classes based on the quartiles of the distribution of each variable. We kept the variables *length, polysemy, arousal, familiarity, aoa, semsize, gender* and we also decided once again to use the two variables **emotion** (given by the average of *dominance* and *valence*) and **visual** (given by the average of *concreteness* and *imageability*) already used in the clustering section.

4.1. Frequent pattern

Extracting the most frequent itemsets it is immediately evident that using a low support (0.02) the first itemsets of the list (a large list of 2225 itemsets) all contain the variable *polysemy* with value 'Polysemic'. By raising the support to 0.1 the list is reduced (13 itemsets) and if the variable in question appears it has the value 'Not Polysemic'. This is also due to the fact that most of the dataset is composed of non-polysemic words. This tendency, in fact, is also found by observing the number of **maximal itemsets** of the two values of the variable as the support increases, as shown in the graph (Fig. 43).

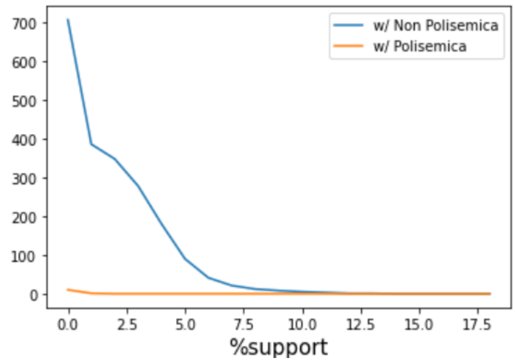


FIGURE 44. TREND MAX INTEMSETS FOR THE VARIABLE 'POLYSEMY'

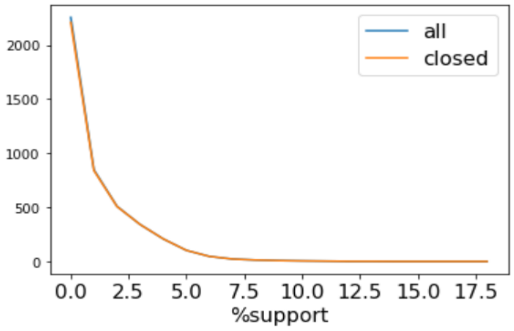


FIGURE 43. TREND OF TOTAL FREQUENT ITEMSETS AND CLOSED ITEMSETS

The minimum number of items per item set was set to 3. It can be immediately observed that the maximal itemsets containing 'Polysemous' have a very low support, and that with support values slightly above zero there are no itemsets containing 'Polysemous'. It is interesting to note that plotting the number of frequent itemsets, with varying support, and those that have the property of being **closed itemsets** (none of its neighboring supersets have the same support) we observe that the two sets practically coincide. Moreover, it is emphasized that with a support level of 0.1 the number of *frequent patterns* is very close to zero (exactly there are 13 patterns with support greater than 0.1).

Some of these itemsets with support greater than 0.1 are given below in order to comment on them.

	PATTERN		SUPPORT
(0.999, 3.606]_gender,	(5,797, 8,147]_emotion,	'Non-Polisemic'	10.440
(5.97, 6.925]_visual	(1,374, 3,438]_semsize'	'Non-Polisemic'	11.219
(1.2180000000000002, 3.114]_aoa'	(5,969, 6,939]_familiarity'	(1.999, 5.0]_length'	10.084
(5,152, 6,971]_aoa'	(1.6460000000000001, 4.706]_familiarity'	Non-Polysemic	15.182

FIGURE 45. TABLE FREQUENT ITEMSETS WITH SUPPORT

The first of the patterns informs about the possible correlation between a low value of *gender*, a high value of *emotion* and non-polysemy. This tendency of the words classified as feminine to present themselves with a high emotional charge had already been found in the clustering process (see section 2.1) and in the preparation of the data (Figure 16 and Figure 17 in which we visualize the correlation between *dominance* and *valence* by highlighting the gender of the words).

The second itemset reports that high values of *visual* correspond to low values of *semsize* and lack of polysemy. This kind of negative correlation between *visual* and *semsize* would mean that the more semantically small a word is, the easier it is to visualize. Considering, moreover, that our *visual* variable is formed by averaging the values of *concreteness* and *imageability*, we find confirmation of negative linear correlation between *semsize* and the other two variables in the heatmap in section 1.5.

The third and fourth patterns again confirm some of the evidence found in the data preparation phase. It had already been pointed out, in fact, that words learned as children tend to be short, and that, moreover, they possess a high level of familiarity. Figure 21 of section 1.5.4 allows us to see how polysemic words are concentrated in an area with high familiarity and low learning age. It is therefore correct to identify a frequent pattern with high enough support containing a high learning age class, low familiarity and the characteristic of non-polysemy. Vice versa, lowering the threshold of support, we find the itemset below, in which we find the same variables with opposite values. The fact that support is low is always thought to be due to the low presence of polysemous words in the dataset.

PATTERN	SUPPORT
Polysemic, 1.2180000000000002, 3.114]_aoa, 5,969, 6,939]_familiarity'),	10.440

4.2. Association rules

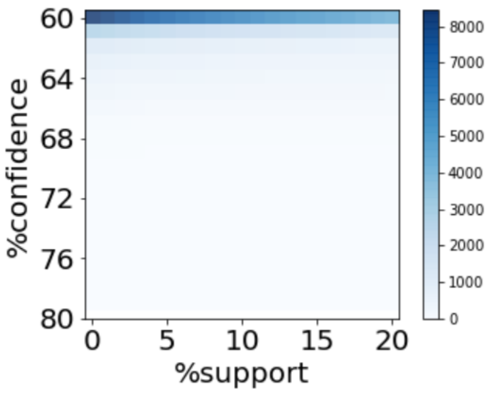


FIGURE 46. RULES ACCORDING TO THE PERCENTAGE OF CONFIDENCE AND

With regard to the extraction of association rules, it was decided to use 0.1 as the support threshold, 60 as the minimum confidence value and with a minimum number of 3 sets per item set to avoid trivial rules by inserting too low thresholds.
With these parameters we find 45 rules, moreover, as shown in the graph in Fig. 46, using a higher threshold of confidence the number of rules found would drop significantly. We also point out that wanting to search for rules with 4 item sets and the same other parameters, we find only more 5 rules. Observing the histograms of the confidence (Fig. 47), it can be observed that after the peak around 0.8 there are few rules with a high confidence, as regards the lift (Fig. 48), instead, it is underlined how a low negative correlation characterizes many rules (lift values less than one). Moreover, it is possible to notice some rules with a lift value greater than one, indicating a positive correlation between the terms of the rule.

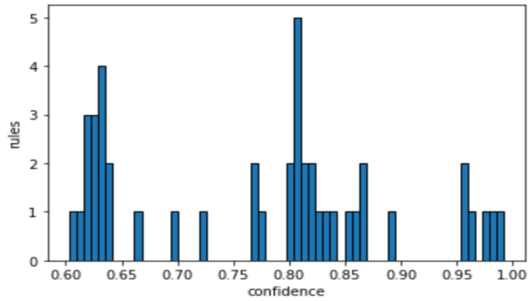
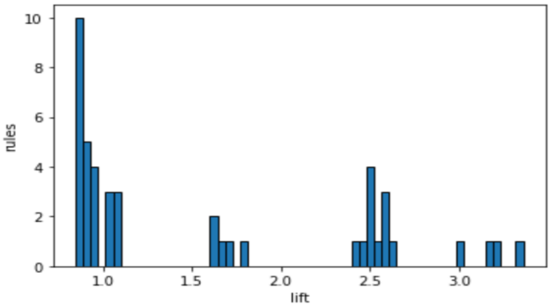


FIGURE 48. CONFIDENCE HISTOGRAM



FIGURES 47HISTOGRAM OF LIFT

Some rules are given below to comment on the semantics and any matches with other sections of the project.

RULES	SUPP	SUPP%	CONF	LIFT
''(5.419, 8.177]_arousal'',('(0.999, 3.606]_gender'', ''(5.797, 8.147]_emotion'', 'Non-Polysemic'	6,5%	6,5%	0.622	2.582
'Non-Polisemic',('(0.999, 3.606]_gender', '(5.797, 8.147]_emotion'	10,44%	10,44%	0.953	1.038
'Non-Polysemic,('(5.97, 6.925]_visual', '(1.218000000002, 3.114]_aoa'	9,93%	9,93%	0.854	0.93
'Non-Polysemic'('(1.2180002, 3.114]_aoa','(5.969, 6.939]_familiarity','(1.999, 5.0]_lenght	8,12%	8,12%	0.805	0.878

TABLE 10. RULES WITH RESPECTIVE VALUES

Regarding the extraction of association rules it was decided to use 0.1 as a support threshold, 60 as a minimum confidence value and with a minimum number of 3 sets per item set to avoid trivial rules by inserting thresholds that are too low.
The reported rules are to be read in this way Y » X and, for each of them, the values of absolute support (number of transitions), support in percentage, confidence and lift have been reported. Looking at the value

of lift, which indicates the ratio between the support observed and that expected if X and Y were independent (confidence normalized with the support of Y), we could immediately discredit the rules in which this value is close to 1, because it indicates that the probability of the antecedent and the consequent are independent, so they could be associated randomly. Regarding the first rule, in which a high lift is observed, it can be noted that the variables *gender* with a low value and *emotion* with a high value are present again (a trend already seen), the presence of *polysemy* with a negative value can be attributed again to the unbalance of the dataset while, on the other hand, on *arousal*, no considerations can be made since it is a variable little used in the rest of the work done. The last rule is one of the most interesting, since it contains more trends already found previously, moreover, the lift value less than one indicates that, if a word is short with a high familiarity and is learned from a young age, probably this word will not be polysemic (negative correlation). This trend is seen again in Figure 21 in section 1.5.4.

4.3 Predicting POLYSEMY

The choice of the target variable that we decided to predict fell on the polysemy variable, since it is the only binary variable in the dataset. Observing the association rules where this variable appears at the Y, however, we immediately notice that with the parameters used we cannot find any rule with a positive value of the variable, a factor due to the unbalance of the dataset. In addition, looking at the lift values of the rules where the variable appears at the Y with a negative value, there are predominantly negative correlations or independence conditions. Therefore, for the negative value, it was decided to carry out an interpretation work of the rules based on the lift value. For the positive value, on the other hand, a rule was first sought by lowering the support, coming to no result. Lowering, instead, also the confidence value up to 0,4 % finally appears the following rule:

RULE	SUPP	SUPP%	CONF	LIFT
'Polisemica',('(1.2180000000000002, 3.114]_aoa', '(5.21, 5.797]_emotion', '(4.121, 4.656]_gender', '(1.999, 5.0]'	23	0.51%	0.442	5.369

This rule, obviously, is not sufficient, given the scarcity of attestations (23). We therefore decided to aggregate an observation linked to the results obtained from the extraction of the frequent pattern, which was also found in the Data Understanding section. In fact, as mentioned above, polysemic words frequently report high values of familiarity and are learned in childhood. Applying what has been described the number of words without prediction was however high, for simplicity and with the awareness of having already a sufficient number of records classified with positive value, all the non-predicted words were classified as negative.

Thus, the results obtained are as follows:

dataset / prediction	**Polysemic**	**Non-polysemic**
Polysemic	TP = 87	FN = 283
Non-polysemic	FP = 380	TN = 3742

The accuracy of the model is 0.852.

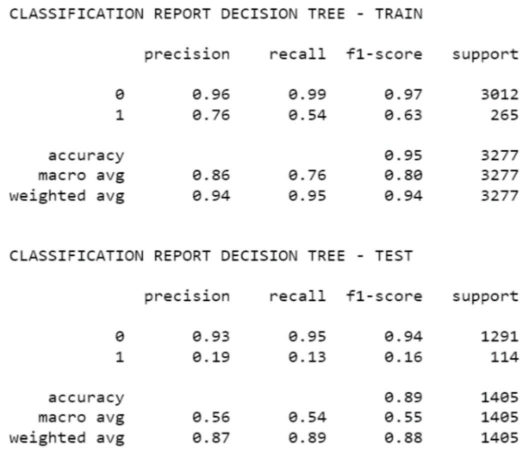
	PRECISION	**RECALL**
Polysemic	0.186	0.235
Non-polysemic	0.929	0.907

The F-measure of the model for predicting **Polysemic** is 0.207 while the F-measure for predicting **Non polysemic** is 0.918. These results were obtained using the whole dataset, to compare them with the results of a classification algorithm we divided the dataset into a train part and a test part (through a random splitting as for the classification) and applied the rules on the test set. Here are the results.

The accuracy of the model is 0.865.

	PRECISION	RECALL	FMEASURE
Polysemic	0.230	0.288	0.256
Non-polysemic	0.936	0.915	0.926

It is readily observed that the results on the test set are slightly better than those found on the entire dataset. Finally, we report the results of the DecisionTree algorithm obtained by attempting to rank the *polysemy* variable.



```
CLASSIFICATION REPORT DECISION TREE - TRAIN

              precision    recall  f1-score   support

           0       0.96      0.99      0.97      3012
           1       0.76      0.54      0.63       265

    accuracy                           0.95      3277
   macro avg       0.86      0.76      0.80      3277
weighted avg       0.94      0.95      0.94      3277

CLASSIFICATION REPORT DECISION TREE - TEST

              precision    recall  f1-score   support

           0       0.93      0.95      0.94      1291
           1       0.19      0.13      0.16       114

    accuracy                           0.89      1405
   macro avg       0.56      0.54      0.55      1405
weighted avg       0.87      0.89      0.88      1405
```

FIGURE 49. CLASSIFICATION REPORT DECISION TREE

In the end, it is observed that the classification algorithm achieves better results than the association rules based model, except for the precision and recall values of the positive value in the test set. Our model achieves higher results in classifying polysemous words from the Decision Tree when tested on the same test set. Nevertheless, the accuracy of Decision Tree is still higher than the other model.

Conclusions

For the analysis of polysemic and non-polysemic words in the dataset, different data mining techniques were used in order to extrapolate meaningful results. Often, however, due to the unbalanced nature of the dataset, suboptimal results were obtained but still useful for analysis. Section 1 of Data Understanding was fundamental in order to understand, based on statistical values and correlations, which variables could have been the object of the following sections. Semantics, distributions and data quality were analyzed and the few missing values present were eliminated. In the section of 2Clustering, we chose the variables considered useful for the search of groups with similar points and eliminated the outliers of these variables and then applied the algorithms of K-means, Hierarchical and DBSCAN noting that, on the first, we obtained better results although not optimal compared to the other algorithms. In section 3 of Classification, we opted for the creation of the binary variable 'gender' as a target variable, as the only binary variable in the dataset ('polysemy') contains classes clearly unbalanced for the Decision Tree. We used 10-fold-cross validation obtaining good results (mainly on train) and then we analyzed the results of accuracy, f1-score and cv of other classification algorithms, obtaining better results with Random Forest and SVM. In the 4Pattern mining section, it was found that frequent itemsets are completely composed of closed itemsets. To select the frequent itemsets we used different values of support and then we chose as threshold 0.1. Also for the rules extraction we compared the results obtained with different confidence levels and then we chose, as threshold, 60 and set 3 as minimum number of sets per item. The results of the rules were then compared with the previous sections of the report finding a linearity in the concepts and, finally, the target variable *'polysemy'* was predicted obtaining good results.