Machine Learning – Programming Assignment

I affirm that this report is the result of my own work and that I did not share any part of it with anyone else except the teacher.

**Marko Pandža**

Given a data, which was already distributed into 3 parts: training (10k), test(1k) and validation data(1k). Data consisted labels (business, health, entertainment, science), news titles, and a publisher. The task was to label each news article corresponding to a topic(label). This was multiclass (4) problem.

The process of the program:

1. Read data and make a vocabulary by using nltk tokenizer on publisher and titles
2. Remove stopwords from vocabulary
3. Perform lemmatization and remove duplicates
4. Save vocabulary as a file
5. Load the data and vocabulary
6. Change classes (business, health, entertainment, science), into integers (0,1,2,3)
7. Save the change into file
8. Add publishers to title
9. Lemmatize the data and make a BoW representation
10. Stack it and save it
11. Load the data and perform training (Multinomial NB) on training data
12. Check the accuracy on all 3 data sets
13. Plot confusion matrix

In the end with 94.95% training accuracy, 90.8% testing accuracy and 90,4 validation accuracy it should be decent result for a 4-classification problem using MNB model. I’ve tried: SVM OvO, SVM OvR and perceptron, however I decided on MNB, because SVMs were taking extremely long. Confusion matrix is a good visual representation of the results, which was also needed instead of accuracies. Figure 1. shows a CM of training data. When it is misclassified, vast majority of times it is classified as science. The best predicated class is entertainment with nearly 97%. The worst predicated class is business with nearly 7.5% error rate. Most of times it was misclassified as entertainment and almost equally probable with health.

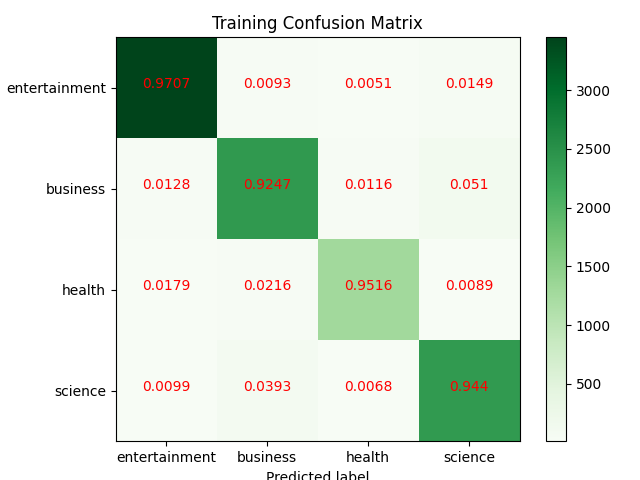


Figure 1 CM of Training

Next figure shows CM of test set. Again, the similarity is that entertainment is the class that is the most accurately predicted class. Also, a lot of times when it was misclassified it was classified as science. However, the class which had the worst accuracy is science with nearly 14,4% of an error rate. Mostly it was misclassified as business.

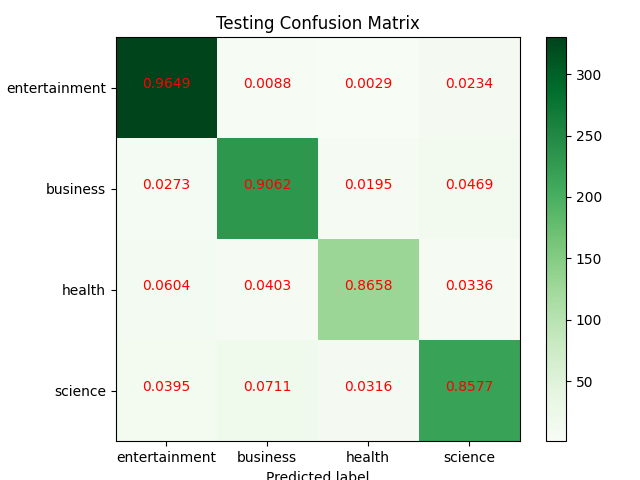


Figure 2 CM of Testing

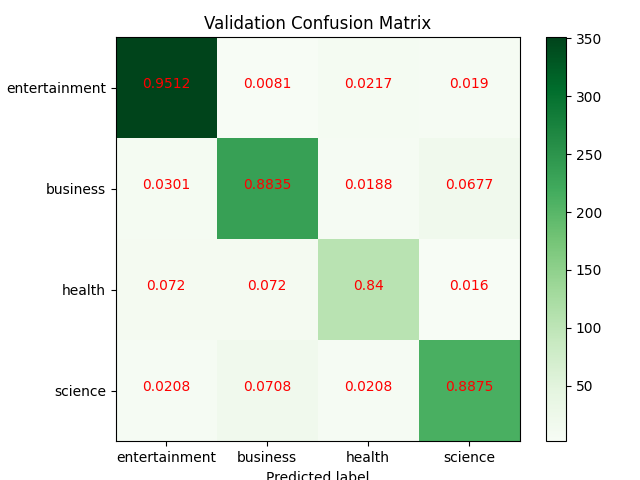
Lastly, the figure 3 shows CM of a Validation set. Where the best set again is the entertainment. The most misclassified class is health. It is often misclassified as entertainment.  
  


Figure 3 CM of Validation

After looking at the data and the results a conclusion can be made. Of course, a better model can be made with even better data representation, but generally the data is very misleading. Entertainment has very good accuracy. The data have a problem, because it was classified by a human and a human has a certain error rate and then on the data that already has an error rate a model is trying to learn. If it was classified by another human, it wouldn’t have the same labels. Let’s look at the examples. The sentence from test set “T. Rex Arrives At Smithsonian National Museum Of Natural History” was labeled as science. If I were to label this it would be labeled entertainment, because they made a hologram or a sculpture of a T.Rex and it’s in a museum, at least for me it doesn’t make a lot of sense it being labeled as science. Another example a title from train data: “S.Korean stocks rise for 3rd day on Samsung Elec rally, won flat” is labeled as health. I’m not sure why it is labeled as health, because I’m sure it should be business. Depending on where data is, model is either learning wrong from wrongly labeled data, or getting “good” results but it is thought of wrong, since its labeled incorrectly.

Conclusion: the model and the data representation and preprocessing could be better as always, but the data here, more precisely, the labels are distributed by humans, which is damaging for the model.