NLP. Determining the Sentiment of Documents

Julia Belova

Preprocessing

1. Getting rid of NaN
2. Data cleaning, lemmatization
3. Using doc2vec -> **300 features**
4. Data balancing: sampling 3000 observations for each class -> (15 000, 300)

Bayes

Bernoulli:

	precision	recall	f1-score	support
-2	0.13	0.48	0.20	473
-1	0.46	0.27	0.34	2746
0	0.69	0.42	0.52	4180
1	0.15	0.40	0.22	523
2	0.07	0.45	0.12	105
accuracy			0.37	8027
macro avg	0.30	0.41	0.28	8027
weighted avg	0.53	0.37	0.42	8027

KNN

Params:

```
[588]  1 knn = KNeighborsClassifier(n_neighbors=1)
```

✓ 1 сек.

Result:

	precision	recall	f1-score	support
-2	0.46	0.63	0.53	473
-1	0.65	0.41	0.50	2746
0	0.75	0.47	0.57	4180
1	0.14	0.78	0.24	523
2	0.50	0.69	0.58	105
accuracy			0.48	8027
macro avg	0.50	0.59	0.48	8027
weighted avg	0.65	0.48	0.52	8027

Catboost

Params:

```
model = CatBoostClassifier(iterations=2000,
                           learning_rate=0.05,
                           depth=7,
                           loss_function='MultiClass',
                          #  verbose=True,
                           task_type="GPU",
                           eval_metric='TotalF1',
                           grow_policy='Depthwise')
```

Result:

	precision	recall	f1-score	support
-2	0.49	0.62	0.55	473
-1	0.57	0.65	0.61	2746
0	0.76	0.62	0.68	4180
1	0.46	0.65	0.54	523
2	0.53	0.67	0.59	105
accuracy			0.63	8027
macro avg	0.56	0.64	0.59	8027
weighted avg	0.66	0.63	0.64	8027

SVM

Params:

```
1 svc = svm.SVC(gamma="scale", kernel='poly', C=10, degree=5)
```

Result:

	precision	recall	f1-score	support
-2	0.50	0.66	0.57	473
-1	0.63	0.52	0.57	2746
0	0.71	0.73	0.72	4180
1	0.52	0.67	0.58	523
2	0.47	0.76	0.58	105
accuracy			0.65	8027
macro avg	0.56	0.67	0.60	8027
weighted avg	0.65	0.65	0.65	8027

Random Forest

Params:

```
model = RandomForestClassifier(n_estimators=1500,criterion='gini', max_depth=15)
```

Result:

	precision	recall	f1-score	support
-2	0.55	0.62	0.58	473
-1	0.59	0.61	0.60	2746
0	0.74	0.69	0.71	4180
1	0.55	0.63	0.59	523
2	0.51	0.70	0.59	105
accuracy			0.66	8027
macro avg	0.59	0.65	0.62	8027
weighted avg	0.66	0.66	0.66	8027

RNN

Params:

Layer (type)	Output Shape	Param #
dense_38 (Dense)	(None, 1024)	308224
reshape_12 (Reshape)	(None, 1, 1024)	0
simple_rnn_6 (SimpleRNN)	(None, 512)	786944
dropout_9 (Dropout)	(None, 512)	0
dense_39 (Dense)	(None, 128)	65664
dense_40 (Dense)	(None, 32)	4128
dense_41 (Dense)	(None, 5)	165

Result:

	precision	recall	f1-score	support
-2	0.70	0.33	0.45	1020
-1	0.50	0.57	0.53	2412
0	0.59	0.73	0.65	3344
1	0.68	0.36	0.47	975
2	0.80	0.30	0.44	276
accuracy			0.57	8027
macro avg	0.65	0.46	0.51	8027
weighted avg	0.59	0.57	0.56	8027

Best model: Random Forest

Params:

```
model = RandomForestClassifier(n_estimators=1500,criterion='gini', max_depth=15)
```

Result:

	precision	recall	f1-score	support
-2	0.55	0.62	0.58	473
-1	0.59	0.61	0.60	2746
0	0.74	0.69	0.71	4180
1	0.55	0.63	0.59	523
2	0.51	0.70	0.59	105
accuracy			0.66	8027
macro avg	0.59	0.65	0.62	8027
weighted avg	0.66	0.66	0.66	8027

Code

https://colab.research.google.com/drive/13cSAtYs4_MAejOoZ1pgxd3uqGa8zYm21?usp=sharing