UNIVERSITY OF PISA

DEPARTEMENT OF INFORMATICS

DATA MINING

THE GLASGOW NORMS

Chiara De Nigris (586013)

Chiara Giurdanella (560686)

Simona Sette (544298)

ACADEMIC YEAR 2021/2022

CONTENTS

1. Data understanding

The Glasgow Norms dataset is made by the evaluation of a 4682 words corpus on nine psycholinguistic measures. This analysis will consider three more dimensions than the original set: Length, Polysemy and Web Corpus Freq.

1.1 Data semantics

Attribute	**Data type**	**Description**	**Range of value**
Arousal (AROU)	Float	Word's internal activation. The higher arousal a word has, the more exciting it is.	1-9
Valence (VAL)	Float	Value or worth. The higher valence a word has, the more positive it is.	1-9
Dominance (DOM)	Float	Degree of control one feels. The higher dominance a word has, the more dominant it is.	1-9
Concreteness (CNC)	Float	Degree to which something can be experienced by our senses.	1-7
Imageability (IMAG)	Float	Effort involved in generating a mental image of something.	1-7
Familiarity (FAM)	Float	Word's subjective experience.	1-7
Age of acquisition (AOA)	Float	Age at which that word was initially learned.	1-7 (2 years periods from 0–12 years and a final 13+ period)
Size (SIZE)	Float	Semantic dimension of a word.	1-7
Gender (GEND)	Float	How strongly a world meaning is associated with male or female behavior. The higher gender a word has, the more masculine it is.	1-7
Length	Int	Length of each word.	2-16
Polysemy	Int	Ability of a word to convey multiple meanings.	0-1
Web_corpus_freq	Float	Frequency of a word in the Google Newspapers Corpus.	127700 - 20224598480

Table 1 Data semantics.

1.2 Distribution of the variables

The first step in the analysis has been the exploration of numerical variables distribution, which were graphically analyzed using histograms. Most of the data presents a normal Gaussian distribution, except for three variables: *familiarity*, *concreteness* and *imageability*.

From Figure 1, it is possible to appreciate an initial exponential growth of the variable *familiarity,* that reflects a major frequency of words classifiable as medium-high familiar, followed by a drastic slope of frequency for the most familiar elements.

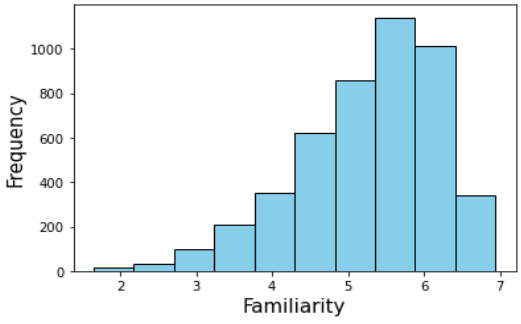


Figure 1 Distribution of variable *familiarity*.

As attested in Figure 2, variables *imageability* and *concreteness* present a bimodal behavior, analyzed as follow:

- for *imageability*, the trend reveals that the most frequent words are also those which require less effort in generating a mental image of them;
- for what *concreteness* is concerned, it seems that the most frequent words are both those which are easier to be experienced by our senses and those which are in a middle way between abstract and concrete.

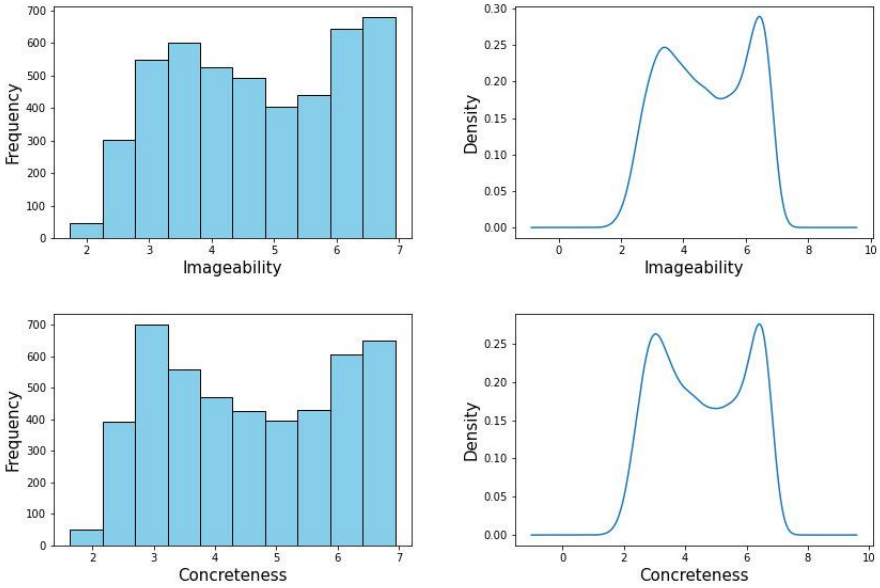


Figure 2 Distribution of frequency for variables *imageability* and *concreteness*. On the right side there are also reported the KDE.

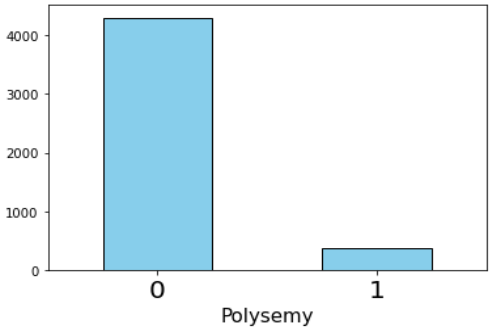


Figure 3 Distribution of polysemic words.

The only categorical feature in the set is the variable *polysemy,* that has been analyzed using a bar chart. Figure 3 clearly points out that this variable has a sharply unbalanced and asymmetrical distribution. Indeed, polysemic words are only 379, which is 8.09% of the whole set.

1.3 Data quality

This section of the labor has been centered on the research of duplicates, missing values and outliers and aims to increase the quality of the data. The first step has been researching duplicates in order to reduce the dimension of the dataset, but there wasn't any evidence of them among the records. Quality measures applied on outliers and missing values problems were more deeply explored in the following sections.

1.3.1 Missing values

According to what was reached after the research of missing values, it is possible to assert that the Glasgow Norms dataset is complete, syntactically and semantically correct. Indeed, the only attribute that presents empty cells is *web_corpus_freq*, with 14 empty instances. From the moment that these values can be classified as *Missing Completely At Random* (MCAR), empty instances of *web_corpus_freq* has been filled with the median of itself, grouping it by the discretized value of the most correlated attribute, *familiarity*.

1.3.2 Outliers

For what outliers are concerned, analysis has highlighted that only seven attributes present this problem: *arousal*, *length*, *valence*, *dominance*, *gender*, *familiarity* and *web_corpus_freq* (outliers of the latter are shown in Figure 5).

From Figure 4 it is possible to see that the number of outliers for each variable is quite irrelevant: it is never higher than the 4% of the sample size and so they cannot be considered noisy. Considering this observation and the fact that deleting outliers would mean to lose useful information to highlight important features of the dataset, data have been preserved.

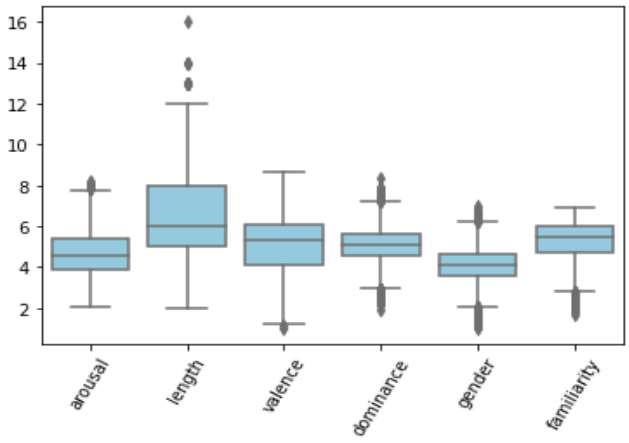


Figure 4 Variables outliers.

1.4 Data Transformation

This phase is focused on the transformation of variables to make data suitable for further analysis, such as clustering methods.

Considering that all the variables in the set are already distributed in ranges from 0 to 16, the only attribute that has been transformed was *web_corpus_freq* because of its wider range of values. Records have been handled performing a natural logarithmic transformation, obtaining a range of values that goes from approximately 10 to approximately 20 (excluding outliers). In Figure 5, it is shown the attribute *web_corpus_freq* before and after normalization.

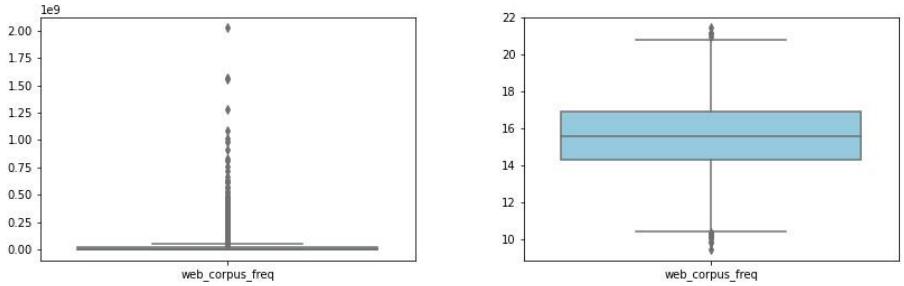


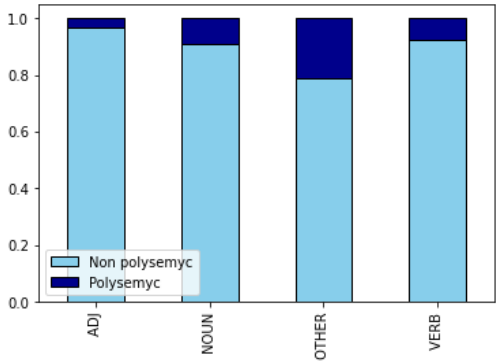
Figure 5 Distribution of *web_corpus_freq* before and after normalization. Pre-normalized values are represented on an exponential scale.

1.4.1 POS

To provide a wider and more appealing data analysis, it has been added a new categorical variable named *POS*, which categorizes the part of speech for every word. The classification has been realized using the NLP Stanford Stanza library[1]. The first step in this analysis has been to explore the distribution of the *POS*. As shown in Figure 6, the dataset is mostly composed by names. Other less numerous parts of speech, such as pronouns or appositions, have been collocated under the label *Other*.

Figure 6 POS distribution.

[1] https://stanfordnlp.github.io/stanza/



Data obtained have been used also to provide other interesting observations, such as the correlations between *POS* and Polysemic words displayed in Figure 7. Even if they are the fewer in the set, *Other* parts of speech reveal to be the more polysemic among all the *POS*.

Figure 7 Distribution of polysemic part of speech.

Using *POS*, a further analysis has been provided for the variables *familiarity* and *concreteness*. From Figure 8, it appears that nouns tend to be generally familiar and concrete, whereas adjectives are mostly familiar but less concrete than nouns and verbs.

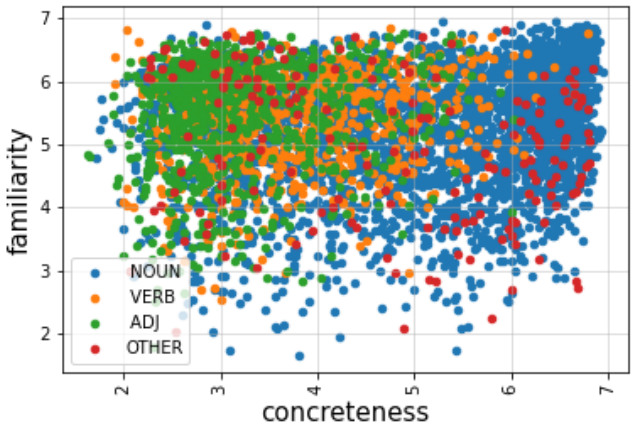


Figure 8 Distribution of *familiarity* and *concreteness* for *POS*.

1.5 Correlations

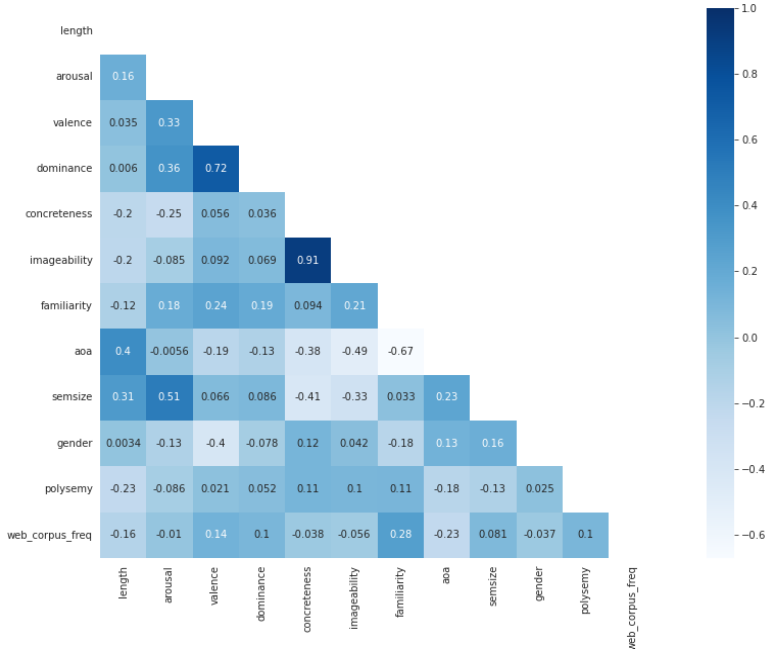


Figure 9 Correlation matrix.

Figure 9 shows correlations between the variables. As it is evident, *concreteness* and *imageability* are strongly correlated. This was an expected observation, from the moment that the more concrete a word is, the easier it is to imagine. As a confirmation of the fact that the more positive a word is, the more it provokes feelings of dominance, also *valence* and *dominance* appear to be solidly connected.

2. Clustering

This section is centered on cluster analysis and preliminary phases for its implementation. In the following sections, there will be explored and compared three different cluster techniques: center-based clustering through *K-means* algorithm, *Hierarchical* clustering in an agglomerative way and density-based clustering using *DB Scan*.

The first step has been to make data suitable for the clustering algorithm. For this reason, all the attributes which presented a non-linear distribution of values have been normalized through the min-max normalization. After this passage, the inspection has been centered on the selection of attributes to include in this phase of analysis, for the purpose of having a plainer representation of clusters. From the moment that these three features are the more representative from a psycholinguistic point of view, analysis has been focused just on three variables: *arousal*, *valence* and *dominance*.

2.1 K-means

To apply *K-means* algorithm to the subset, the first operation has been to select the number of desired clusters. To choose the right number of clusters, they have been implemented different trials which results have been evaluated with two different metrics: *Silhouette* and *SSE*.

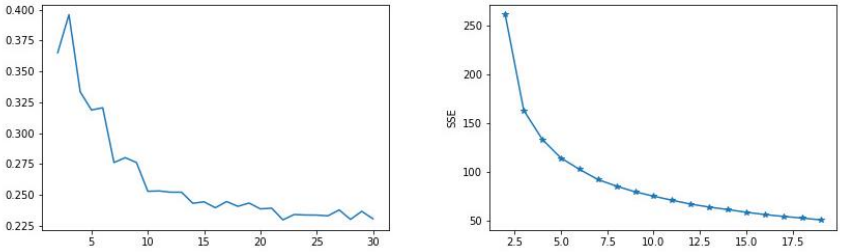


Figure 10 *Silhouette* distribution and *SSE* distribution.

From the moment that it's possible to visualize a peak in correspondence of point 3 in both graphs in Figure 10, the number of clusters (K) has been defined as 3. Furthermore, from the first graph it is perceivable that *Silhouette* score is a number near 0 for all the values and this result suggests that clusters won't be well separated but overlapping.

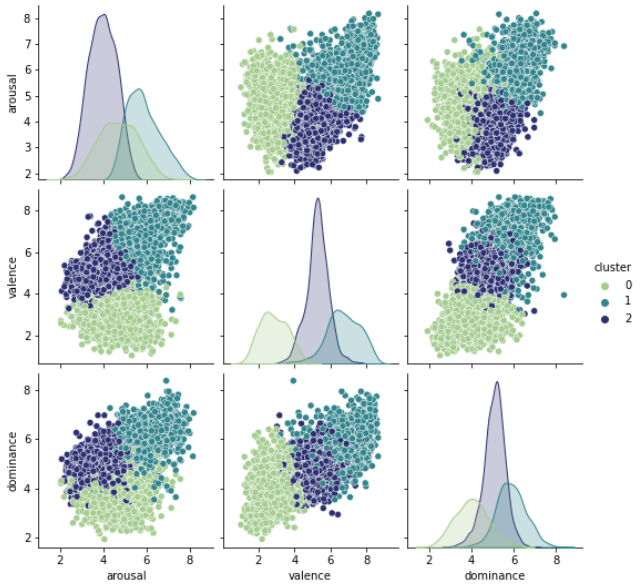


Figure 11 Clustering with *K-means*.

K-means algorithm outputs are three clusters, a bigger one and two made by nearly the same number of elements, as shown in Table 2.

Cluster	Number of elements
0	2222
1	1321
2	1139

Table 2 Elements for each cluster after the first execution of K-means.

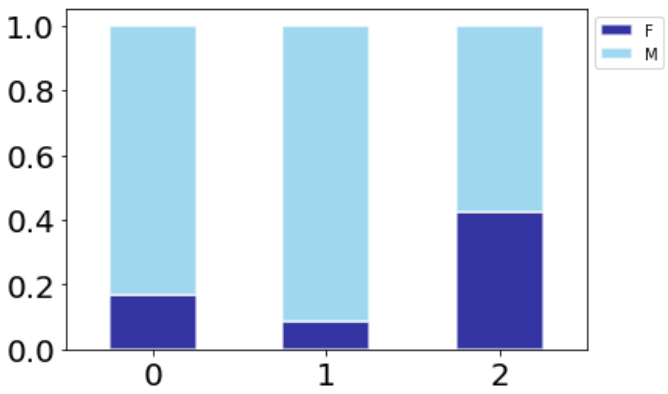


Figure 12 Distribution of the target variable *gender* among the clusters.

To provide a further analysis, it has been investigated the distribution of the variable *gender* among the clusters. To proceed in this task, the first step has been the transformation of the variable from continuous to categorical, getting, in this way, a binary classification of female (F) and male (M) words. From the example in Figure 12, obtained by plotting the results for the variables *arousal* and *valence*, it's possible to visualize the prevalence of male words among all the three clusters and a particularly unbalanced distribution in the first two.

Figure 13 presents the centroid's distribution. From the graph it is possible to notice that centroids for *valence* attribute are the ones furthest between each other, whereas *arousal* and *dominance* centroids are closer.

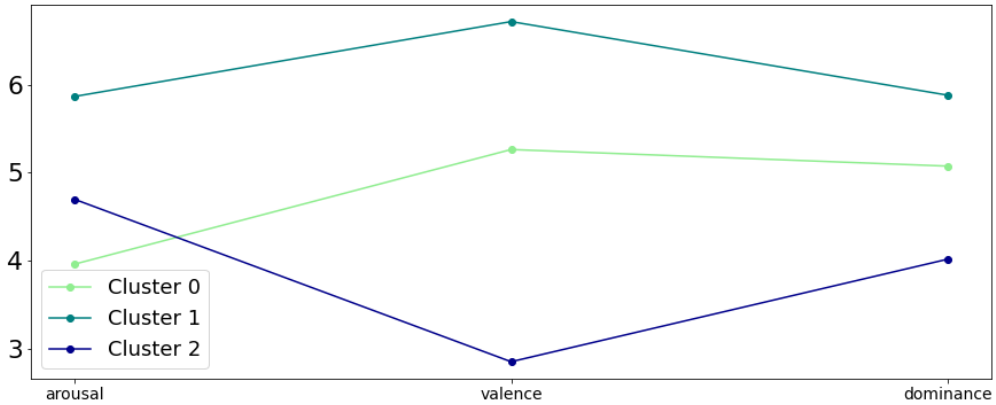


Figure 13 Distribution of cluster's.

Considering *Silhouette* values and the case that *K-means* it's a partitional cluster algorithm that needs well separated clusters to perform optimally, data show that this algorithm is not suitable for the globular-shaped Glasgow dataset.

2.2 Hierarchical clustering

For what *Hierarchical* clustering is concerned, analysis has been developed on highlighting how different data in the subset are in terms of a single number, focusing on dissimilarity or distance.

To calculate distance, the selected metric has been the *Euclidean distance*, from the moment that this measure is more suitable to work on continuous variables as the ones contained in the Glasgow dataset. Furthermore, the algorithm has been implemented using four different methods (*single*, *ward*, *average* and *complete*), aiming to compare results and define the more adequate between them.

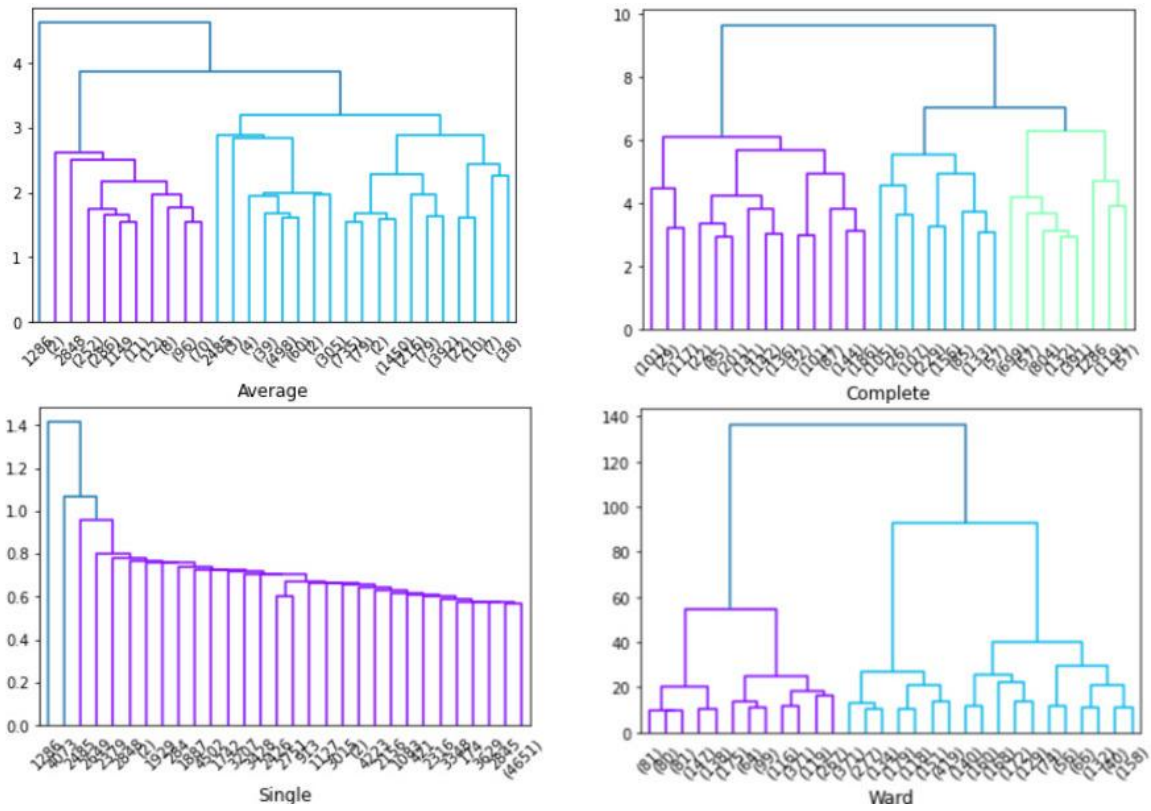


Figure 14 Hierarchical clustering methods.

To provide a valuation of the performance of the different methods, it has been analyzed elements distribution inside the clusters and the *Silhouette* scores. For the first task, it has been provided a quantification of elements inside the clusters taking a sample of 4 clusters for each method.

Method	Clusters distribution	Silhouette
Average	[3030, 912, 739, 1]	0.35
Complete	[2083, 1524, 898, 177]	0.22
Single	[4679, 1, 1, 1]	0.30
Ward	[1649, 1295, 1211, 527]	0.28

Table 3 Comparison between Hierarchical clustering methods.

Both clusters distribution and *Silhouette* values have been selected as metrics to define the best method and, after a balanced evaluation, the *Ward* method has been identified as the most suitable for the dataset. As expected, according to the evidences that the set is noisy and clusters have globular shapes, result obtained through *Single* method can be considered meaningless from the moment that clusters are highly unbalanced.

2.3 Density based clustering

Density based clustering has been implemented using *DB Scan* algorithm, which needs the tuning of two hyper-parameters: *epsilon (ε)*, which value corresponds to the radius of neighborhood around a certain point, and *min-points,* that is the minimum number of neighbors within "*eps*" radius.

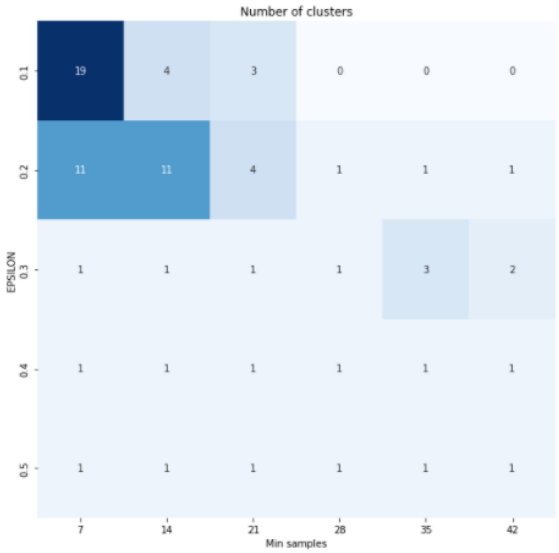


Figure 15 Number of clusters for different tunings of *eps* and *min_samples*.

To choose the best value for both, it has been explored a practical approach to the problem, trying to detect which was the composition of the clusters for different values of ε in a range from 0.1 to 0.5 and of *min_samples* from 7 to 42. As is clear from Figure 15, most of combinations return only one cluster, whereas smallest values of ε combined with smallest values of *min_samples* are the ones that reveal a more interesting clusterization. Exploring the composition of clusters, it is possible to observe that most of them are highly unbalanced, exception for few cases.
For example, for ε =0.3 and *min_samples*=35, the distribution of elements inside the clusters is the most relevant.

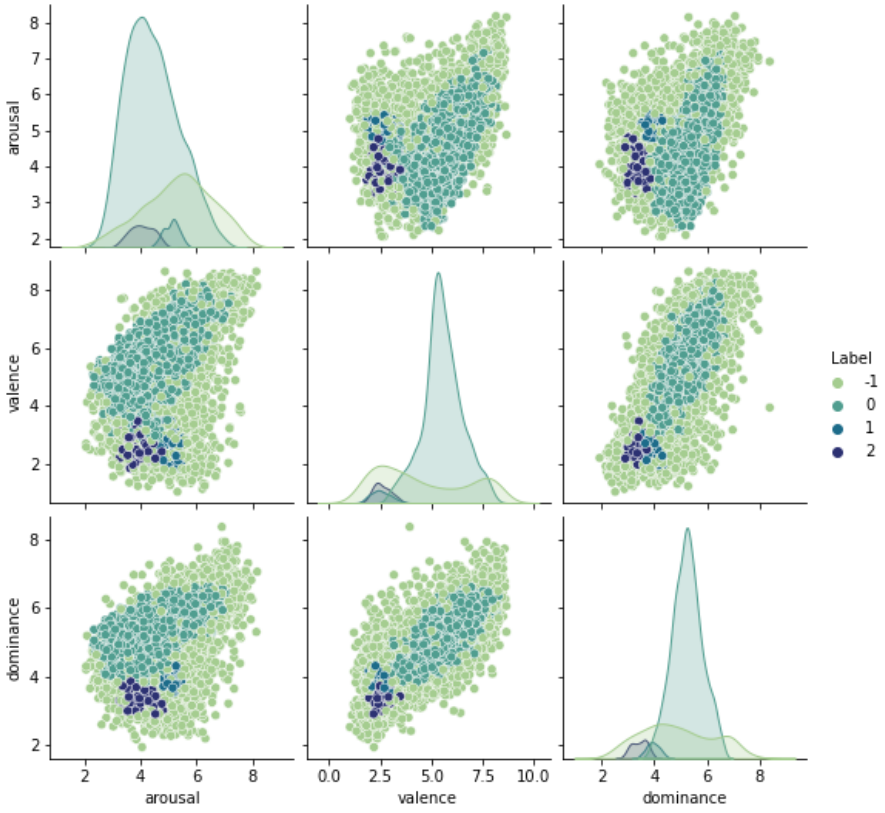


Figure 16 *DB Scan* algorithm with ε =0.3 and *min_samples*=35.

Indeed, the distribution of elements inside the clusters is the following:

Cluster	Number of elements
0	3131
1	111
2	157

Table 4 Distribution of elements inside clusters.

Moreover, 1283 elements have been classified as noise (represented in Figure 16 by the -1 label). However, from this data we can still notice one bigger cluster and two smaller ones.

From Figure 16, is clear that *DB Scan* algorithm is not actually suitable for Glasgow Norms dataset for its dimension and clusters globular shapes.

2.4 Final discussion

Clustering analysis reveals that this dataset is not actually suitable for this task or, at least, for the three explored algorithms. The reason of the failure can be individuated in the Glasgow Norms dataset's dimension and in globular-shape and not well-separated resulted clusters.

Algorithm	Silhouette	Number of clusters	Distribution of elements inside clusters
K-means	0.4	3	[2222, 1321, 1139]
Hierarchical	0.28	4	[1649, 1295, 1211, 527]
DB Scan	0.16	3	[3131, 111, 157]

Table 5 Comparison between clustering algorithms. Results for Hierarchical Algorithm refer to the ones obtained through *Ward* method.

As shown in table 5, *DB Scan* algorithm does not present neither a high silhouette value neither a balanced distribution among clusters, so the final evaluation of the best algorithm is reduced between *K-means* and *Hierarchical*.

K-means and *Hierarchical* both present a similar distribution of elements inside the clusters but, considering that *Hierarchical* has been implemented with a major number of clusters but still has a lower value of Silhouette, *K-means* algorithm appears to be the best choice for clustering analysis on Glasgow Norms.

3. Classification

This section is centered on the classification task, with the aim to build a model able to make predictions based on the dataset attributes. To achieve this goal, they have been explored two different algorithms: *Decision Tree* and *K-Nearest Neighbors* (*K-NN*).

3.1 Pre-processing

To make the data more suitable for the task they have been provided some transformations. All the categorical attributes have been directly transformed through a *One-hot encoding*. From the moment that they have considered to be not interesting for the classification purposes the attributes *word*, *length*, *POS* and *web_corpus_freq* have been dropped from the data set.

The target variable selected for the classification has been the *gender* attribute in its categorical form, keeping the transformation proposed in section 2.1. Exploring its distribution it has been noticed that the values of the target variable present a negligible displacement: the Glasgow Norms contains 1032 female words and 3650 male words.

In order to proceed with a *hold out validation* and evaluate the results, the first step of the analysis has been the division of the data in two parts: the train set, composed by 70% of the dataset, and the test set, composed by 30% of the original set.

	N. of elements	**N. of attributes**
Train set	3277	9
Test set	1405	9

Table 6 Distribution of train and test set.

3.2 Decision Tree

The first step in the implementation of the *Decision Tree* algorithm has been the hyper-parameters tuning in order to maximize the value of the *F-measure* score. This criterion has been selected because it is a harmonic mean of *Precision* and *Recall*, suitable for the evaluation required on this set. To decide the best values for *max_depth min_samples_leaf* and *min_samples_split* it has been implemented the *Grid search* technique with different ranges for each parameter. Computation has been developed for both *Gini* and *Entropy* metrics:

- *max_depth*: [2, 5, 10, 15, 20, 25, 30, None],
- *min_samples_split*: [2, 5, 10, 20],
- *min_samples_leaf*: [1, 5, 10, 20].

The best values for the two metrics provided by the *Grid search* are the ones shown in the following table:

Criterion	**Max_depth**	**Min_samples_leaf**	**Min_samples_split**	**F-measure Score**
Gini	5	5	2	0.702
Entropy	15	10	2	0.700

Table 7 Best tuning for hyper-parameters with *Gini* and *Entropy* metrics.

The *Grid search* shows that the *Gini* metric provides better results than *Entropy,* even if results are similar. To be sure to find the optimal values for hyper-parameters tuning, it has been provided also an evaluation through another metric: the *accuracy* measure, that is the number of predictions correctly classified by the model. Possible values of *accuracy* have been plotted on a line graph for each value of *max_depth* in a range from 0 to 30 for both *Entropy* and *Gini* rate.

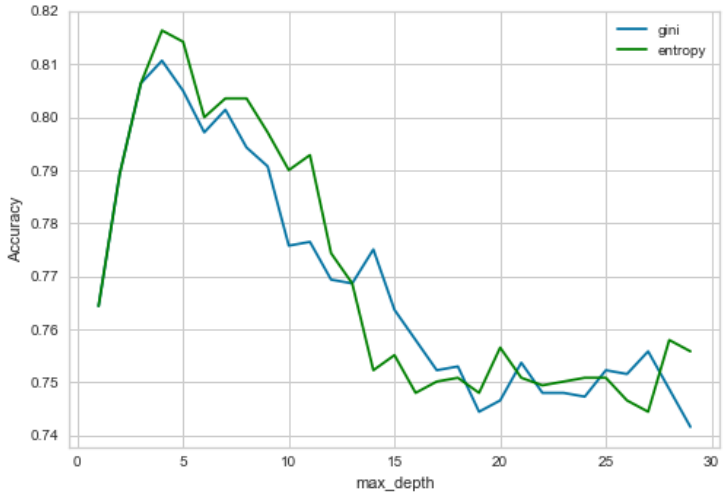


Figure 17 Values of *accuracy* for different tuning of *max_depth*

The graph in Figure 17 shows that 4 for *max_depth* with *Entropy* metric is the best value to maximize *accuracy* and, so, the grid search has been computed again to explore the best values of *min_samples_split* and *min_samples_leaf* with a *max_depth* equal to 4. Results are shown in Table 8.

Criterion	Max_depth	Min_samples_leaf	Min_samples_split	F-measure Score
Entropy	4	1	20	0.697

Table 8 Best tuning for hyper-parameters with *max_depth*= 4.

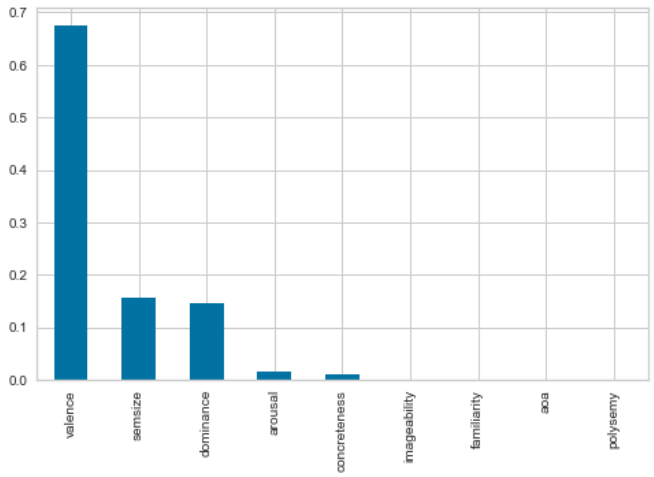


Figure 18 Informative variables computed on the decision tree.

Using *max_depth* equal to 4, the *F-measure* score decreases with a few loose but the *accuracy* value improves. Taking stock of these results, it has been decided to preserve anyway 4 as value of *max_depth*.

After the tuning, the analysis proceeded with the computation of the more informative variables to classify words by gender according to the selected hyperparameters values.

As is shown in the plot in Figure 18, with these parameters tuning only four variables are excluded from the analysis of informative variables. The attribute *valence* has the predominance on the others, as a matter of fact, this confirms the tendency of words of being affected by gender in the classification as more positive or negative. Indeed, as shown in the correlation

matrix in section 1.5, *gender* and *valence* are inversely proportional, from the moment that the more positive words (with a high value of *valence*) tend to be classified as female (so with a low value of *gender*).

Once the hyper-parameters have been tuned, the analysis has proceeded with the evaluation of the proposed model for both train and test set. The results on the train set have been evaluated as the performance of the model on the dataset, whereas the performances on the test set are an index of the ability of the model to generalize its task of classification, from the moment that it is working on data never seen before.

The following tables summarize the scores for *precision*, *recall* and *F-measure* scores on the train and the test sets, calculated on the confusion matrix.

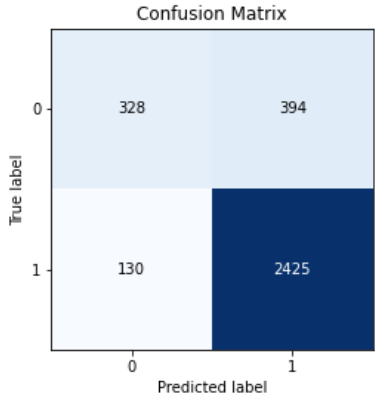


Figure 19 Confusion Matrix for the train set.

Class	Precision	Recall	F1 Score	Support
0 (female)	0.72	0.45	0.56	722
1(male)	0.86	0.95	0.90	2555

Table 9 Evaluation metrics for *Decision Tree* on train set.

On the train set values, the model reaches a value of *accuracy* equal to 0.84, with 2753 elements correctly classified.

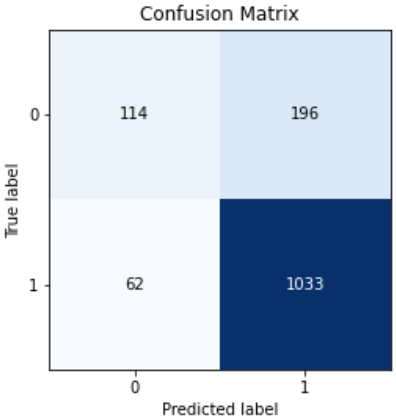


Figure 20 Confusion Matrix for the test set.

Class	Precision	Recall	F1 Score	Support
0 (female)	0.65	0.37	0.47	310
1(male)	0.84	0.94	0.89	1095

Table 10 Evaluation metrics for *Decision Tree* on test set.

On the test set values, the model reaches a value of *accuracy* equal to 0.82, with 1147 elements correctly classified.

A further evaluation has been provided through *ROC curve*, shown in Figure 21, which represents the relationship between True positive and False positive predictions. It has also been calculated the *AUC* (area under the curve) value, which reaches the value of 0.76.

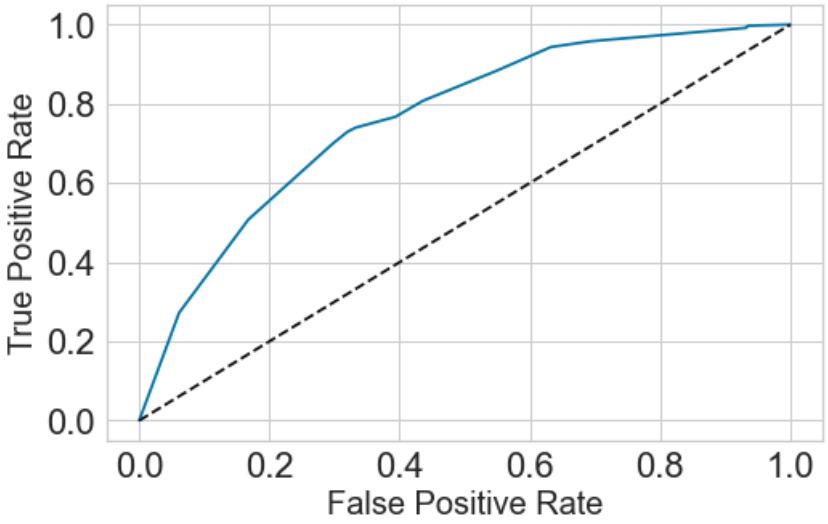


Figure 21 *ROC curve* on the test set.

As last step of the classification analysis with the *Decision Tree* algorithm, to provide another metric for the evaluation of the model performance, results have been compared with a baseline provided by a *Dummy Classifier*, a random guesser implemented with different strategies. It has been provided the implementation of three different guesser, one which generates predictions uniformly at random (labeled in Table 11 as *Baseline1*) and other two which predict first only the male class and then only the female class (respectively in the table as *Baseline2* and *Baseline3*).

	Decision Tree		**Baseline1**		**Baseline2**		**Baseline3**	
Accuracy	*0.82*		*0.49*		*0.78*		*0.22*	
	0 (female)	*1(male)*	*0 (female)*	*1(male)*	*0 (female)*	*1(male)*	*0 (female)*	*1(male)*
Precision	0.65	0.84	0.20	0.76	0.00	0.78	0.22	0.00
Recall	0.37	0.94	0.46	0.49	0.00	1	1	0.00

Table 11 Evaluation metrics for the *Dummy Classifier*.

Our model, compared with these baselines, appears to be more accurate and achieve better the task of classification for both the classes.

Figure 22 shows the tree build by the *Decision Tree* algorithm.

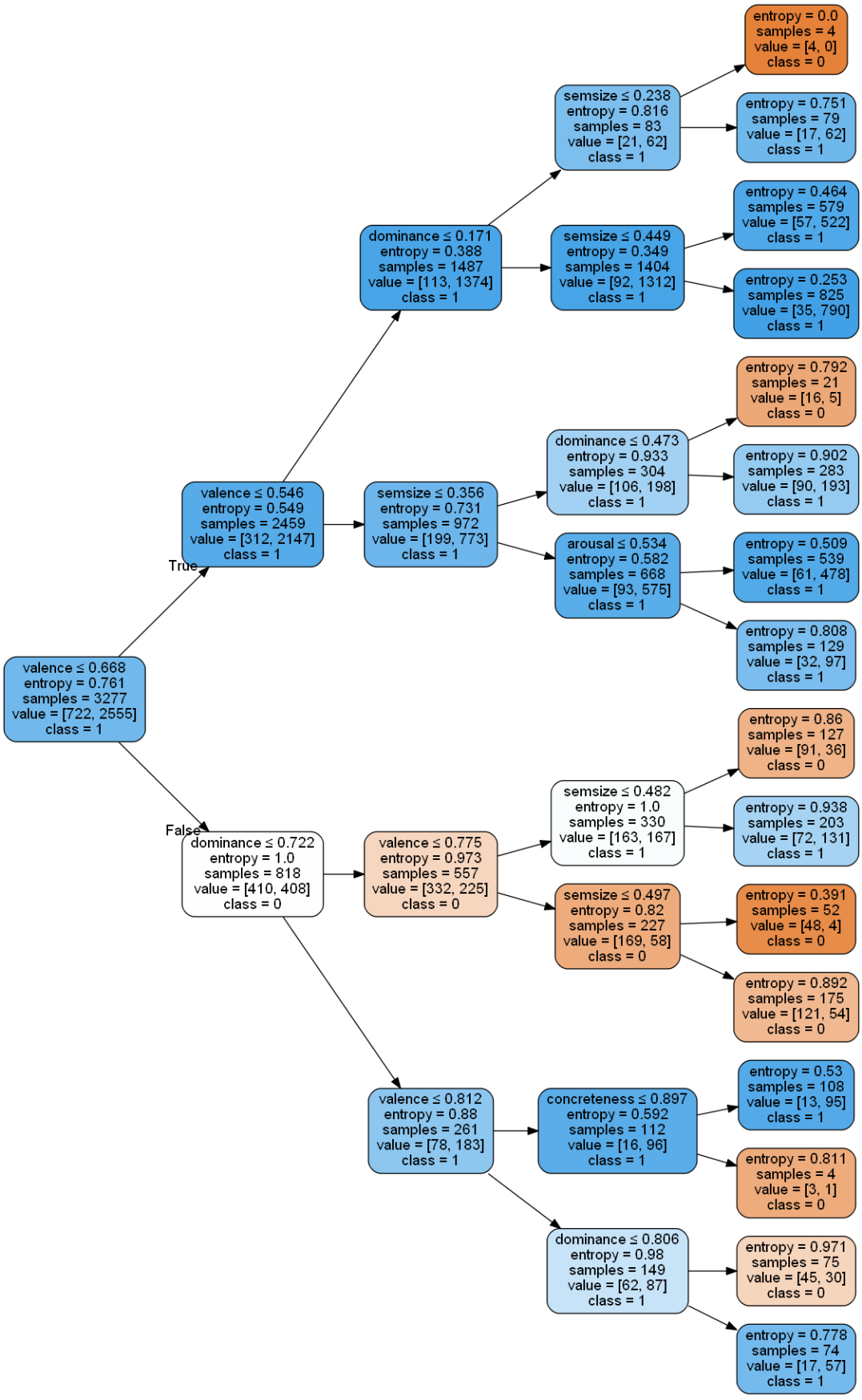


Figure 22 Decision Tree for target variable *gender* with *max_depth*=4.

3.3 KNN

After the implementation of the *Decision Tree* algorithm, it has been tested an Instance-based learning algorithm: *K-Nearest Neighbors* (shortened in *KNN*).

As known, the algorithm requires a proximity measure to determine the similarity (or distance) between instances and a classification function that returns the predicted class of a test instance based on its proximity to other instances. To implement *KNN*, is necessary to set the number of k, the neighbors that will be part of a certain class. For this reason, the first step in the analysis has been the research of the best value of k for different metrics using the *Grid search* technique. The grid has been developed to find the optimal value between a range of k from 0 to 30, three different distances (*Euclidean*, *Manhattan* and *Minkowski*) and two different kinds of weights (*distance* and *uniform*).

At the end, the best result on the test set has shown to be the *Manhattan* metric with 19 neighbors and the *uniform* weights. The *KNN* algorithm has been tested both on the train and on the test set. The evaluation scores proposed are the same tested on *Decision Tree* algorithm and are shown in Table 12.

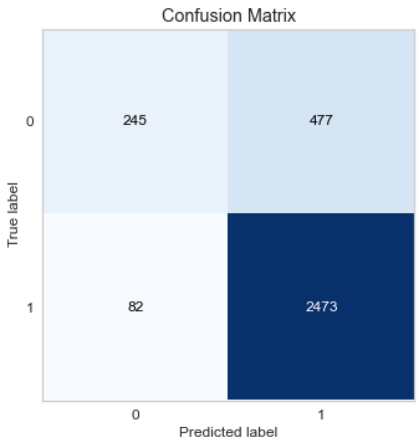


Figure 23 Confusion matrix for *KNN* on train set.

Class	Precision	Recall	F1 Score	Support
0 (female)	0.75	0.34	0.47	722
1(male)	0.84	0.97	0.90	2555

Table 12 Evaluation metrics for *KNN* on train set.

On the train set, the model reaches an *accuracy* value of 0.83, with 2718 elements correctly classified.

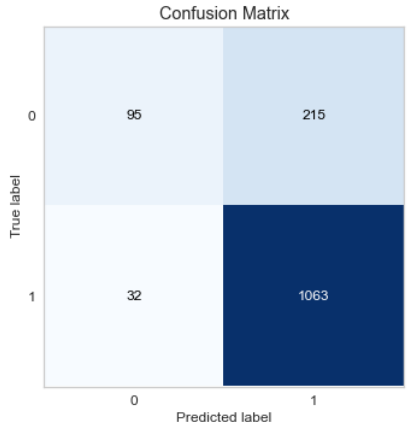


Figure 24 Confusion matrix for *KNN* on test set.

Class	Precision	Recall	F1 Score	Support
0 (female)	0.72	0.32	0.45	310
1(male)	0.83	0.97	0.90	1095

Table 13 Evaluation metrics for *KNN* on test set.

On the test set, the model reaches an *accuracy* value of 0.82, with 1158 elements correctly classified.

3.4 Comparison between Decision Tree and KNN

From the moment that evaluation metrics on *Decision Tree* and *KNN* algorithms returned very similar values, to provide another comparison between the two models they have been calculated the learning curves for both.

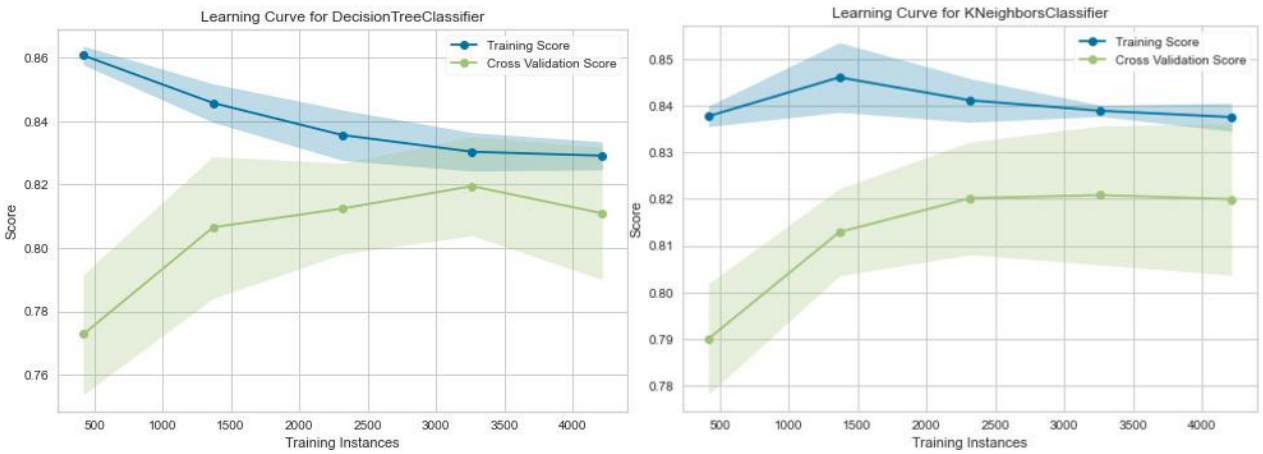


Figure 25 Learning curves for the *Decision Tree* and *KNN* algorithms.

From the learning curves presented in Figure 25 it is possible to notice that both models suffer of variance problem. However, it's obvious from the observable gap between *Training score* and *Cross validation score* that the *KNN* algorithm appear to be affected by a higher variance problem than the *Decision Tree*. Theoretically, both models might perform better in the classification task getting more training data, but with some differences:

- for what *Decision Tree* is concerned, the optimal performance value appears to be reached already with the training set composed for the experiment (70% of the dataset, as explained in section 3.1). This evidence confirms that adding more data could improve the performances but only if they do not exceed a certain amount, in order to avoid another gap's rise;

- whereas, for *KNN* the distance remains steady from a certain point, not showing a significant of improvement with the increasing of training sample. However, the soft decrease of the training score at the end of the curve suggests that it could be observed an improvement by adding more data.

According to the previous observations, *Decision Tree* can be defined as slightly more suitable for the classification task, even if both the models reach a good level of *accuracy* and can be considered satisfying for the prediction of the *gender* variable.

4. Pattern Mining

This section is focused on the pattern mining analysis in order to identify, extract and study frequent patterns and association rules in the Glasgow Norms dataset. The extracted rules have also been used to achieve the task of *gender* variable prediction.

4.1 Pre-processing

To make the data suitable for the analysis and implement the *Apriori algorithm*, attributes have been pre-processed in the following way:

- the columns *word*, *POS* and *web_corpus_freq* have been dropped because not informative for the task;
- all the variables have been discretized into a number of buckets based on sample quantiles, coherent with its distribution. *Valence*, *Arousal* and *Dominance* have been discretized into 5 bins and *Concreteness*, *Imageability*, *Familiarity* and *Semsize* into 4 using the *q_cut* function. The other two variables, *Aoa* and *Length*, have been respectively discretized into 4 and 3 bins:
  - *Aoa* has been divided into smaller ranges according to the age in which a word has been learned: *0-4*, *5-8*, *9-12* and *12+*;
  - as far *Length* is concerned, words have been divided into *short*, *medium* and *long*.

The only attribute whose hasn't been transformed is *Polysemy*, which was already categorical and which values have only been renamed from 0 and 1 to *not polysemic* and *polysemic*.

	Gender	Valence	Arousal	Dominance	Concreteness	Imageability	Familiarity	Semsize	Length	Aoa	Polysemy
0	M	(0.9, 3.6]_Val	(3.7, 4.3]_Arousal	(4.3, 4.9]_Dom	(4.5, 6.0]_Conc	(3.5, 4.7]_Imag	(1.5, 4.7]_Fam	(4.2, 4.9]_Semsize	medium	12+	not polysemic
1	F	(5.5, 6.3]_Val	(2.0, 3.7]_Arousal	(4.3, 4.9]_Dom	(4.5, 6.0]_Conc	(4.7, 6.0]_Imag	(1.5, 4.7]_Fam	(4.9, 6.9]_Semsize	medium	9-12	not polysemic
2	M	(5.0, 5.5]_Val	(2.0, 3.7]_Arousal	(4.9, 5.3]_Dom	(3.2, 4.5]_Conc	(1.5999999999999999, 3.5]_Imag	(4.7, 5.4]_Fam	(1.2999999999999998, 3.4]_Semsize	long	9-12	not polysemic
3	M	(3.6, 5.0]_Val	(3.7, 4.3]_Arousal	(4.3, 4.9]_Dom	(3.2, 4.5]_Conc	(1.5999999999999999, 3.5]_Imag	(1.5, 4.7]_Fam	(4.2, 4.9]_Semsize	medium	12+	not polysemic
4	M	(3.6, 5.0]_Val	(3.7, 4.3]_Arousal	(4.3, 4.9]_Dom	(3.2, 4.5]_Conc	(1.5999999999999999, 3.5]_Imag	(1.5, 4.7]_Fam	(4.9, 6.9]_Semsize	long	12+	not polysemic

Figure 26 New dataset resulted from the transformation of the variables.

None of the attributes used for this task presented missing values, so it hasn't been necessary to transform anything in this regard.

After the transformations, the dataset has been converted into a list.

4.2 Frequent patterns extraction

The following phase of the analysis has been the frequent itemsets extraction. It has been decided to consider only the itemsets with a minimum length higher or equal to 3, in order to include only the more interesting results in the study. The first step has been quantifying how the number of founded itemsets varied with different thresholds of *support*, as shown in Figure 27.

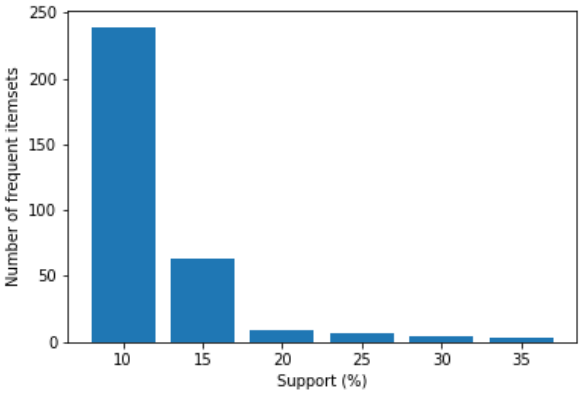


Figure 27 Number of frequent itemsets at varying of *support*.

Results shows that increasing the value of *support* the number of itemsets founded decreased. From the moment that the number of itemsets inspected with a percentage of *support* of 10% is widely dispersive and with values of *support* higher to 25% tend to zero, the analysis has been centered on the frequent itemsets extracted in the range from 15% to 25%. This process has been implemented also for *closed* and *maximal* itemsets, which present roughly the same attitudes, with values contained in Table 14.

Support (%)	Frequent itemsets	Closed itemsets	Maximal itemsets
15	63	63	50
20	9	9	5
25	7	7	3

Table 14 Number of *all*, *closed* and *maximal* frequent itemsets with different percentages of *support*.

At this point, frequent patterns have been extracted for different thresholds of *support*, as shown in Table 15. *Closed* and *maximal* itemsets have been omitted from the table because are equal to the ones found for *all* frequent itemsets.

Support (%)	Frequent patterns
15	**1.** ' (3.7, 4.3]_Arousal', 'M', 'not polysemic'; **2.** '(0.9, 3.6]_Val', 'medium', 'not polysemic'; **3.** '(2.0, 3.7]_Arousal', 'M', 'not polysemic'.
20	**1.** '(1.5, 4.7]_Fam', 'M', 'not polysemic'; **2.** '5-8', 'medium', 'M'; **3.** '5-8', 'M', 'not polysemic'.
25	**1.** '5-8', 'medium', 'not polysemic'; **2.** '9-12', 'medium', 'M'; **3.** '9-12', 'medium', 'not polysemic'.

Table 15 First three frequent itemsets with different percentages of *support*.

4.3 Association rules extraction

The study proceeded with the extraction of association rules, based on the values of *confidence* and *lift*. The first operation to accomplish has been to tune the parameters of *support* and *confidence* for the *Apriori* algorithm. For the *support* threshold, the optimal value has been identified as 20%, because, after this point, it is possible to notice a high decrease in the number of generated itemsets (as discussed in section 4.2). For what *confidence* is concerned, it has been quantified the varying of number of extracted rules at the enhancing of the measure, as shown in Figure 28. From the plot it is possible to observe that increasing the value of confidence the number of rules slowly decreased. The optimal value appeared to be between 70% and 80%. Association rules with these min values of *confidence* have been further filtered considering the value of *lift*. Indeed, rules with a *lift* higher than 1.0 are the ones positively correlated and so appeared to be the more interesting for the following explorations.

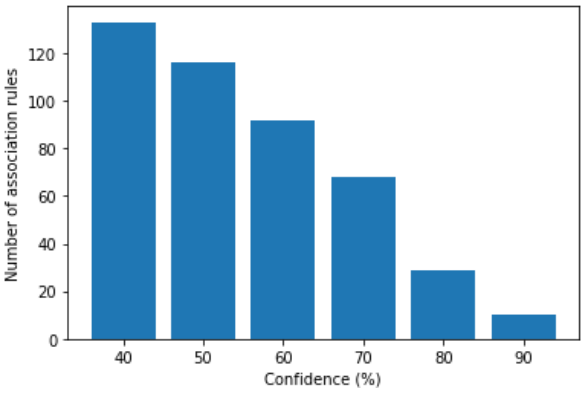


Figure 28 Number of association rules at varying of confidence.

Rules extracted are the ones with a *min_confidence* of 80% and a minimum value of *lift* of 1.0 and are shown in Table 16.

Association rule	Support (%)	Confidence (%)	Lift
('(6.0, 6.9]_Conc', 'not polysemic') => (6.0, 6.9]_Imag'	19.17	86	3.44
('(1.5, 4.7]_Fam', 'not polysemic') => 'medium'	19.39	80	1.10
('(1.5, 4.7]_Fam', 'not polysemic') => 'M'	20.48	85	1.09
('9-12', 'medium') => 'not polysemic'	35.62	96	1.04

Table 16 Extracted association rules with *min_conf*=80% and *lift* > 1.

From Table 16, it is possible to appreciate that the extracted rule with the higher value of *lift* (3.44) is also the more predictable one since it is intuitive that the more concrete words are also the easier to imagine.

Moreover, from the moment that the variable *Gender* presents more *M* values, it has not been possible to extract association rules to predict female gender with the *min_confidence* used in the previous extraction. For this reason, the *confidence* and the support values have been decreased respectively at 52% and at 10%.

Association rule	Support (%)	Confidence (%)	Lift
('(6.3, 8.6]_Val', 'medium', 'not polysemic') => 'F'	7.50	53	2.43
('(6.3, 8.6]_Val', 'not polysemic') => 'F'	9.95	52	2.37

Table 17 Extracted association rules to predict female words with *min_conf*=52 and *lift*> 1.

4.4 Prediction of target variable

Extracted association rules could be adopted for different task as the filling of missing value in the data or the label prediction for the target variable. From the moment that none of the attributes taken in consideration for the Pattern Mining analysis presents missing values, the more informative rules extracted have been exerted to predict word's gender.

Selected rules, chosen according to their *lift* and *confidence* values, are the following:

- *('(1.5, 4.7]_Fam', 'not polysemic') => 'M'*: for the male values (appliable to 1126 records in the dataset);
- *('(6.3, 8.6]_Val', 'medium', 'not polysemic') => 'F'*: for the female values (appliable to 654 records in the dataset).

TN: 351	**FP:** 167
FN: 303	**TP:** 959

Table 18 Confusion Matrix for results obtained through association rules.

Results obtained reached an *accuracy* of 0.735. As shown in the confusion matrix in Table 18, correct prediction for male words (*TP*) are higher than correct predictions for female words (*TN*). As discussed in previous sections, this displacement is due to the majority of the male words in the training data comparing to the female ones.

4.5 Final considerations

As resulted from the analysis returned from Pattern Mining task, the prediction of gender's words using the association rules does not return general meaningful results: it is not possible to define an objective rigorous rule to predict word's gender or, more precisely, it is not possible to apply these rules to all the existing words. Results provided from the explorations discussed in this section are reliable only in the context of this specific dataset but are not generalizable for other sets of words.
Moreover, rules extracted are bounded to the environment in which data has been collected: for what gender prediction is concerned, also the method of composition of the Glasgow Norms dataset influenced the partial results returned by the label prediction provided by association rules. Indeed, results have been gathered and inserted in the dataset from the answers of some questionnaires provided by more than twice women than men: 599 against 230 [2]. This displacement might have involved an impairment of the data from the moment that people might have suffered the influence of their gender in classifying a word as more associable to a male or a female meaning.

[2] SCOTT, Graham G., et al. The Glasgow Norms: Ratings of 5,500 words on nine scales. Behavior research methods, 2019, 51.3: 1258-1270. Pag. 4 (https://link.springer.com/article/10.3758/s13428-018-1099-3/tables/2)