Master thesis : Fake news detection using machine learning

Review 2 Draft

Simon Lorent

Academic year 2018 - 2019

1 Models

1.1 Model comparaison

Since the last review I've introduced a few new models in addition of the Nave-Bayes classifier. These new models are decision tree classifer, linear svm classifier and ridge classifier. I have also tried to use lasso classifier and non linear svm but these medols complexity increase too fast with respect to the sample size and thus are not possible to terminate in a reasonable amount of time. It exists libraries that run svm on GPU but I have not manage to make it works yet. In order to avoid the problem of over represation of two single domains, I've chose to discard them in the following analysis and to use them as test set.

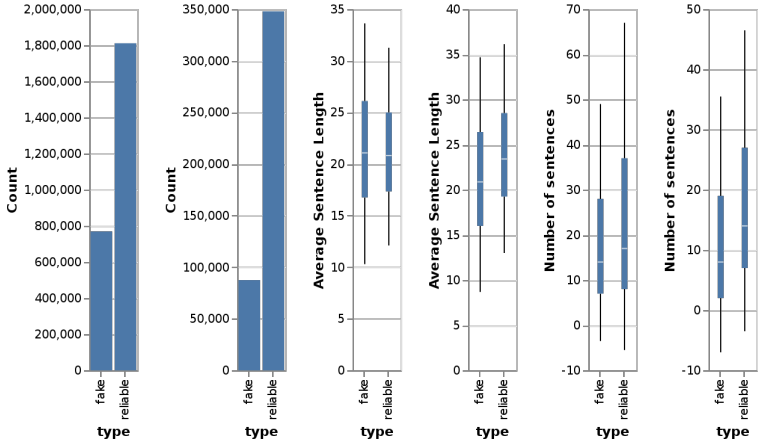

Figure 1: Summary statics on dataset, with and without downsampling.

The linear svc shows a good improvement over the basic Nave-Bayes. I've also try to limit the maximum number of features of the TF-IDF model and it also allows to have some improvement. This can be seen at **Figure 2**.

We can see that Nave-Bayes has the worst recall for fake news but it is the one that perform the best on reliable news.

1.2 Linear SVC

Linear SVC is a special case of SVM that fit linear model and is a lot faster than traditional SVM. **Figure 6** shows the recall with respect to the paramter C, which is defined as Penalty parameter of the error term. It can be seen that this parameter does not have a lot of influence on precision or recall. It should be noted that these values are the ones of the 3-folds cross validation and not from the validation test score. See **Section 1.1** for the performances on the validation set.

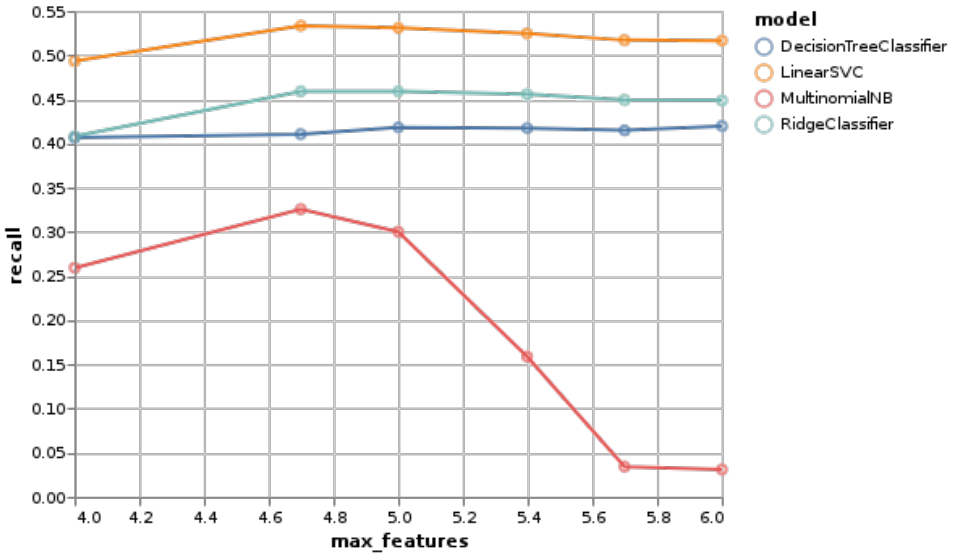

Figure 2: Model comparison with respect to the number of TF-IDF features (log scale) on fake.

2 SMOTE

As opposed to primilary results, SMOTE method does increase the recall for fake detection, but on the other hand is lowering the one for reliable detection, it is also lowering precision. This can be seen at **Figure 8** to **11**.

3 LDA

I have made some experiment with LDA analysis but this does not show any interesting facts, I plan to run on numerous LDA topics numbers but I need some way of mathematicaly compare the two distribution in order to find the optimal number of topics. For instance, **Figure 13** shows the distrubtion of fake and reliable news for 25 topics and **Figure 12** for 10 topics.

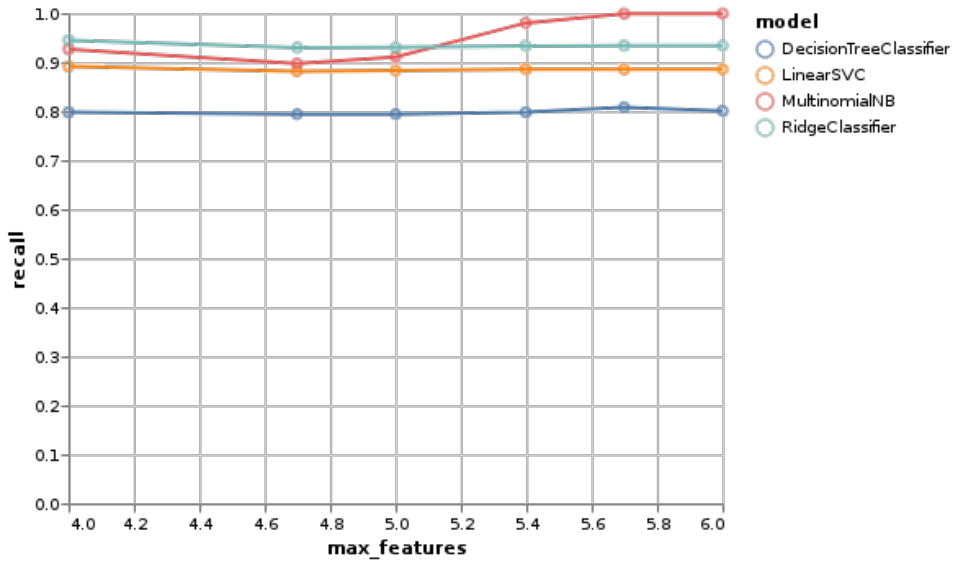

Figure 3: Model comparison with respect to the number of TF-IDF features (log scale) on reliable.

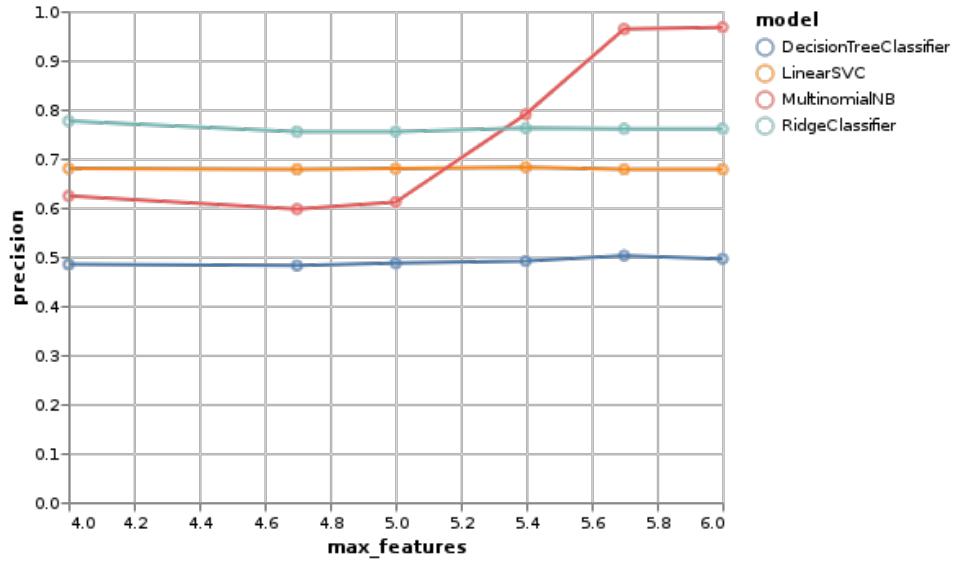

Figure 4: Model comparison with respect to the number of TF-IDF features (log scale) on fake.

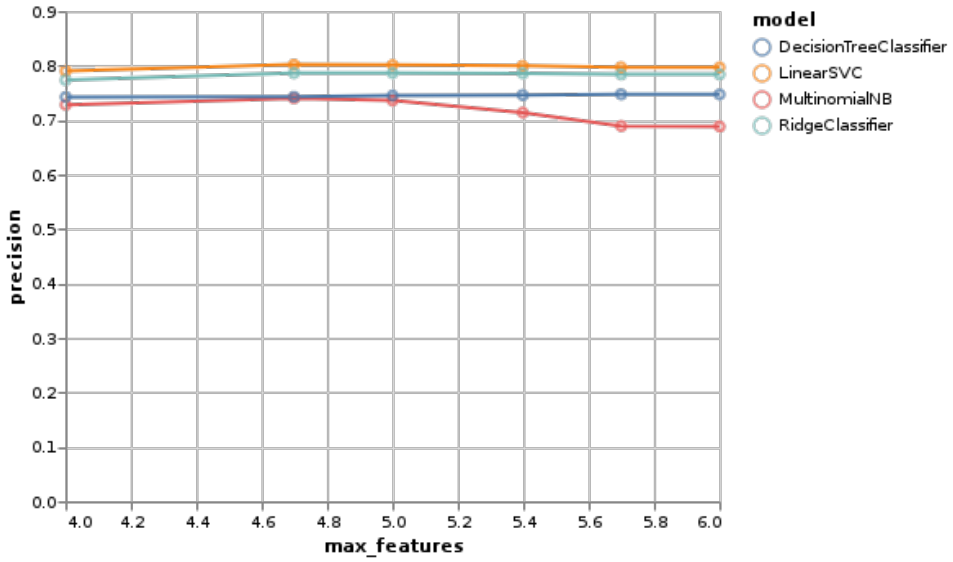

Figure 5: Model comparison with respect to the number of TF-IDF features (log scale) on reliable.

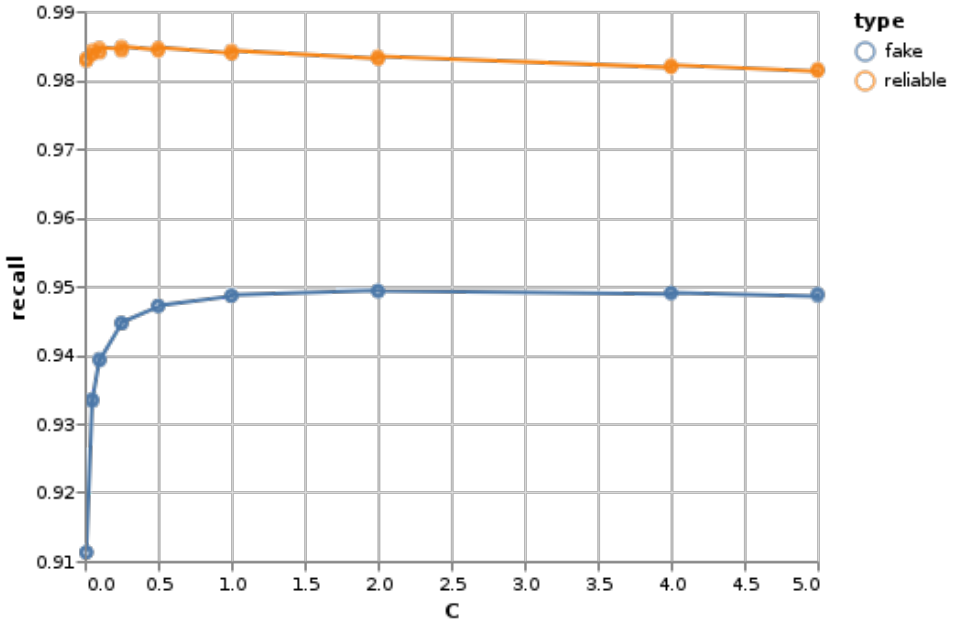

Figure 6: Recall with respect to the parameter C

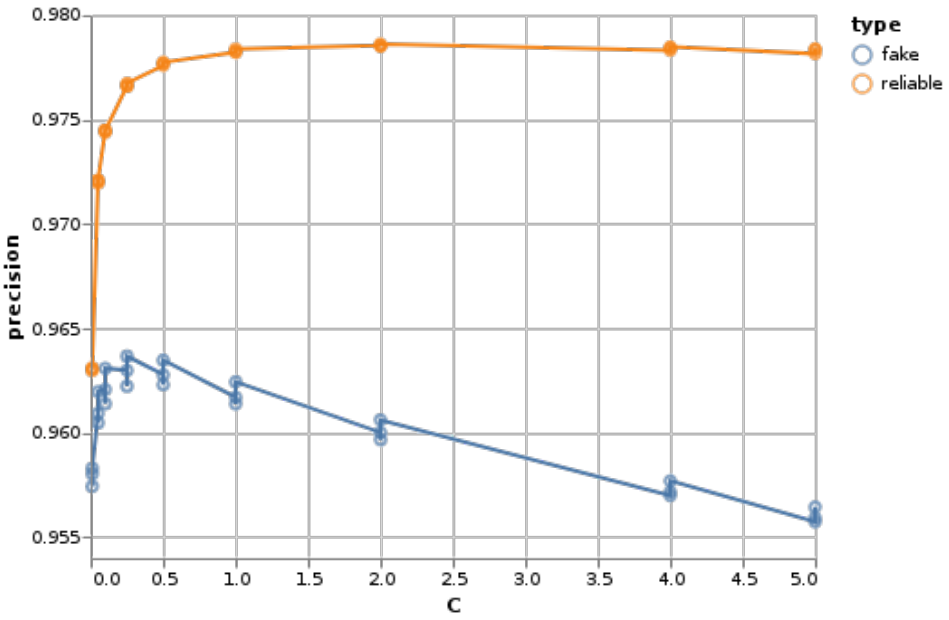

Figure 7: Precision with respect to the parameter C

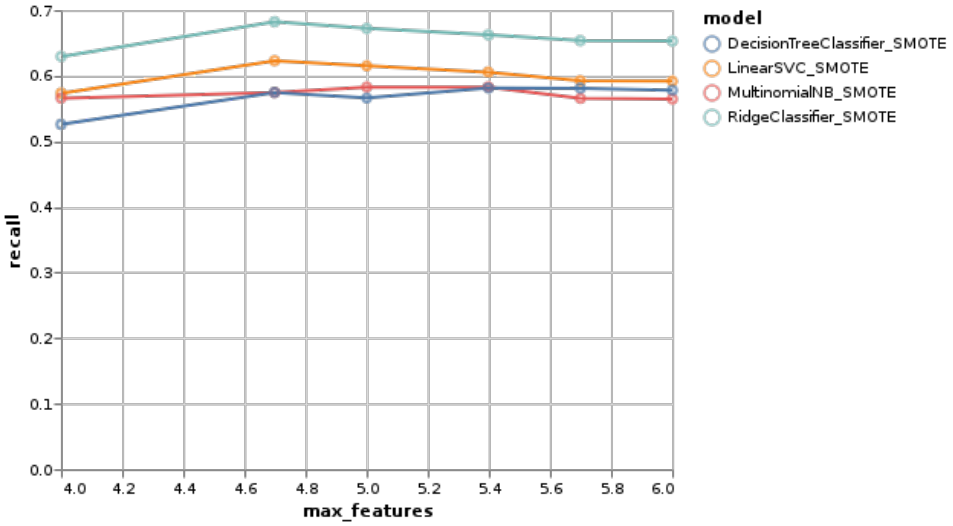

Figure 8: Model comparison with respect to the number of TF-IDF features (log scale) on fake.

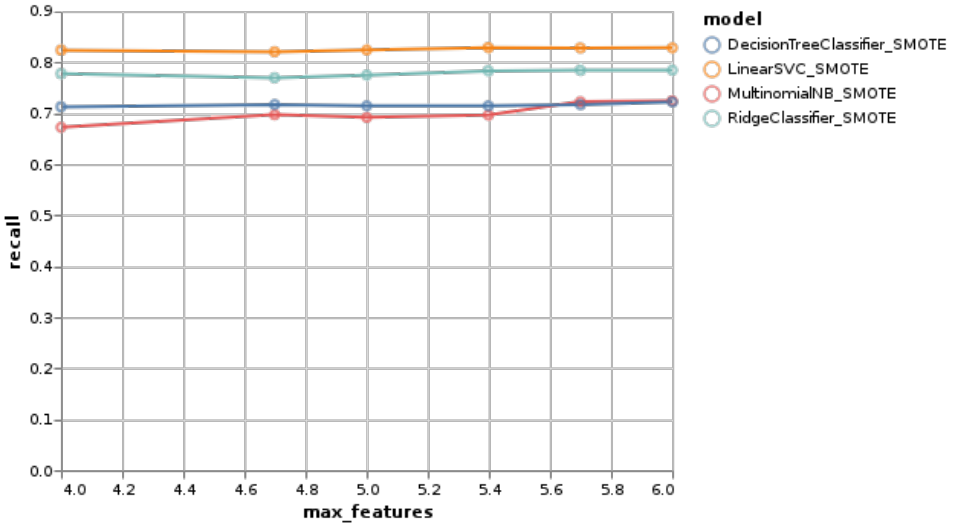

Figure 9: Model comparison with respect to the number of TF-IDF features (log scale) on reliable.

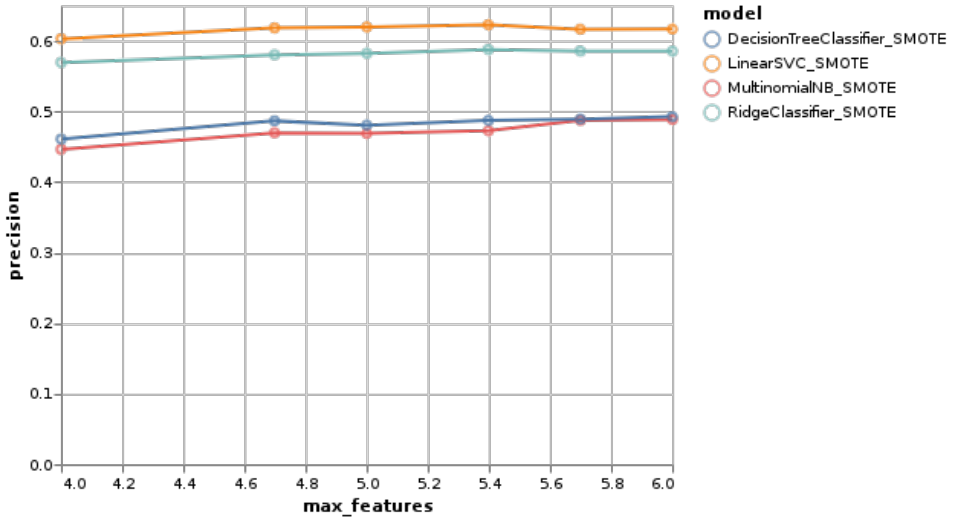

Figure 10: Model comparison with respect to the number of TF-IDF features (log scale) on fake.

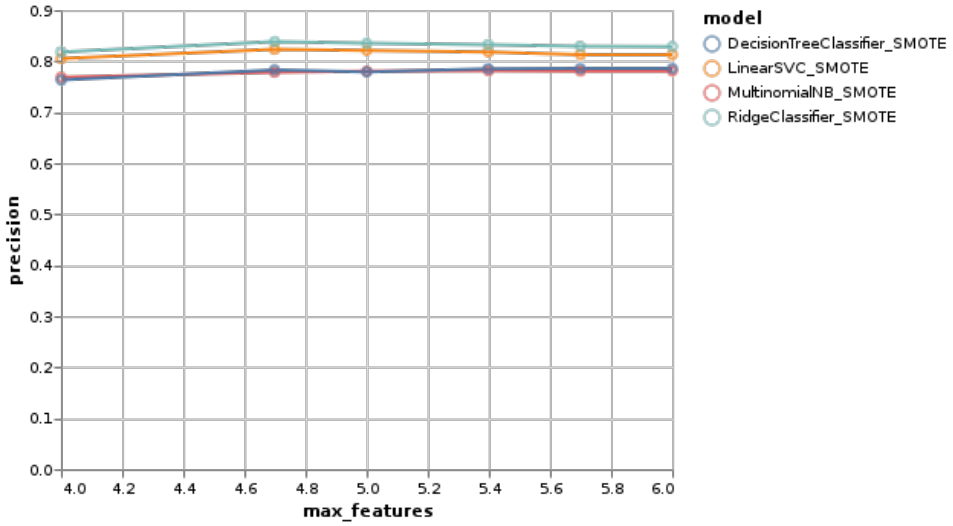

Figure 11: Model comparison with respect to the number of TF-IDF features (log scale) on reliable.

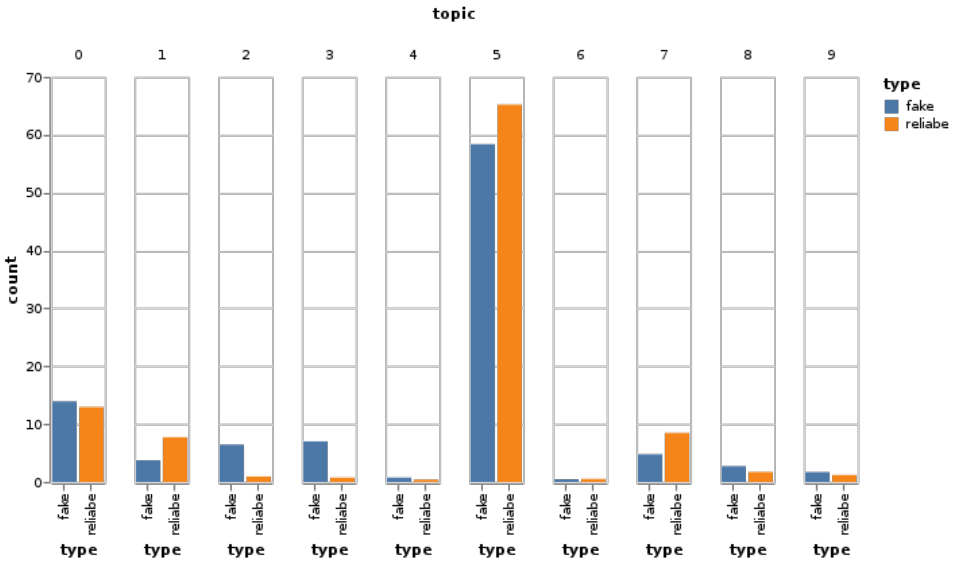

Figure 12: News count per topics for 10 topics

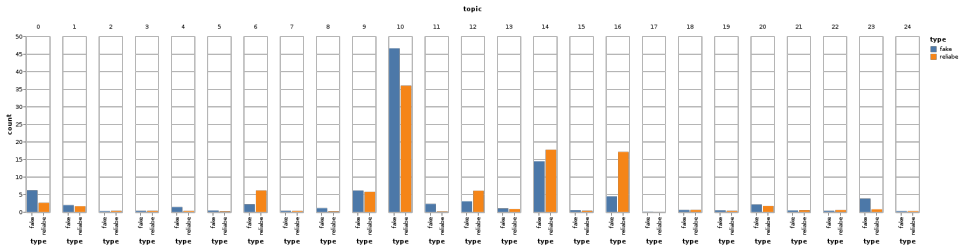

Figure 13: News count per topics for 25 topics