Master thesis : Fake news detection using machine learning

Review 1

Simon Lorent

Academic year 2018 - 2019

1 Data

Topic modeling avec LDA

+ LASSO

1.1 Sources

The data is comming from Fake News Corpus which compiles news from different sources that have been built according to OpenSources initiative. It is made of around 9.5 millions news, classified in 11 categories. A second data set is used as a test set and is comming from FakeNewsNet and is quite smaller.

1.2 Preliminary data exploration

In order to have some insight of the first data set, some statistical exploration have been made. The first thing that have been made is data cleaning, as it comes as a 30GB csv file, loading the complete file in momery is not possible. In order to overcome this difficulty, mongodb have been used. As the csv file is not entierly formated correclty, badly formated lines are thrown away. After this very early cleaning, it lefts exaclty 8.125.732 news in the database distributed as follows (see **figure 1**): Currently, only

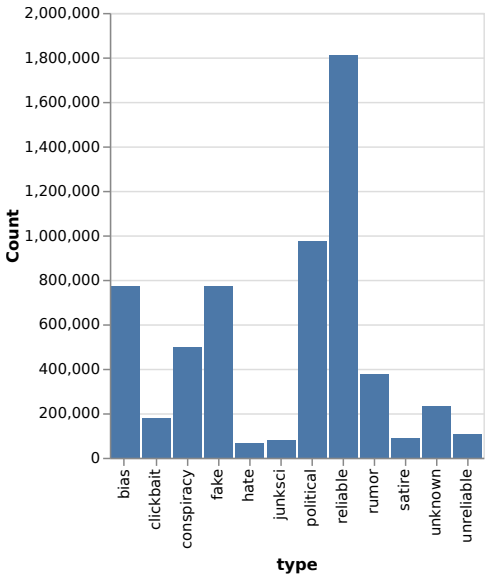

Figure 1: Distrubtion of news by type.

fake news and reliable news have been used.

Some statistics such as the number of sentences in the news and the average number of words per sentence have been computed and gave the follow result:

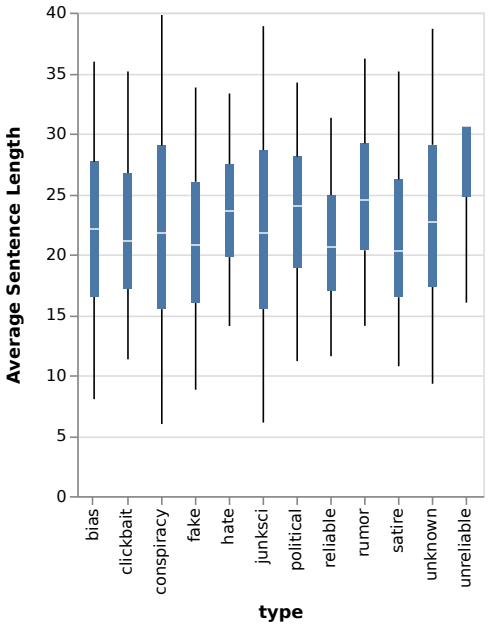

Figure 2: Average number of words per sentences.

Outliers have been removed from the boxplot because some text formating prevent conting these quantities correctly.

2 First Model : Naive-Bayes

This model uses sklearn vectorizer in order to preprocess texts. Using k-folds cross validation, the vectorizer is fitted on the train data, and test data is simply transformed by the vectorizer. The model give accuracy : 0.8921, 0.8924, 0.8922 for an average accuracy of 0.8981 on the train data and 0.8923 on test data. Training the model on the complete corpus and testing it on the second data set give the following results: Which is not bad

type	precision	recall	f1-score
fake	0.752	0.33	0.46
reliable	0.5714	0.890	0.696

smote

for a simple method.

3 Bias in the data set

After trying to explain the reason behind the results on the test set, it appears that the data set in biased. Indeed, must of the news came from the

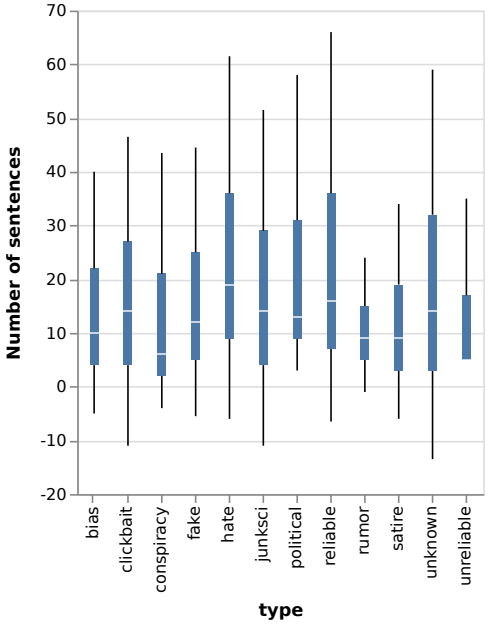

Figure 3: Number of sentences

same sources. For instance, fake news moslty comes from *beforeitsnews.com* and reliable news comes from *nytimes.com*.

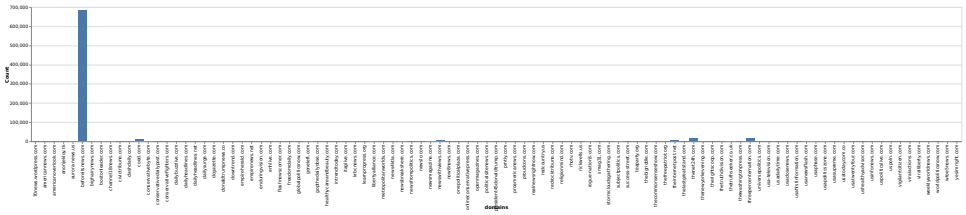

Figure 4: Fake news sources

This bias need to be overcome.

4 Next objectives

- Word embedding using word2vec Char embedding
- Making models on this embedding (Neural Networks, CNN, LSTM) + SVM
- Implementing C-LSTM [1]

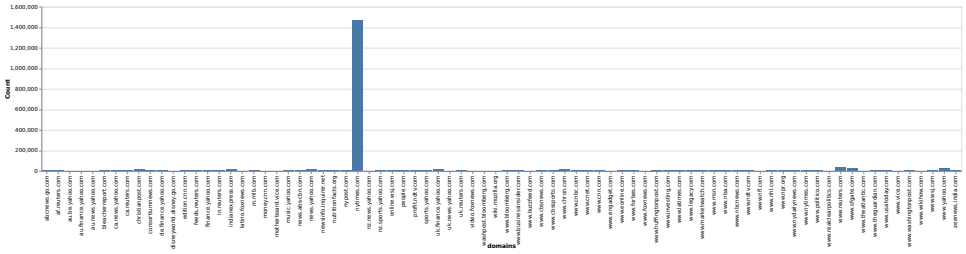

Figure 5: Reliable news sources

Bibliography

[1] Chunting Zhou, Chonglin Sun, Zhiyuan Liu, and Francis C. M. Lau. A c-lstm neural network for text classification.