IML – Hackathon

Benzaquen Rebecca 340938372, Dan Boujenah 341339901

We have chosen to solve the Task1. In the pre-processing step, we decided to choose only letters and numbers in the tweet text and to delete all other characters in order to focus the learning algorithm to significant words.

After doing the first stage, we transformed the given data, texts (of tweet) to a feature vector. For this target, we used the countVectorize function that changes all the data to make it more manipulable by the algorithm. Here, we also test two functions and choose the best of them, for example TfidfVectorizer.

The next step is to split the data into training and test set. We want to output a number between 0 and 9 that represents the celebrity who posted the tweet. We tested many partitions between the training and test set and the we take finally 0.27 of the data for the test set. Graph n°1

We after searched for the best algorithm for solving the problem. We thought to use Random Forest method, and the rate of success was 70% that is worse than with the present algorithm.

At this step, the result of the algorithm was 77%, with the MultinomialNB classifier.

Next, we thought how to improve this result, We analyzed when our classifier wrong:

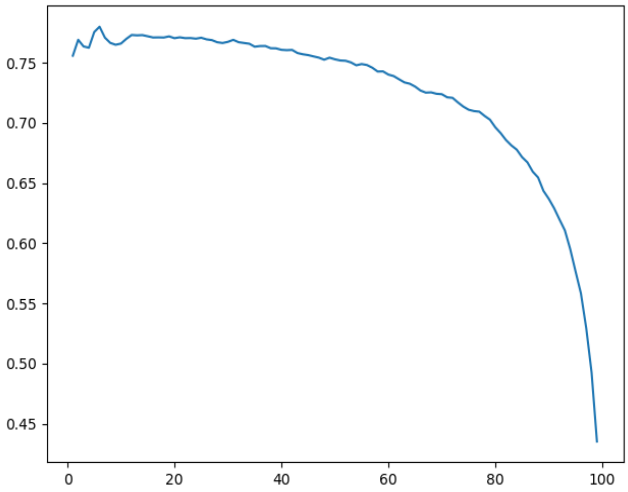
When 0 false	[0, 46, 83, 16, 23, 13, 33, 5, 67, 26]
When 1 false	[4, 0, 12, 2, 2, 2, 7, 0, 8, 4]
When 2 false	[4, 8, 0, 13, 18, 29, 24, 9, 32, 11]
When 3 false	[32, 13, 67, 0, 144, 58, 67, 78, 104, 40]
When 4 false	[1, 0, 15, 19, 0, 16, 15, 17, 21, 5]
When 5 false	[3, 3, 24, 11, 17, 0, 10, 23, 21, 9]
When 6 false	[6, 8, 31, 20, 25, 21, 0, 20, 40, 21]
When 7 false	[4, 5, 10, 5, 7, 20, 20, 0, 5, 14]
When 8 false	[13, 2, 34, 35, 17, 6, 14, 5, 0, 8]
When 9 false	[11, 26, 35, 23, 21, 36, 15, 24, 35, 0]

So we thought to analyze the probability of predictions for each personality, and we build a matrix of all the probability for each tweet and each personality. And add a column of tweets.

We perform a RandomForestClassifier on this data, such that features are the probabilities and label personalities.

Thanks to this second classifier we get a accuracy of 90%.

Graph n°1 :



This graph represents the percentage of error according to the percentage of test into the data.