MUNI
FI

Automatic identification of speakers and parties in steno protocols of the Czech Parliament

Similarity of sentences with sentences of members of Czech parlament.

Dataset - ParlaMint 2.1

- Multilingual set of 17 corpora containing parliamentary debates mostly. (Including Czech parlament.)
- https://www.clarin.si/repository/xmlui/handle/11356/1432

- Czech corpus:
- From 2015 to mid-2020.
- 22,104,199 words (26,695,173 tokens)
- 1,479,990 sentences
- 60 people with 100,000+ tokens
- 113 people with 50,000+ tokens
- 275 people with 10,000+ tokens

MUNI FI

Data analysis

- Vertical format.
- Data only from Czech parlament.
- Every speach marked with party name and unique speaker name.
- Sometimes speaker is visitor of parlament with party "Nezařaz". 18064 sentences.
- Sometimes speaker is visitor of parlament with party unset. 87039 sentences.

Preprocesing

- Extracted sentences from dataset.
- Lematization and tagging.
- Syntax analysis
- Czech stopword list.

- Futures:
- Sentences
- Sentences length, rare characters count
- N-grams
- Word embedings (fasttext)

Learning methods

- Decision tree classifier
- Naive Bayes
- SVM
- Random forests
- embedings with KNN, LSVM, SVM, Random forest, Naive Bayes

Result classification

- 80:20
- Crossvalidation (When learning is not longer than day.)
- F1 score
- Data, validation, testing (Enables model parameters tuning.)

Baseline

```
Classification report for classifier DecisionTreeClassifier(class_weight=None, criter
ion='gini', max_depth=50,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=42,
            splitter='best'):
             precision    recall  f1-score   support

          0       0.24      0.17      0.20     34047
          1       0.28      0.15      0.20     35463
          2       0.22      0.73      0.33     49651
          3       0.41      0.32      0.35     37418
          4       0.36      0.03      0.06      5695
          5       0.35      0.21      0.26     25682
          6       0.30      0.06      0.10     17245
          7       0.47      0.12      0.19     22492
          8       0.37      0.02      0.04      3539
          9       0.00      0.00      0.00        54
         10       0.18      0.01      0.02      3689
         11       0.29      0.02      0.04      6994
         12       0.44      0.08      0.14     17882
         13       0.53      0.13      0.21     13358

avg / total       0.33      0.26      0.23    273209
```


MUNI FI

Resutls of party identification

embeddings_lsvc	24	25	5
embeddings_naive_bayes	16	21	8
embeddings_random_forest	37	30	1
embeddings_svc	DNF	DNF	DNF
simple_bag_of_words	29	9	3
lemma_bag_of_words	27	28	1
lemma_bag_of_words_limited250_balanced_ngrams1-3	29	24	0
lemma_bag_of_words_limited50	33	27	1
lemma_bag_of_words_limited500	27	28	1
lemma_bag_of_words_limited500_balanced_ngrams1-3	27	25	0
lemma_bag_of_words_limited50_balanced	35	14	0
lemma_bag_of_words_limited50_balanced_ngrams1-3	38	14	0
lsvc	35	35	1
lsvc_all_params	38	38	1
lsvc_all_params_balanced	38	38	1
lsvc_balanced	35	34	0
naive_bayes	49	32	1
naive_bayes_all_params	53	30	5
random_forest	53	26	1
random_forest_all_params	52	28	3
random_forest_all_params_balanced	35	20	1
random_forest_all_params_balanced_limited250	30	18	0
random_forest_all_params_balanced_limited50	30	18	0
random_forest_balanced	31	19	1
simple_bag_of_words	29	9	4
svc	19	18	11
svc_all_params	DNF	DNF	DNF
svc_all_params_balanced	DNF	DNF	DNF
svc_bag	43	29	2
svc_bag_all_params	46	26	5
svc_bag_all_params_balanced	46	26	5
svc_bag_balanced	43	29	5
svc_balanced	19	18	11

MUNI FI

Tuning for party identification

Model folder name	precision (%)	Recall (%)	Classes with precision and recall 0 %
random_forest	53	26	1
random_forest_tuning_1	53	26	1
random_forest_tuning_2	52	27	1
random_forest_tuning_3	53	25	3
random_forest_tuning_4	53	26	1
random_forest_tuning_5	54	26	1
random_forest_tuning_6	53	26	1
random_forest_tuning_7	51	27	1
random_forest_tuning_8	51	27	1
random_forest_tuning_9	54	26	1
random_forest_tuning_10	51	27	1

MUNI FI

Results of member identification

Model folder name	precision (%)	Recall (%)	Classes with precision and recall 0 %
tokens_100000_embeddings_knn	31	29	0
tokens_100000_embeddings_lSVM	23	16	0
tokens_100000_embeddings_naive_bayes	12	15	38
tokens_100000_embeddings_random_forest	30	26	0
tokens_100000_embeddings_SVM	33	26	0
tokens_100000_naive_bayes_all_params	44	22	38
tokens_100000_random_forest	43	20	28
tokens_100000_random_forest_all_params	43	21	30
tokens_100000_random_forest_all_params_balanced	35	23	0
tokens_100000_random_forest_balanced	32	20	0
tokens_100000_simple_bag_of_words	26	27	0
tokens_100000_svc	42	33	0
tokens_100000_svc_all_params	DNF	DNF	DNF
tokens_100000_svc_all_params_balanced	DNF	DNF	DNF
tokens_100000_svc_balanced	36	30	0

MUNI FI

Tuning for member identification

Model folder name	precision (%)	Recall (%)	Classes with precision and recall 0 %
tokens_100000_random_forest	43	20	28
tokens_100000_random_forest_tuning_1	45	20	28
tokens_100000_random_forest_tuning_2	47	21	20
tokens_100000_random_forest_tuning_3	43	18	34
tokens_100000_random_forest_tuning_4	45	20	27
tokens_100000_random_forest_tuning_5	42	19	30
tokens_100000_random_forest_tuning_6	47	22	17
tokens_100000_random_forest_tuning_7	46	21	23
tokens_100000_random_forest_tuning_8	47	21	20
tokens_100000_random_forest_tuning_9	45	21	25
tokens_100000_random_forest_tuning_10	45	21	24

Best result for party identification

Classification report for classifier RandomForestClassifier(criterion='entropy', min_samples_leaf=20, n_estimators=150, n_jobs=4, random_state=42):

	precision	recall	f1-score	support
0	0.22	0.81	0.35	40595
1	0.65	0.02	0.05	28372
2	0.26	0.39	0.31	39722
3	0.43	0.26	0.33	29936
4	0.80	0.00	0.00	4553
5	0.71	0.06	0.11	20544
6	0.57	0.00	0.01	13796
7	0.83	0.05	0.10	17994
8	0.69	0.01	0.02	2834
9	0.00	0.00	0.00	43
10	1.00	0.00	0.01	2952
11	0.96	0.00	0.01	5596
12	0.94	0.02	0.04	14305
13	0.92	0.02	0.05	10687
accuracy			0.26	231929
macro avg	0.64	0.12	0.10	231929
weighted avg	0.54	0.26	0.18	231929

MUNI FI

Best result for member classification

Classification report for classifier RandomForestClassifier(min_samples_leaf=13, min_samples_split=3, n_estimators=120, n_jobs=4, random_state=42):

	precision	recall	f1-score	support
0	0.15	0.79	0.25	13296
1	0.20	0.90	0.33	17036
3	0.43	0.17	0.25	8241
4	0.54	0.04	0.08	7271
5	0.43	0.03	0.06	5630
6	0.71	0.16	0.26	7373
7	0.73	0.10	0.17	6776
8	0.64	0.20	0.31	7535
9	0.53	0.07	0.13	3151
11	0.76	0.09	0.16	5588
12	0.53	0.01	0.03	2937
13	0.47	0.00	0.01	2241
14	0.65	0.02	0.03	2101
15	0.39	0.01	0.02	1980
16	0.00	0.00	0.00	2392
17	1.00	0.00	0.00	2143
18	0.44	0.03	0.05	1846
19	0.81	0.01	0.02	2045
20	0.53	0.06	0.11	1844
21	0.88	0.17	0.28	4033
22	0.73	0.31	0.43	2201
23	0.41	0.04	0.07	1599
24	0.37	0.06	0.10	6748
25	0.00	0.00	0.00	1319

Best result for member classification

55	0.00	0.00	0.00	847
56	0.00	0.00	0.00	879
57	0.20	0.00	0.00	851
58	0.63	0.11	0.19	8967
59	0.00	0.00	0.00	778
accuracy			0.22	163983
macro avg	0.44	0.06	0.07	163983
weighted avg	0.47	0.22	0.15	163983

MUNI FI

Future work

MUNI
FI

Thank you for your attention

MUNI
FI