Report Kaggle Competition

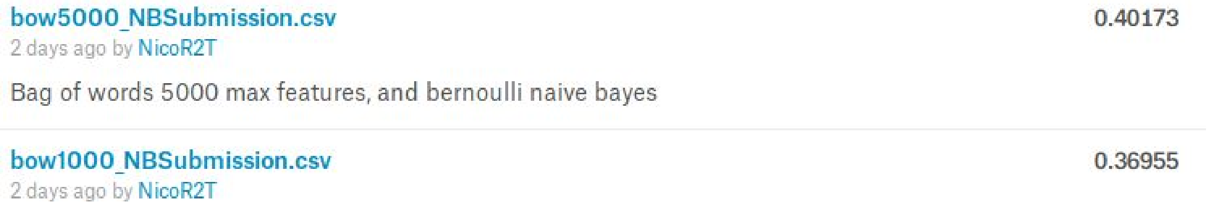
Nicolas TIREL - X1085033

I) Loading data

To load the data, I had to load the data given with the competition. All the tweets are in a json file, and the identification of the data is in a csv file. It was a little bit complicated at first to understand how to manage a very huge file (more than 1 million lines in the json file) in a low computation time, but I found the solution by indexing the data identification with the ID of the tweet, which increases a lot the computation time. I decided to store the data in a pickle file to increase the time of loading for the next time.

II) Naive Bayes classifier

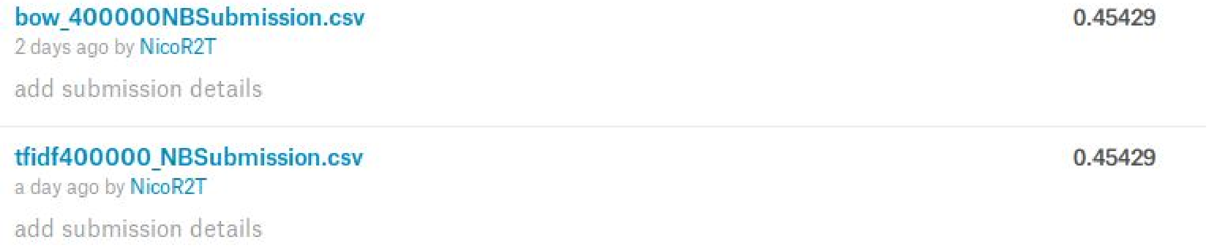
To obtain the best score, I tried first to use Bag Of Words (BOW) and Term Frequency - Inverse Document Frequency (TF-IDF) vectorizers without any pre-processing to see if the score is interesting. With those vectorizers, I decide to use the Naive Bayes classifier applied with Bernoulli equation because I had better result in the AI Cup with this classifier. For the first submissions, I used only 1000 and 5000 features for the Bag Of Words, and I see that the score increases with the 5000 features.



bow5000_NBSubmission.csv 0.40173
2 days ago by NicoR2T
Bag of words 5000 max features, and bernoulli naive bayes

bow1000_NBSubmission.csv 0.36955
2 days ago by NicoR2T

That's why I decided to try again until I reach the best score. When I had reached my best score with 400.000 features (out of 1 million features), I tried with the TF-IDF vectorizer and I notice that the score is exactly the same with TF-IDF or BOW vectorizer.



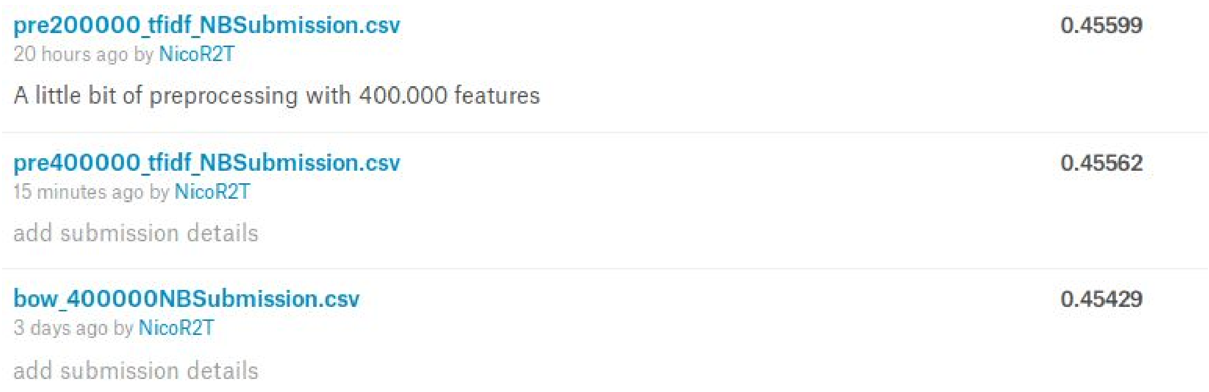
bow_400000NBSubmission.csv 0.45429
2 days ago by NicoR2T
add submission details

tfidf400000_NBSubmission.csv 0.45429
a day ago by NicoR2T
add submission details

At this point, I decided to try some pre-processing techniques to improve my result, and applied four steps:

- Convert all tweets to Lowercase
- Remove special characters
- Remove repetitions (replace "noooooooo" by "no")
- Remove stop words

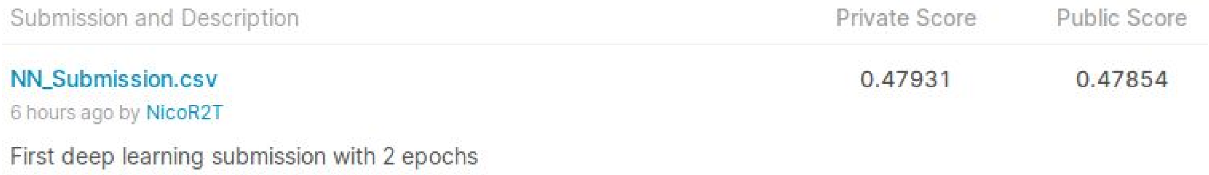
And I upload my submission to the kaggle competition and see that the result get better but it was not a very good improvement. I will keep it as an input for the deep learning algorithms.



pre200000_tfidf_NBSubmission.csv 20 hours ago by NicoR2T A little bit of preprocessing with 400.000 features	0.45599
pre400000_tfidf_NBSubmission.csv 15 minutes ago by NicoR2T add submission details	0.45562
bow_400000NBSubmission.csv 3 days ago by NicoR2T add submission details	0.45429

III) Deep learning

For this part, I decided to use the pre-processing techniques, bag of words vectorizer and the fitting model provided by the keras library. The input layer was the tweets after pre-processing, the 1st and 2nd hidden layers used the dense and relu functions from keras, and finally the output layer uses the softmax function. I use 2 epochs and a batch size of 32, which took me a long time to compute but finally, the result was better than with the Naive Bayes or Decision Tree Classifier.



Submission and Description	Private Score	Public Score
NN_Submission.csv 6 hours ago by NicoR2T First deep learning submission with 2 epochs	0.47931	0.47854

IV) Conclusions

For this competition, I prefered to submit as much as possible a lot of different tests with different size and different features in order to know which works the best. As I expected, my first tries using classifier was worse than with deep learning, which must be logic because the neural networks has usually better performance to treat that kind of data. I didn't find a big differences between the bag of words and the TF-IDF vectorizer, maybe because tweets are limited in number of characters and that means it's difficult for TF-IDF to obtain a better score than bag of words, and I found that the Naives Bayes classifier with Bernoulli was a little bit better than Decision tree. This competition can be useful know to understand better how to implement data mining technique to real dataset and how to classified texts into labels, which will be very useful for the final project competition.