Automatic detection of alcohol intoxication

ELEC-E5510. Speech recognition

Period II - 2021/2022

13/12/2021

Iria Durán Lusquiños

Student nº : 100087541

DNI: 53819555S

Javier Muinelo Monteagudo

Student nº : 100084638

DNI: 29222644V

Index

Introduction

The present project belongs to the course Speech Recognition D.

The task proposed is to identify automatically through speech if a person is inebriated or not. For that, we will compare different models of machine learning (classical models and neural networks). The approach taken is to understand the problem as a binary classification, using labels of *sober* and *intoxicated*. To train the models the data set being used is The Alcohol Language Corpus (ALC), that will be described in the following section.

One of the main challenges that it is being faced is, as in any paralinguistic task, to recognize in the audio some characteristics that go beyond the message being spoken: alterations on the speech, vocalization, accent… This causes a big problem, as the models will have to distinguish between alcohol intoxication and other kinds of state alterations, for example excitement, sadness, stress or anger.

Automatic alcohol intoxication detection can have interesting applications. On the one hand, it could be used as a driver's control software, in which the technology would detect through an audio input if the driver is inebriated or not, and from that point make some recommendations or even block the car if the models have enough reliability. On the other hand, it could be interesting to prove some points in the justice field. As the trials are done months after the crime, it is not possible to employ blood or breath controls to see if the person was intoxicated, so through videos and audios, these models could make a difference.

The code of the project is available in the repository
https://github.com/javiermuinelo/Automatic-Alcohol-Detection

Literature study

Recognizing emotion and state from speech may seem like a fairly simple thing, but can be very difficult for AI models. This is especially because the individualistic nature of human emotional expression means that in order for a model to be accurate over a large set of people, it needs to lie in a sweet spot between being precise enough to classify accurately and broad enough to avoid speaker-influenced errors (Miller 2018).

While there are some general speech pattern changes when a person consumes alcohol, including speech degradation and slurred speech, the same rules with emotions apply to intoxication because the severity and type of impairment varies from person to person greatly (Mal, Sharma, Kumar.

2014). Any data used should also account for different accents and ways of speaking to allow generalization.

Detecting the intoxication state of a person has many potential use cases, including detecting intoxication level from a phone call, creating a non-invasive method of determining if a person is drunk or preventing drunk people from driving vehicles or using online services while drunk.

Automatic alcohol intoxication detection is a challenging area due to several factors that must be studied carefully. The big challenging part of this task is trying to identify some paralinguistic characteristics, this means that the models have to focus their attention not on the messages that the locutor is producing, but on the emotions and sounds that the locutor is doing while talking. One of the main problems is the kind of feature selection and model.

Several studies have been carried out on how to deal with this problem. There are different approaches both in the extraction of characteristics and in the training of predictive models.

Mel-frequency cepstral coefficients have been used successfully in the features selection. Since a drunk person tends to make great variations in his voice, peaks are formed in log Mel Filter Bank (Mal, 2014) which allows us to identify this type of behavior. Although MFCCs are generally found to outperform formant features across many speech-based classification tasks, normalize hierarchical features (Bone,2011), phonotactic features such as prosodic events, phonotactics and phonetic (Biadsy,2011), prosodic features (phrasal units) (Levit,2002) were shown to be in the top feature set in a related paralinguistic. Other studies investigated changes in the waveform of glottal pulses estimated from speech by applying the Iterative Adaptive Inverse Filtering (IAIF) (Sigmund,2011). For the studies focused on glottal pulse features, the “excited” vocalizations due to alcohol can be confused with some kind of emotional manifestation or voluntary changes on the speech. For this type of approach, maybe it would be necessary to have some complementary training for differentiating these two possible options.

Regarding the models for the identification of alcohol intoxication in speakers, the Liljencrant-Fant (LF) approach (Sigmund,2011) stands out with a success rate of between 69.3% and 77.0%, and a fusion method based on systems using a set of acoustic supervectors Gaussian mixture model (GMM) (70.54%, precision) (Bone,2011). In addition, vector quantization (VQ) using the LBG algorithm is used when speech feature vectors are classified using k-means and identified by distance measurement using Euclidean distance. It has a 95% recognition with each of the 8 speakers (Mal, Sharma, Kumar. 2014). Other approximations have been obtained using a convolutional neural network (Miller 2018)(Levit,2002).

A totally different approach stands up from the others: it focuses on the drowsiness of a speaker, induced by alcohol intoxication or sleep deprivation (Zhang1, 2017). For this, both the Alcohol Language Corpus and the Sleepy Language Corpus were used, which allowed a considerable increase in training data. For the extraction of characteristics they used The ComParE set of supra-segmental acoustic features (prosodic (PROS), Mel Frequency Cepstral Coefficients (MFCC), spectral (SPEC), and voice quality (VQ) features) and a Support Vector Machines (SVMs) model was created with which yields up to 60.3% unweighted average recall were obtained, which is significantly above-chance (50%) and highly notable given the mismatch in the training and validation data.

Finally, it is also interesting to say that there are not many extensive datasets that could be used to train models for these tasks. One of the most remarkables is The Alcohol Language Corpus (ALC), a dataset comprising intoxicated and sober speech of 162 female and male German speakers.

Methodology

Data set

The Alcohol Language Corpus (ALC) is the data set used to solve this task (version 4). It is provided by the Bavarian Archive for Speech Signals (BAS) and consists of a set of recordings of German native people between 21 and 67 years old. The corpus was built under different motivations; first, the non-existence of a large enough data set of spoken recordings for any kind of speech recognition task; second, the ones that existed were really limited in terms of gender and age distribution. In relation with alcohol intoxication, previous data sets measured the percentage of alcohol in blood with mathematical approximations or with breath controls and any of these methods provide the accuracy of a blood control. That was the one used in ALC. Moreover, ALC wanted to provide some kind of spontaneous speech samples that were not present in previous corpuses.

ALC consists of speech recording samples of 162 people (77 females) in two standing automobiles to ensure a constant acoustic environment in both situations: sober and after drinking as much as they wanted. Each person has performed several recording types. On the one hand, read speech recordings, such as reading a digit string simulating a telephone number or a credit card, spelling some word, reading an address or a tongue twister. On the other hand, it has been tried to emulate spontaneous speech through picture descriptions, commands, dialogues and answering questions.

Each one of the ALC samples comes with an annotation file that contains several labels describing its conditions such as the following ones: alcoholized - non-alcoholized, gender, age group, drinking habits, breath alcohol content (BrAC), blood alcohol content (BAC), emotional state, speech type , content, etc.

Features extraction

The recordings cannot feed directly to the models since most of them expect numerical vectors with a fixed size. As mentioned above, the algorithm should not focus on the message of the audio but on its paralinguistic features. We extracted these features through the openSMILE toolkit (“OpenSMILE”).

The resulting feature set contains 4370 common acoustic low-level descriptors (LLDs). It includes MFCCs, log magnitude of Mel-frequency bands (MFBs), fundamental frequency (f0), energy, jitter, and shimmer, among other features. The final base feature set is produced by computing 'global' static functionals (e.g., mean, standard deviation) across each of these LLD streams. These huge amounts of characteristics led to several redundancies that were avoided by applying dimension reduction .

Train- test division

The resulting data frame was split into two datasets: one for training and the other , independent, for drawing final conclusions as well as for the selection of the best model . The sample was shuffled to make sure that the datasets reflect reality and the 80-20 division rule was applied . Notice that the training test was not divided also in the validation set as cross validation was used for searching for the best parameters for the models.

Dimension reduction: PCA

Having a high number of characteristics does not always improve the prediction since in many cases the variables explain the same variability. It only complicates the model's training. The unsupervised statistical method Principal Components Analysis condenses the information into a few variables, while preserving a significant part of its variability. Thus, all those original features that present a strong correlation and, therefore, measure the same thing from different points of view, will form a single latent variable called the main component.
Highlight the fact that the same PCA has to be applied to the test sample since the input data of the models must be the same.

Models

The goal of the project was finding out the best classification system for the task proposed in the study: binary classification of being drunk or not. In order to do so , different types of models were trained and tested .
First, two classic machine learning models were trained: Logistic Regression and decision tree.

- **Logistic regression** is a classification model that learns a linear relationship from the given data set and then introduces a nonlinear relationship with the logistic function. Specifically, the optimization problem is solved with the "liblinear" algorithm and the one-vs-rest (ovr) scheme is used for training. Furthermore, the inverse of the regularization force (C) was 0.4. It is needed to say that logistic regression models learn linear boundaries and struggle to find more complex relationships. Also, that tend to overfit the data if the number of features overpass significatively the number of samples, although this issue has been solved with dimension reduction. On the other hand, this model is easy and fast to train, so it can provide rapid results and give a general idea of how the samples are behaving.
- The **decision tree** model is a non-parametric supervised learning method whose objective is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the characteristics of the data. Specifically, it is a model of the maximum depth of the tree is 5 and minimum number of samples required to be at a leaf node is 4.

When decision trees are used, it is needed to take into account that they are sensitive to small changes in data, because they modify the structure of the tree and lead to instability.

On the other hand, two more complex deep learning models were implemented: our own neural network and a Multi-layer Perceptron classifier.

- The **neural network** used was simple. It was built with four layers, three of them with RELU activation functions. The last layer was feeded with the sigmoid activation function, as it is a recommended approach for binary classification tasks.
- The **multilayer perceptron** is an artificial neural network (ANN) made up of multiple layers, in such a way that it has the ability to solve problems that are not linearly separable. Specifically, stochastic gradient descent is used as an optimizer of the weights and a logistic activation function is applied ”. The random_stage was set as 1 and the maximum number of interactions was 100.

Finally we decided to apply two ensembles: bagging and boosting that combined different learning techniques being the resulting system much more robust than the models than the models separately . The classic models that have been combined are those with which the best results have been previously described.

- In **Bagging** (averaging method), you fit multiple models each with a different subset of the training data.
- **Boosting** (improvement method) consists of sequentially training multiple simple models so that each model learns from the mistakes of the previous one. En concreto un *AdaBoost* .

Metrics.

To observe and evaluate the behavior of the models, the following metrics will be taken into account:.

- **Accuracy**: precision of the trained model, that is, measure of all the correctly identified cases.

$$\mathrm{N} \qquad \mathrm{TN} + \mathrm{FP}$$

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{P} + \mathrm{N}} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

-

- **F1-score**. Weighted average of accuracy and recall, where a score reaches its best value at 1 and the worst score at 0.

$$\mathrm{F}_1 = 2 \cdot \frac{\mathrm{PPV} \cdot \mathrm{TPR}}{\mathrm{PPV} + \mathrm{TPR}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

-

It was decided to use these two metrics because accuracy is used when the True Positives and True negatives are more important while F1-score is used when the False Negatives and False Positives are crucial (https://medium.com/analytics-vidhya/accuracy- vs-f1-score-6258237beca2). In this way we can evaluate the model in both possible situations. Likewise, accuracy can be a somewhat dangerous metric when the sample is unbalanced because it does not detect whether the model tends to predict only the majority class. Instead, f1 gives a better measure of the incorrectly classified cases. practically balanced, so it is expected that for this aspect there is no difference between the two metrics.

Experiments

For extracting audio features, we used the tool OpenSMILE and a configuration file of one project that used the Alcohol Language Corpus before : *INTERSPEECH 2011 Speaker State Challenge.*
OpenSMILE is a software open source tool for audio feature extraction. The configuration file used allowed us to extract a feature vector per each wav file. Our part was simple: iterate along the database extracting the characteristics and merging everything in a single csv file that would allow us to preprocess and organize the data easily.

Once we had our features, we went through all the annotation files that were with the recordings to obtain if the speaker was drunk or not. However, some audios were the same but with different qualities . As they were not classified, they were not included in the present study. On the other hand, a third class called "cna" was found that corresponds to the control group. Thus, some volunteers were randomly selected to redo the drunk session, so they are intoxicated. They were replaced by "a" label.

The final dataset, after merging the OpenSmile result and the labels, had 15180 rows and 4370 columns. We split it into train (80%) and test (20 %) sets but first we shuffle the sample.

After that, we standardize the features as they were in different scales and have different units.Esto implica que aquella variable que posee una escala mayor que el resto será la que más aporte en el PCA , es decir, en la que se centrará el análisis . In this way, the size differences do not influence the PCA results.

We then conduct a study to determine the number of principal components of PCA. To do this, we graph the accumulated variability explained by each of the components as can be seen in graph 1.

In this case, we decided to select 1500 variables because, as we can see in the graph, it is the point where the curve flattens. It means that from this point, the rest of components tend to explain the noise in the data. Nevertheless, we all carry out the entire following training process for different numbers of main components. Initially we selected about 40 components since they were the ones that explained the most variability but only represented 60% of the total variability. We also tested with 1000 variables, but the best results were obtained with 1500. However, being such a high number it is impossible to explain which are the characteristics that most influence the analysis and which are not significant.

To finish the preprocessing of the data, it was necessary to transform the labels into 0: alcoholic and 1 nonalcoholic to be able to introduce it into the neural network.

Once we obtained the feature vectors to introduce the models, we carried out several experiments with the GridSearchCV function from the sklearn (python) library in order to find the selection of parameters that maximize the metrics. This algorithm automatically checks the different combinations to observe the general behavior of each model. To avoid overfitting using the test data, when adjusting the parameters of each combination, the Cross-validation function of the same library was used. This technique divides the train dataset into k subsets (k = 10 was used in this study). It uses k-1 folds to train the model and the resulting model is validated on the remaining part of the data. The performance measure reported is then the average of all the cases.

(https://scikit-learn.org/stable/modules/cross_validation.html) In this way, the results were more adjusted to reality.

Once we found the parameters of the models that best fit the data, we trained the models with all the training data and used the test data to compare the behavior of each model. It should be noted that we apply the PCA to the test sample since the input data of the models must be the same

.

Results

The results obtained are collected in the following table:

MODEL	ACCURACY	F1
Multilayer Perceptron	0.75	0.74
Logistic Regression	0.73	0.73
Adaboost - Log. Regr	0.73	0.73
Adaboost - Decision Trees	0.64	0.63
Decision Tree	0.64	0.60
4 Layer Neural Network	0.75	-
Bagging - Log.Regr	0.73	0.73
Bagging - Decision Trees	0.73	0.73

Note: All the confusion matrices can be consulted on the section *Annex* at the end of the report.

In it, we observe that both metrics present very similar results in all cases except for the decision tree. In this case, it is observed that the model tends to fail more in False Negatives and False Positives. Thus, from the confusion matrix, it is observed that the model tends to always predict that the subject is not drunk, so it is not a suitable model for the objective task. Furthermore, its predictions are little better than those made randomly.

On the other hand, we can highlight the behavior of those most complex models: neuronal network and multilayer perceptron for reaching 75% accuracy. Taking into account its confusion matrices, we observe that the neural network better predicts those cases of drunk subjects while the multilayer perceptron classifies non-intoxicated people with 0.8% of correct answers. Therefore, depending on the application, one model will be more interesting than the other. In the case of using it as an alert for the driver of a vehicle, it is better the case in which more false positives are committed than negatives, so the neural network would be the best model.

Regarding the assembly models, it should be noted how both improve the results of the decision trees, reaching an accuracy in both cases of 73%. As seen in the confusion matrices attached in the annex, the decision tree system better predicts those cases of intoxicated people (0.4 TP in Adaboosting and 0.5 in Bagging).

Conclusions

The results obtained so far are way better than random guessing, and below the current state-of-the art (around 80%). Studying each one of the machine learning methods, we have observed that some of them are more accurate in predicting one of the labels, and other ones are more accurate in predicting the other one. The use of several distinct approaches helps us to differentiate the way of learning of every algorithm used.

We contemplate the possibility of extending the study combining another meta-information (labels) that could increase the accuracy of the model. One possible approach could be using the BrAc and BAC values to clip the data with lowest percentages of alcohol and change it to non alcoholic people, increasing the margin between both labels. Another idea would be to make use of the labels of emotional state and drinking habits to find new relations and patterns, or to extend the task to other types of drug intoxication, although that would mean to make use of another dataset appropriate to the problem. Moreover, it would be interesting to work with the costs of classification, due to the fact that a false negative (prediction of non-alcoholic when it is actually alcoholic) could be dangerou

Automatic alcohol intoxication detection is a paralinguistic task that has not yet reached enough reliability to be applied in contexts as the ones mentioned in the introduction, although it is interesting to see how raw signals can be transformed into mathematical objects in order to identify emotions or physical states.

Division of the labor

The work distribution has followed the initial plan, although the training phase and presentation were done in the last days due to technical problems with the feature extraction. Concretely, as a team we have prepared the literature study from the 3rd of november to the 11th. Also the presentation and report were done collaboratively. Talking about individual tasks, Javier Muinelo has been in charge of dealing with the raw data set, extracting the features, organizing them and merging them into a single file to train the models. Iria Durán has led the project, deciding and designing which models to train, as well as performing the dimension reduction in the preprocessing part.

Acknowledgments

We appreciate the time and help of *Dejan Porjazovski* for his constant follow-up week after week.

References

[1] Risha Mal, R.K. Sharma, Naveen Kumar (2014). Intoxicated Speech Detection using MFCC Feature Extraction and Vector Quantization.
[https://www.ripublication.com/irph/ijeee_spl/ijeeev7n3_11.pdf]

[2] Joshua Miller, Jillian Donahue, Benjamin Schmitz 2018. Speech emotion and drunkenness detection using a convolutional neural network.
[http://www2.ece.rochester.edu/~zduan/teaching/ece477/projects/2018/JoshuaMiller_JillianDonahue_BenjaminSchmitz_ReportFinal.pdf]

[3] Sigmund, Milan & Zelinka, Petr. (2011). Analysis of voiced speech excitation due to alcohol intoxication. Information Technology and Control. 40. 10.5755/j01.itc.40.2.429.
[https://www.researchgate.net/publication/228755948_Analysis_of_voiced_speech_excitation_due_to_alcohol_intoxication]

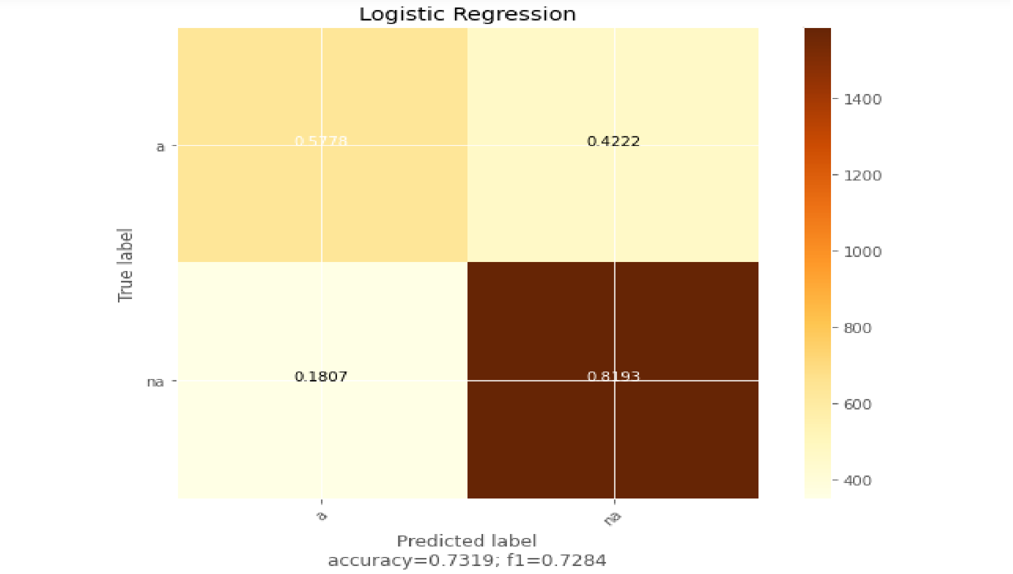
[4]Bone, Daniel & Black, Matthew & Li, Ming & Metallinou, Angeliki & Lee, Sungbok & Narayanan, Shrikanth. (2011). Intoxicated Speech Detection by Fusion of Speaker Normalized Hierarchical Features and GMM Supervectors.. Proceedings of Interspeech. 3217-3220.
[https://www.researchgate.net/publication/221490104_Intoxicated_Speech_Detection_by_Fusion_of_Speaker_Normalized_Hierarchical_Features_and_GMM_Supervectors]

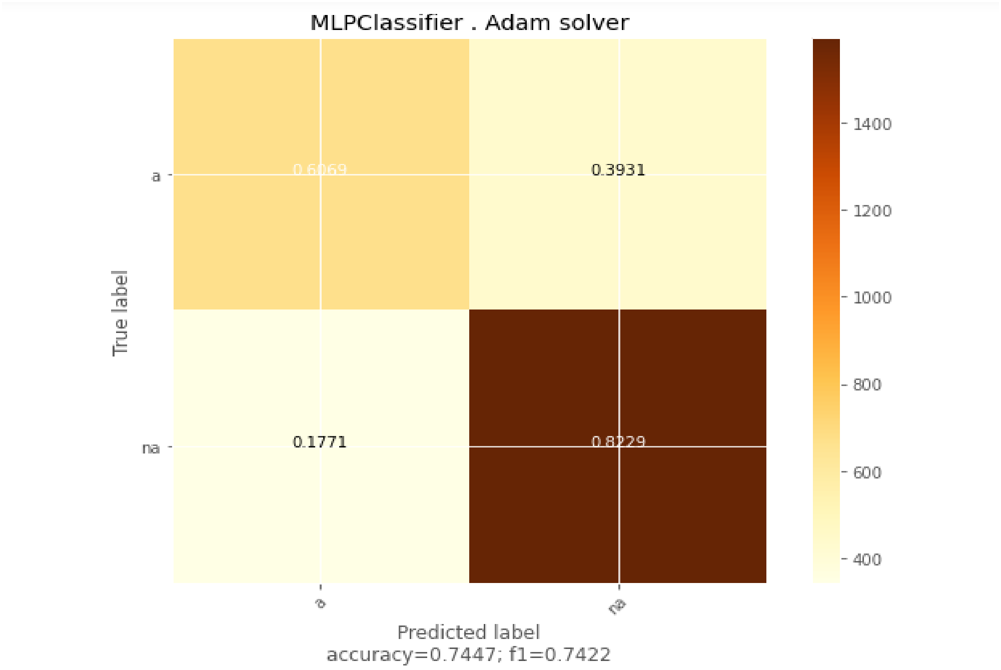
[5] Biadsy, Fadi & Wang, William & Rosenberg, Andrew & Hirschberg, Julia. (2011). INTERSPEECH 2011 Intoxication Detection using Phonetic, Phonotactic and Prosodic Cues. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 3209-3212.
[https://www.researchgate.net/publication/221479488_INTERSPEECH_2011_Intoxication_Detection_using_Phonetic_Phonotactic_and_Prosodic_Cues]

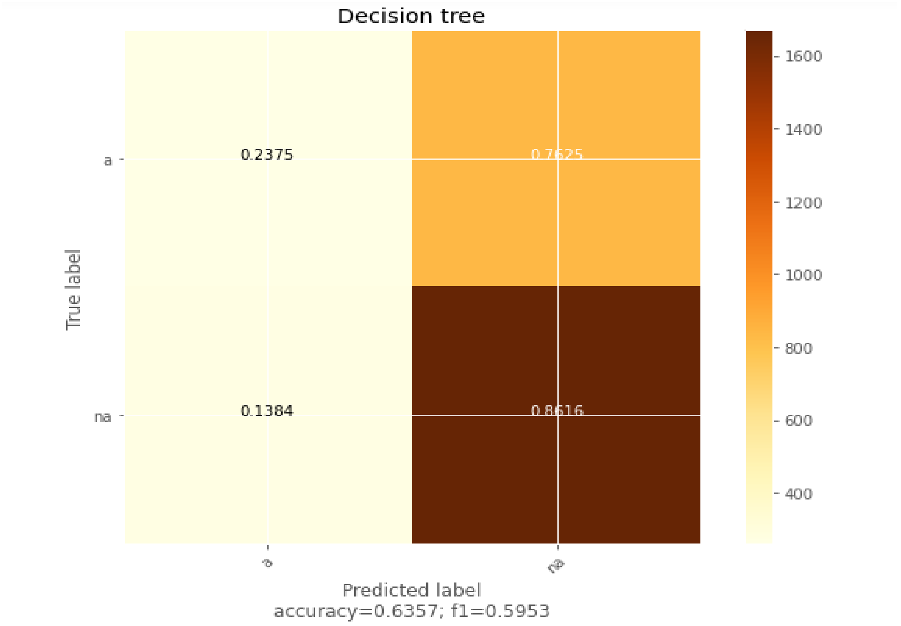
[6]Levit, Michael & Huber, Richard & Batliner, Anton & Noeth, Elmar. (2002). Use of Prosodic Speech Characteristics for Automated Detection of Alcohol Intoxication. Proceedings of the Workshop on Prosody and Speech Recognition 2001. [https://www.researchgate.net/publication/2522570_Use_of_Prosodic_Speech_Characteristics_for_Automated_Detection_of_Alcohol_Intoxication]

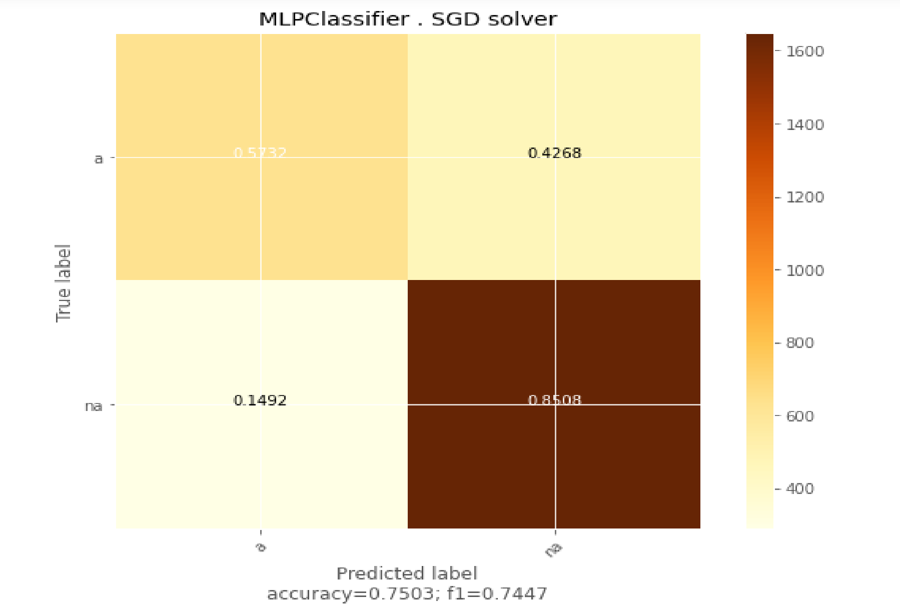
[7] Yue Zhang1, Felix Weninger2, Bj¨orn W. Schuller. (2017). Cross-Domain Classification of Drowsiness in Speech: The Case of Alcohol Intoxication and Sleep Deprivation. Department of Computing, Imperial College London, London, U.K.2Nuance Communications, Ulm, Germany [https://www.isca-speech.org/archive_v0/Interspeech_2017/pdfs/1015.PDF]

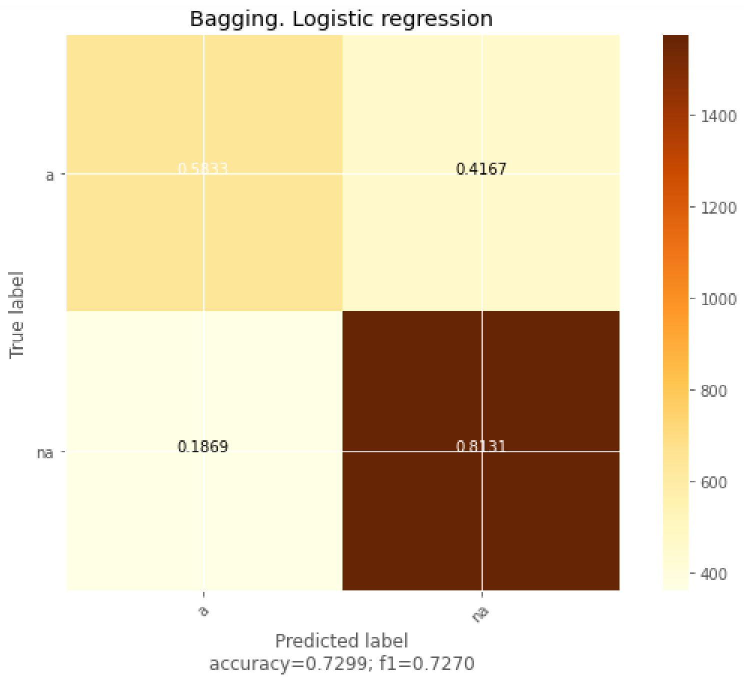
Annex

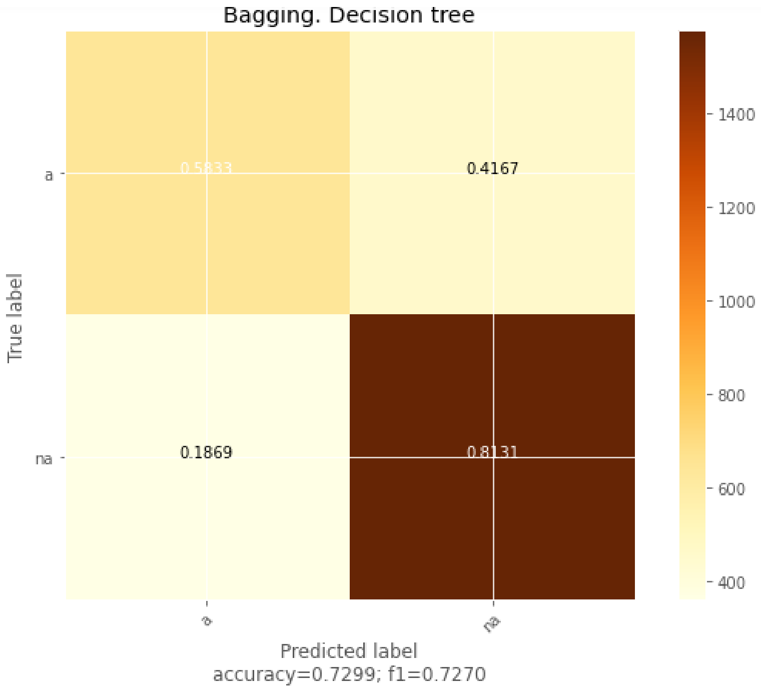


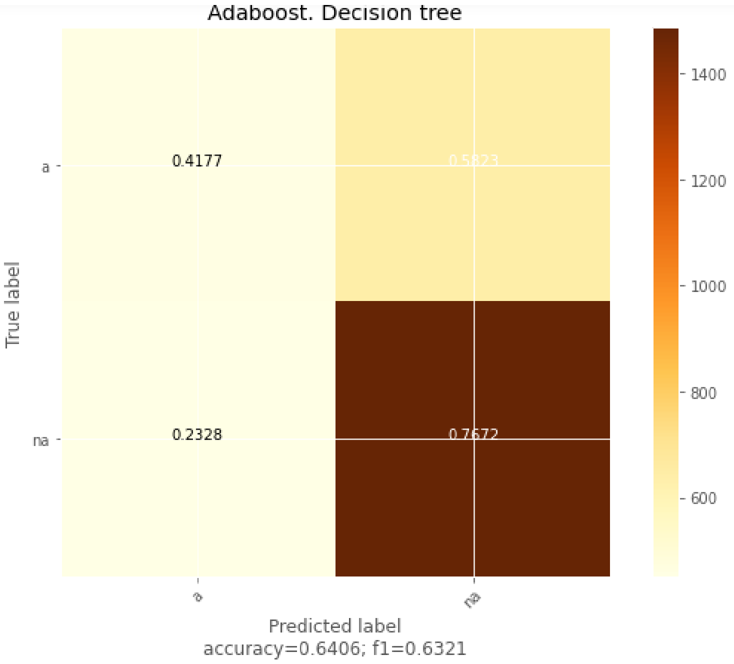

MLPClassifier . Adam solver
True label
a
na
0.6069
0.3931
0.1771
0.8229
a
na
Predicted label
accuracy=0.7447; f1=0.7422
1400
1200
1000
800
600
400


Decision tree
a
na
True label
0.2375
0.7625
0.1384
0.8616
a
na
Predicted label
accuracy=0.6357; f1=0.5953
1600
1400
1200
1000
800
600
400

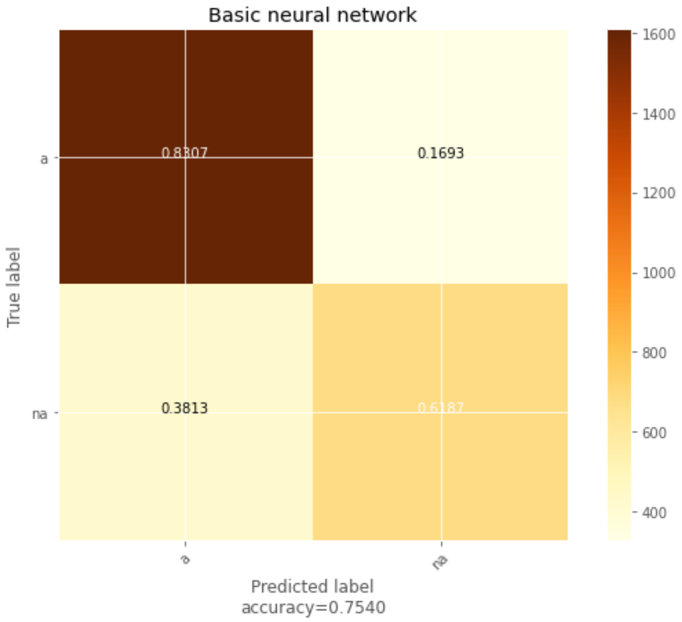

MLPClassifier . SGD solver
a
na
True label
0.5732
0.4268
0.1492
0.8508
a
na
Predicted label
accuracy=0.7503; f1=0.7447
1600
1400
1200
1000
800
600
400


Bagging. Logistic regression
True label
a
na
0.5833
0.4167
0.1869
0.8131
a
na
Predicted label
accuracy=0.7299; f1=0.7270
1400
1200
1000
800
600
400


Bagging. Decision tree
True label
a
na
0.5833
0.4167
0.1869
0.8131
a
na
Predicted label
accuracy=0.7299; f1=0.7270
1400
1200
1000
800
600
400



Adaboost. Decision tree



Basic neural network