Report： Text Processing

Date: 09/01/2021

Introduction

In the following report I explain the Naïve Bayes classifier experiments results and show various improvements and alternatives. My best model after processing obtained: **65.2% for 3 values** sentiment and **41.5% for 5 values** sentiment. Below are listed the processing techniques:

- **Stemming** – one of the single parameters
- **Lemmatization** – group with the stop words
- **Punctuation removal** – in vast majority improves the accuracy score
- **Stop words** – group with lemmatization, because after the testing it was figured out that they corelate well.
- **Pos tag** – it targets only the adverbs and adjectives in all regular, comparative and superlative forms as they are the most likely to be the subjective words
- **Capitalization** – one of the primary techniques, it is grouped with punctuation as both of them were seen to improve the accuracy

The primary evaluation factor is the accuracy, however in the notebook, F1 score, recall and precision were also implemented. The running time is kept around 22 seconds (best accuracy) however depending on the processing used, it can vary(depending on the number and kind of techinques). The most time-consuming stage is the training (cell 14). The running time and the size of the code was significantly reduced by applying functions and methods from the external libraries.

Accuracy (5 values)	Accuracy (3 values)	Capitalization & Punctuation removal	Stop words & lemma	Pos tag	stemming
0.412	**0.644**	0	0	0	0
0.412	**0.652**	1	0	0	0
0.39	**0.64**	0	1	0	0
0.341	**0.558**	0	0	1	0
0.415	**0.648**	0	0	0	1
0.39	**0.641**	1	1	0	0
0.354	**0.562**	1	0	1	0
0.409	**0.641**	1	0	0	1
0.318	**0.545**	0	1	1	0
0.328	**0.538**	1	1	1	0
0.387	**0.634**	1	1	0	1
0.33	**0.542**	1	1	1	1
0.407	**0.637**	0	1	0	1
0.32	**0.55**	0	1	1	1
0.34	**0.556**	0	0	1	1
0.35	**0.561**	1	0	1	1

0 – False (technique not applied) 1 – True (technique applied)

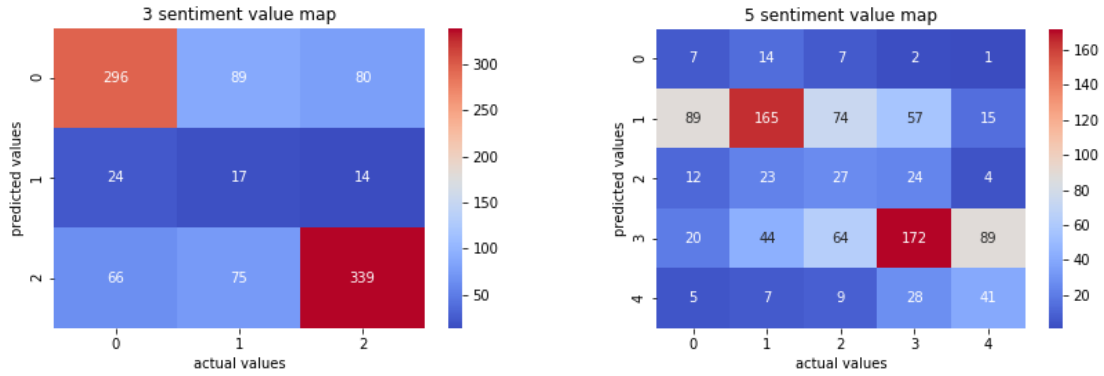
After the results analysis we can see that specific processing parameters can have a significant effect on the results. Green row presents the best results, red the worst and yellow the results that used processing techniques but obtained worse performance than the base score (no processing, only

smoothing). Stemming and capitalization & punctuation removal were the only parameters to be used and obtain the higher than base score. The application of other parameters was harmful to the performance. Stop words & Lemmatization and Pos tags combination gave the worst result. In other cases where this combination was used the results were also poor. All the results in both tables were obtained on model with smoothing, because from earlier testing it appeared that smoothing improves the performance. Looking for the ways to further improve the results, I decided to experiment with other Stemmer and the table below the show the results the obtained. Porter is the base Stemmer used. In overall performance it occurred to be the best, however for the 3-sentiment accuracy Lancaster outperformed rest.

	Porter Stemmer	Lancaster Stemmer	Snowball Stemmer	RSLPS Stemmer
5 sentiment Accuracy	**0.415**	**0.395**	**0.409**	**0.406**
3 Sentiment Accuracy	**0.648**	**0.651**	**0.647**	**0.645**

During this assignment I used several libraries:

- **Pandas** – files reading and writing.
- **Numpy** – made the array manipulation more coherent and quicker
- **Seaborn** – used only for the heatmap generating
- **NLTK** – this is crucial library in the reviews processing
- **Matplotlib** – used for labelling the heatmaps



Based on the confusion matrices above, the highest percentage of the correct predictions was obtained for the corners which represent the utmost positive or negative values (especially 3 sentiment). The highest error rate was noticed for the neutral field as it is the most difficult to certainly state and for 0 and 4 states in the 5 sentiment in favour of their less extreme version 1and3.

How to run my script

The Jupyter Notebook can be run by clicking “run all” at the top bar. There are several parameters which can be adjusted (cell 2), to do it please follow the instructions in the comments, in case of running the Notebook on other files please edit the last cell(Evaluation function) and the df path(cell 2) to train and evaluate model on the new files. Heatmap file names can be changed.

Conclusions

As it can be noticed applying the processing techniques not always improve the accuracy, that is why the extensive testing, understanding and analysis of the dataset and choosing the best performing evaluation measures is crucial and needed. Most of the processing methods decreased the accuracy measure, however the specific combination managed to outperform the base score.