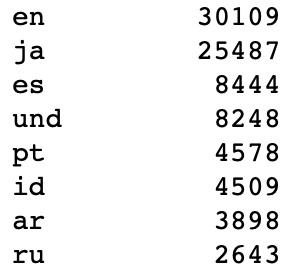
Lab report David Bielik and Debora Beuret

# Data

The data we were given was not yet it its final form. There were more labels than tweets in total. For this reason, the methodology we followed was very simple: create pandas dataframes for both datasets (the labels and the tweets) and then merge them with the IDs. With the module pandas, the process was quick and efficient.

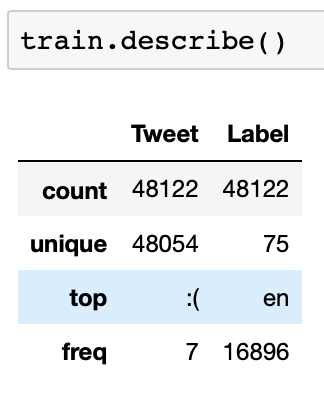
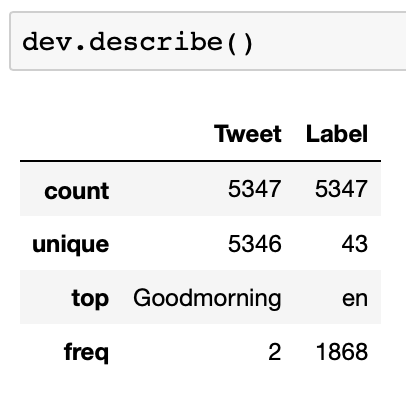
## Properties of the data



The frequency of the languages was not uniform. Almost half of the tweets were written in either English or Japanese. There were a lot of outliers with only a single occurrence in the whole dataset which created some **class imbalance**. Taking into account the distribution of the languages, we have decided to put all of the low frequency labels into one class. The ‘unknown’ class. More specifically, the threshold value to determine inclusion in this class is 10 (i.e. there need to be at least 10 tweets in one language).

Another option would be to simply remove all of the low frequency labels from the training set.

After thoroughly inspecting the labels, we have found that some labels were enclosed in **unnecessary whitespace**. For example, labels ‘ar ’ and ‘ ar ‘ are presumably describing the same language. However, the classifier would consider these as different labels. Thus, we have simply removed all whitespace from every label.

The final training set shows that there are there are 48122 instances in total. Out of those, only 48054 tweets are unique, which means some tweets must have appeared multiple times in the dataset. Those are presumably re-tweets, short statements or simple replies (such as ‘:(‘).

# Methodology

To choose the best performing model, the following methodology was chosen: the test set was set to fit the size and content of the file labels-test.stv (excluding the deleted tweets).

The preprocessed labels-train+dev.tsv set was first shuffled and then splitted: 90% for training set, 10% for development/validation set.

From there, we performed several **GridSearches** (cv = 4) with various parameters on the train set. The top three performers of the GridSearch would be tested with the development set. Since the model had not “seen” this data yet, a poorer performance of the best model would prove the model had overfitted the training data. It also allowed to make sure the number 1 ranking model was really the best.

Our final model would be the one that performed best on the development set. This final model would then be trained on the entire training set (90% train + 10% dev) to not lose any valuable information. At last, we measured and reported its performance on the test set.

# Part 1 – Language identification with linear classification

## Pipelines

We considered quite a few pipelines. Obviously, **ngrams** and **tfidf** came into question and were used in our pipeline. Additionally, we tried adding the **average word length of the tweets** (see class AverageWordLengthExtractor). Initially, it had a negative impact on the performance of each classifier! The solution to this problem was to use a MinMaxScaler on the average word length values. This scaler would scale down the values to a range of [0, 1] – performing a feature scaling technique called normalization. Another side-effect of using this scaler was an obvious speed up in training times in each classifier.

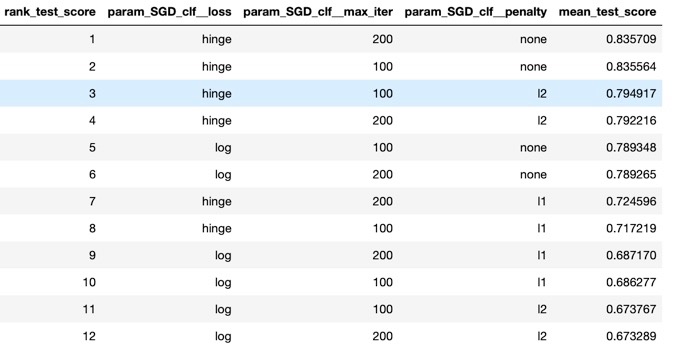
We also considered **removing stop words**, but it probably wouldn’t have been a good idea as they are probably important in the classifier, allowing it to identify a given language. We, for a minute, considered lemmatizing/stemming the words, but since those modules rely on already knowing what language is used, this option soon proved impossible. For this reason, we have a limited number of features.

## SGD Classifier

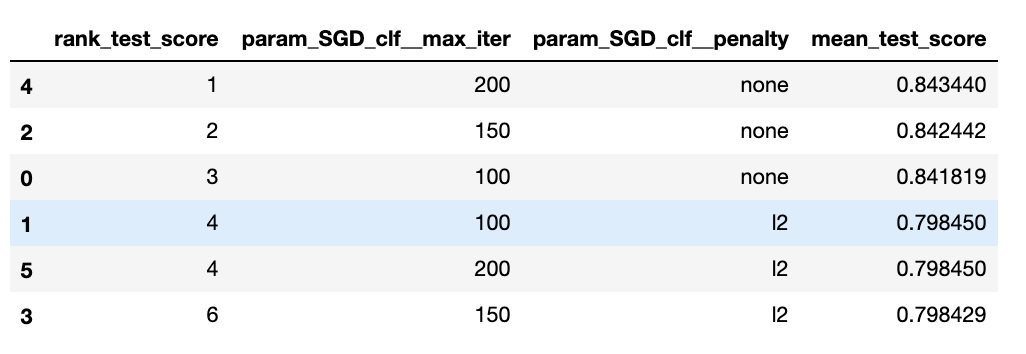
The parameters we tested were those described in the exercise: loss function, regularizations and number of iterations.

In a first GridSearch, we tried a few parameters, which resulted in this ranking.

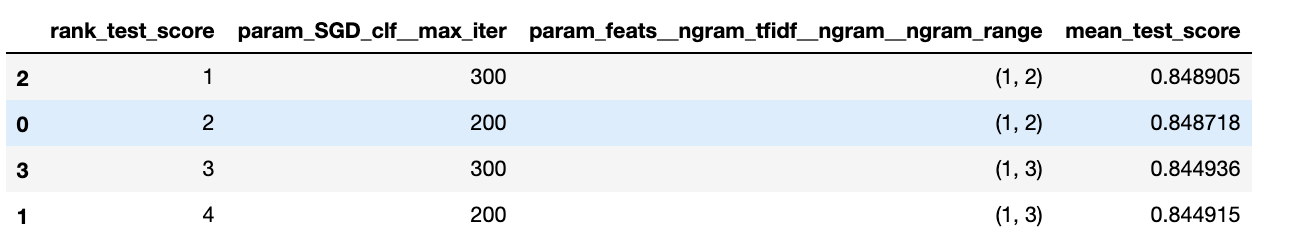
Class imbalance was dealt with by replacing the rare label (frequency: <10) by an ‘unknown’ label.



We realized that the best models would consistently have a hinge loss function. The number of max\_iterations would vary and both no penalty and L2 seemed to provide good results. For this reason, we trained a new GridSearch that always had a hinge loss function. The varying parameters were the penalty (None or L2) and the maximum number of iterations (100, 150 and 200).



In this case, we saw that all the best models had no penalty. Seeing the best models, we see that the best has maximum 200 iterations, then 150 and finally 100. This made us wonder if even more than 200 would do even better, or if it was a sweet spot already. For this reason, we tried another GridSearch with a constant hinge function, no penalty. The maximum number of iterations could vary between 200 and 300. We also chose to work with one more parameter, namely the ngram range. We decided to aim for both (1, 2) and (1, 3) to compare the results.



The results showed that the best model was having a maximum of 300 iterations and an ngram range of (1,2). It was outperforming the previous models.

In a last attempt to see if 300 was our “sweet spot” in terms of maximum iterations, we did a pipeline (no GridSearch) which had 400 iterations and performed very slightly under the winner model shown above.

**For this reason, we chose the model with the following parameters:**

**loss function: hinge**

**regularization: none**

**# of iterations: 300**

**ngram\_range: (1, 2)**

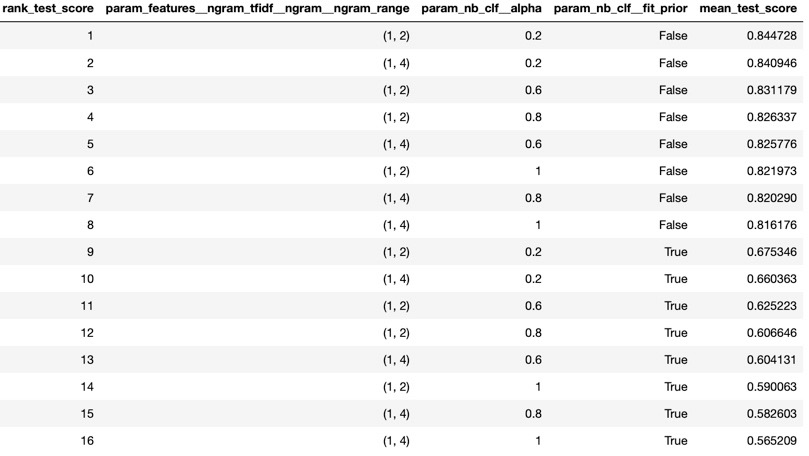
**Accuracy obtained on average: 85.6%**

## Multinomial Naïve Bayes Classifier

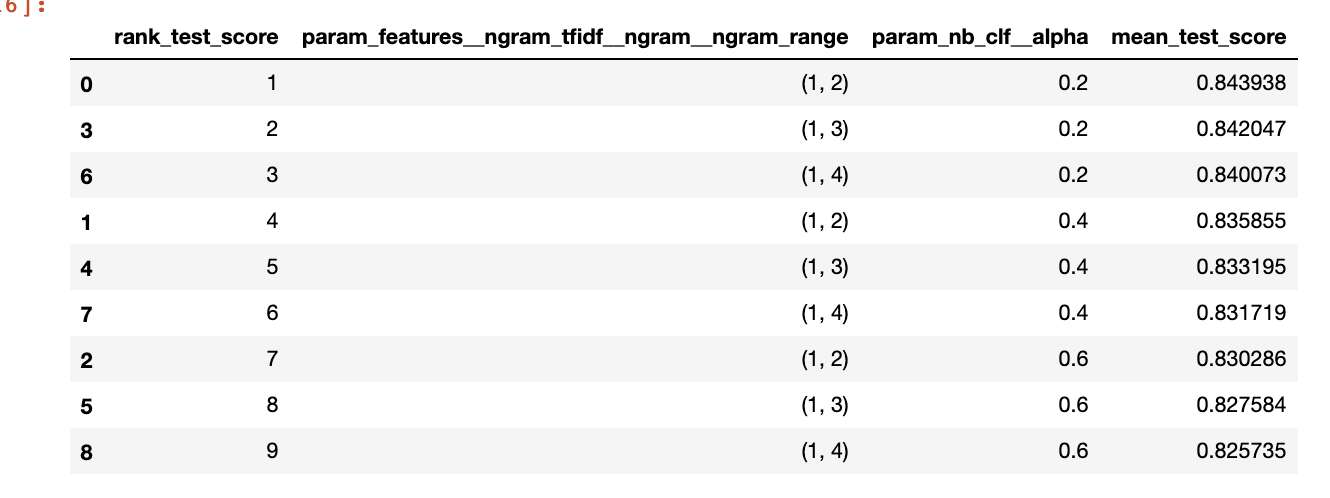
The first parameters we tested were:

* alpha: 0.2, 0.6, 0.8, 1.0
* Fit\_prior: True, False

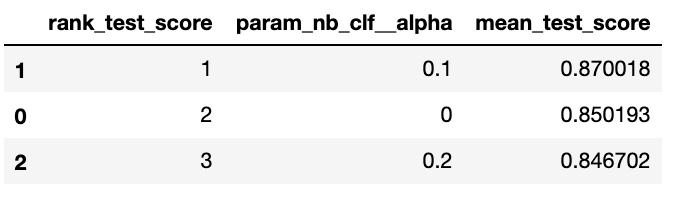
Resulting in this:



We then decided to try a few more models and created a new GridSearch that would always have a fit\_prior set to False, since all the best performers showed that. On the other hand, there was some more play with the ngram range: (1, 2), (1, 3) and (1, 4) and the alpha was varying between 0.2, 0.4 and 0.6. We can see that the best performing model, interestingly, is still the best (although its mean\_test\_score is slightly lower).



Since NB is not computationally expensive, a last test was created, this time focusing on the alpha: since all the good performers had the lowest one of the list, it was legitimate to wonder if even lower would do better. For this purpose, the GridSearch was made with alpha as only varying parameter (0.0, 0.1 or 0.2).



The results show that 0.1 (and even 0) perform better than 0.2. Thus, 0.1 was chosen as final alpha parameter.

**The best parameters thus were:**

**alpha = 0.1**

**fit prior = False**

**ngram\_range = (1, 2)**

**They resulted in an accuracy of 0.876152245019328 on the test set.**

best\_SGD4 = Pipeline([

('feats', FeatureUnion([

('ngram\_tfidf', Pipeline([

('ngram', CountVectorizer(ngram\_range=(1, 2), analyzer='word')),

('tfidf', TfidfTransformer()),

])),

# second feature

('ave\_scaled', Pipeline([

('ave', AverageWordLengthExtractor()),

('scale', MinMaxScaler())

]))

])),

('SGD\_clf', SGDClassifier(loss='hinge', max\_iter=300, penalty=None ))# classifier

])

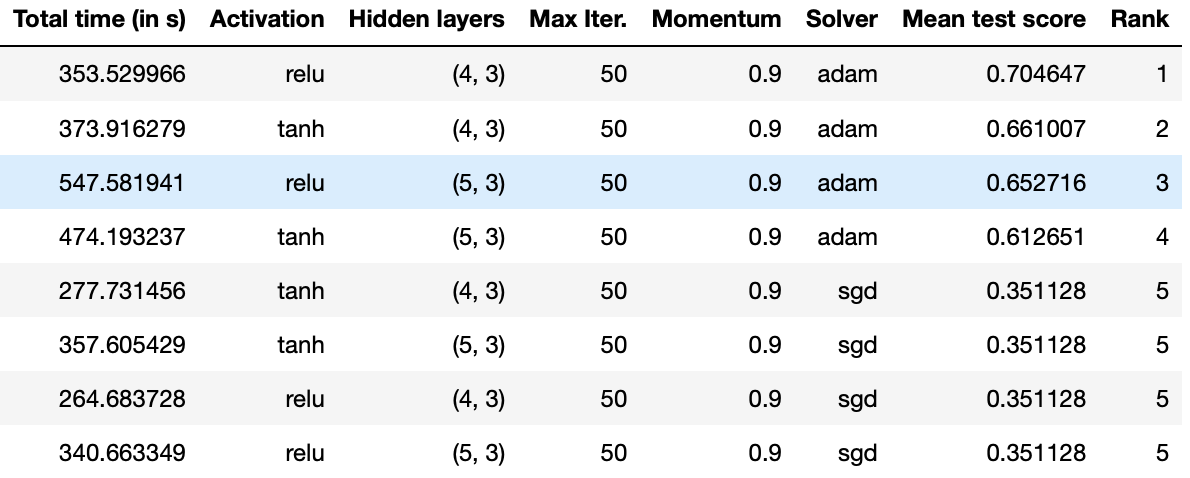
best\_SGD4.fit(X\_train, y\_train)

y\_SGD4 = best\_SGD4.predict(X\_dev)

accuracy\_score(y\_SGD4, y\_dev)

# Part 2 – Multilayer Perceptron (MLP)

We have used the same features for MLP as for the other classifiers. Parameter-wise we have tested different hidden layer sizes, activation functions, solvers, iterations and momentum values. The initial tests were perfomed with crossvalidation via gridsearch. First thing that we noticed were the long execution times. We had to let the gridsearch run overnight in order to get some results. After some discussion we have discarded cross validation and reduced the layer size from (100,) to (5,3) and (4,3). The max iteration times were also reduced to 50. The next gridsearch was performed with a cv value of 3 resulting in this table:



We have concluded, that the next set of tests should be done on a single hidden layer ranging from 3 to 5 hidden neurons without cross validation. The best performance (albeit the slowest, ca 20-25min) was observed on the 5 hidden neurons. Afterwards we have played with the alpha regularization parameter which we have decided to leave at 0.1.

**The best parameters (with a reasonable execution time) were:**

**hidden\_layer\_sizes=(5,),**

**activation='relu',**

**solver='adam',**

**max\_iter=50,**

**momentum=0.9,**

**alpha=0.1**

**Average accuracy on the test set: 78.5%**