Brno Institute of Techonology, FIT

NAIVE BAYES TO NEURAL TEXT CLASSIFICATION

Alexandra Fernandes, xferna17

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

GOALS

- What is text classification?
- Training and testing:
  - Naïve Bayes
  - Logistic Regression
  - Convolutional Neural Network
  - Recurrent Neural Network
- Implementation of one of these models
- Comparison and drawing conclusions

DATASETS

- **20Newsgroups** - has just one field - used on the feature weighting and top-k classifications analysis
- **HuffPost** - has several fields - used to see the impact of the number of fields in classification

THINGS TO EXPLORE

- different feature weighting/representations
- different number of fields during the classification
- getting the top k classifications, instead of just one

RESULTS

Binary weighting - if a word is present in a document, the weight is '1', otherwise the weight is '0'.
Term-frequency of words (counts) - the frequency of a word becomes its corresponding weight.
Term frequency–inverse document frequency (TF-IDF) - words that are unique to a particular document have higher weights compared to words that are less common across documents.

Legend:
NB - Naïve Bayes
LR - Logistic Regression

FEATURE WEIGHTING

NB & TF-IDF feature representation

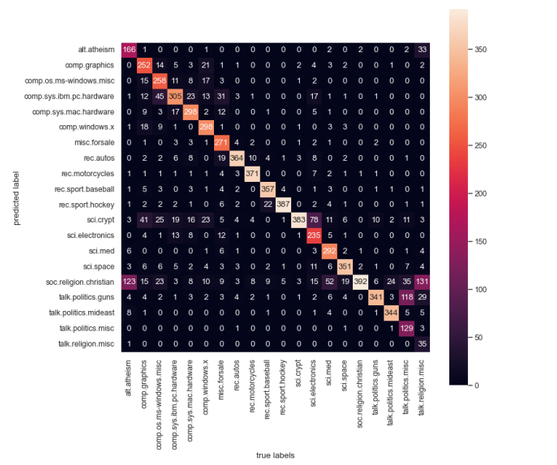

fig. 1 - confusion matrix - NB TF-IDF

LR & binary feature representation

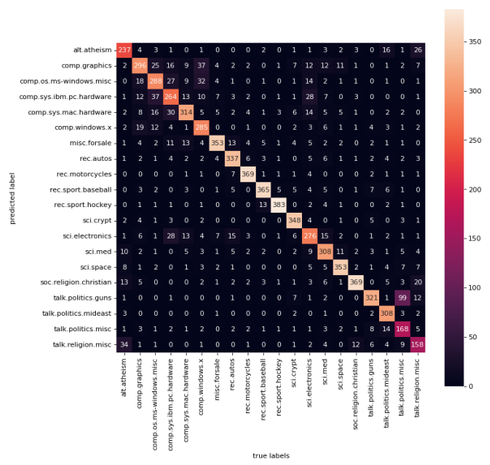

fig. 2 - confusion matrix - LR binary

LR & counts feature representation

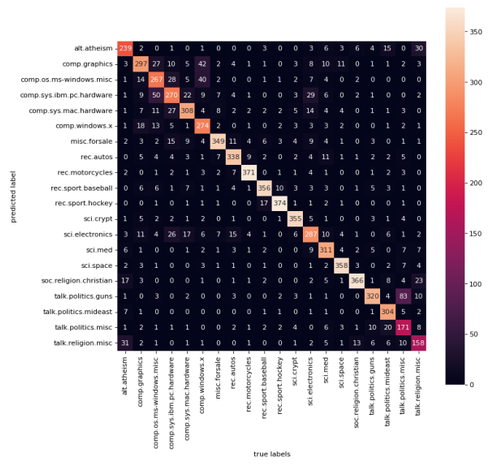

fig. 3 - confusion matrix - LR counts

LR & TF-IDF feature representation

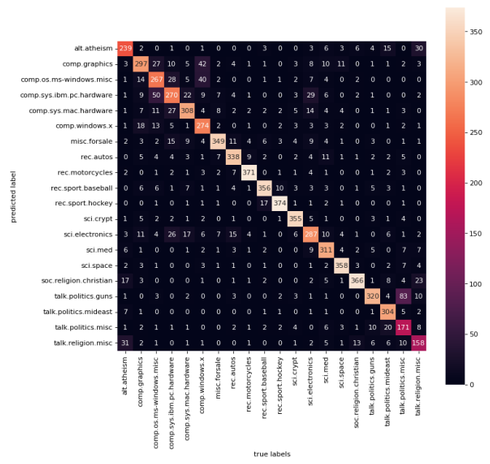

fig. 4 - confusion matrix - LR TF-IDF

	precision	recall	f1-score
accuracy			0.77
macro avg	0.83	0.76	0.76
weighted avg	0.82	0.77	0.77

fig. 5 - several metrics - NB TF-IDF

	precision	recall	f1-score
accuracy			0.81
macro avg	0.81	0.80	0.80
weighted avg	0.81	0.81	0.81

fig. 6 - several metrics - LR binary

	precision	recall	f1-score
accuracy			0.81
macro avg	0.80	0.80	0.80
weighted avg	0.81	0.81	0.81

fig. 7 - several metrics - counts

	precision	recall	f1-score
accuracy			0.85
macro avg	0.84	0.84	0.84
weighted avg	0.85	0.85	0.84

fig. 8 - several metrics - TF-IDF

These results will still be compiled into graphics and the rest of the models also evaluated .
Just by looking at the confusion matrices we cannot directly compare the different models and the impact of the feature weighting, but in all of them we get a sparse matrix and the Naïve Bayes model seems to be the one with more missclassifications (looking at the bottom right corner).

It possible to see that the best model is **LR with TF-IDF** by looking at the accuracy and the averages of precision, recall and f1-score. But since accuracy can be deceiving, it is best to look at f1, given that it balances the values of precision and recall to get a better metrics.

TOP-K CLASSIFICATIONS

	feature_representation	top_k	accuracy	mrr_at_k
0	binary	3	0.927244	0.862962
1	counts	3	0.925518	0.859997
2	tfidf	3	0.953664	0.894782

	feature_representation	top_k	accuracy	mrr_at_k
0	binary	1	0.809878	0.809878
1	counts	1	0.806293	0.806293
2	tfidf	1	0.845061	0.845061

fig. 9 - top-K classifications

Even though the accuracy is bigger using the top-3 classification, it is important to note that the top is unranked, so the first label has the same weight as the third label when computing this accuracy. But the MRR takes the rank of the first correct answer into consideration, so the higher the rank of the correctly predicted category, the higher the MRR.
Taking this into account, using the top-k classifications seems to lead to better results.

NUMBER OF FIELDS

	text_fields	feature_representation	top_k	accuracy	mrr_at_k
0	text_desc	binary	3	0.598086	0.480500
1	text_desc	counts	3	0.595590	0.478441
2	text_desc	tfidf	3	0.630696	0.510859
3	text_desc_headline	binary	3	0.794675	0.679153
4	text_desc_headline	counts	3	0.792147	0.677910
5	text_desc_headline	tfidf	3	0.835893	0.717187
6	text_desc_headline_url	binary	3	0.830101	0.715576
7	text_desc_headline_url	counts	3	0.829621	0.718099
8	text_desc_headline_url	tfidf	3	0.867256	0.751163

fig. 10 - impact of number of fiedls evaluated - on HuffPost dataset

The HuffPost dataset has several articles categorized and has several fields.
From all of the availble, text_desc (text description), headline and url are the three fields selected to use for the classification, without looking at the content of the articles themselves.

Using the top-3 classification, and trying the several combinations of feature weighting, it is possible to check that with more fields analysed the better the MRR (also the accuracy, but as explained before, it does not take into account the rank of labels).
So it seems using more fields is better, when not analysing the articles themselves.

REFERENCES

- chapters 4 and 5 of "Speech and language processing" , Dan Jurafsky and James H. Martin;
- several websites, where some source code was found and theoretical explanations

This work is not yet finished, so the conclusions are incomplete.

TO BE CONTINUED...

contact: xferna17@vutbr.cz