Final Project for Natural Language Processing Lab - Report

Using word-embeddings as features to determine content writer native

Matan Kolath, Merav Mazouz

Under the supervision of
Prof. Shuly Wintner

Computer Science Department
University of Haifa
Israel
August 2019

1 Problem Statement

The purpose of our project is to examine how useful word vectors are for the task of native language recognition. For base results we used the same features described in the article. In our attemp to improve them we used pretrained word embeddings vectors as features, or sentence embedding vectors. In our project, We used two datasets "europe" or **in-domain**- texts which we use for training our classifiers, and "non europe" or **out of domain**- texts which we use to test the quality of our classifier. Our task uses labeled data for which we know the author's native language. Our in-domain criteria for evaluating the quality of our classifier is 10-fold scores, and for out of domain accuarcy.

In our project we used different models of word embeddings:

- fastText
- GloVe
- Word2vec
- Doc2vec

Countries used:

- Australia
- Austria
- Bosnia
- Bulgaria
- Croatia
- Czech Republic
- Denmark
- Estonia
- Finland
- France
- Germany
- Greece
- Hungary
- Ireland
- Italy
- Latvia
- Lithuania
- Netherlands
- Norway
- Poland
- Portugal
- Romania
- Serbia
- Russia
- Slovakia
- Slovenia
- Spain
- Sweden
- Turkey
- UK
- US
- Ukraine

2 Method

In our project we used the in-domain dataset (r/europe) for training our classifier. our classfier is a Logistic Regression classfier.
The labels of the text are known to us by looking at the flair of the author, the flair contains the assumed nationality of the user.
The text is broken into 10 consecutive sentences which form our training unit called chunk.
The chunks are analyzed for the features we are using; either word embeddings, sentence embeddings or for the baseline bag of words, char trigrams, part of speech trigrams and function words frequencies.
After the analysis we randomly select an equal number of chunks for each country, this will cause our training corpus to be of even values for each country.
The classifier performance was evaluated using 10-fold cross validation, where each fold has an equal number of chunks from each country.
After the training has been completed we start to produce results for the out of domain chunks, we go over all chunks of a certain country, extract the relavant features and then test the accuracy using our the classifier we trained on the in-domain chunks.
This step is performed sequentially, country after country, the reason for this is that the number of chunks is very large and performing this step sequentially reduces memory usage.

3 Models Comparison

The table below shows different parameters of each model we used in our project. noteworthy is the large different between the number of words used to train the word-embedding models compared to the Doc2Vec model, as well as the Doc2Vec model much larger size.

In-domain	Word2Vec	fastText	GloVe	Doc2Vec
Word count	100B	600B	840B	139M
Unique Words	3M	2M	2.2M	0.9M
Vector length	300	300	300	500
Dataset	Google News	Common Crawl	Common Crawl	In-domain chunks
Model size	3.5GB	4.4GB	5.5GB	17.0GB

4 Models Results

The trivial baseline for the binary classification task is 50% , for language family classification 20%, and for the language identification task 3.2%.

In-domain	Is Native	Language Family	Native Language
Baseline	84.18%	61.80%	47.47%
word2vec	77.60%	48.61%	29.14%
fastText	81.93%	57.26%	40.08%
GloVe	79.21%	52.84%	35.12%
Doc2Vec	83.85%	60.98%	46.07%

Out-of-domain	Is Native	Language Family	Native Language
Baseline	60.94%	39.40%	31.66%
word2vec	62.82%	40.51%	32.38%
fastText	66.34%	45.31%	36.47%
GloVe	64.13%	43.74%	36.09%
Doc2Vec	66.52%	45.97%	38.52%

Baseline

Results for all countries are ten-fold cross validation:

in-domain:

Country	Is Native	Language Family	Native Language
All	84.18%	61.80%	47.47%
Australia	69.25%	69.25%	69.25%
Austria	83.54%	53.66%	48.49%
Bosnia	95.65%	87.05%	65.27%
Bulgaria	85.86%	58.09%	41.35%
Croatia	85.07%	59.14%	30.50%
Czech Republic	87.78%	59.41%	38.01%
Denmark	89.95%	63.50%	56.22%
Estonia	88.37%	66.83%	44.82%
Finland	82.59%	42.61%	65.27%
France	77.30%	47.64%	35.37%
Germany	83.54%	53.66%	48.49%
Greece	85.05%	49.95%	41.54%
Hungary	90.28%	50.10%	45.86%
Ireland	69.25%	69.25%	69.25%
Italy	84.03%	54.03%	43.63%
Latvia	94.80%	83.20%	57.68%
Lithuania	91.80%	72.55%	46.30%
Netherlands	72.94%	42.04%	26.22%
Norway	77.24%	53.04%	37.60%
Poland	89.33%	66.30%	41.21%
Portugal	80.52%	50.75%	39.10%
Romania	80.54%	41.58%	32.19%
Russia	89.95%	73.81%	49.35%
Serbia	87.80%	63.89%	38.64%
Slovakia	95.27%	80.28%	55.78%
Slovenia	84.74%	55.78%	30.91%
Spain	85.74%	63.89%	53.34%
Sweden	77.26%	48.19%	32.04%
Turkey	92.61%	70.34%	63.56%
UK	69.25%	69.25%	69.25%
US	69.25%	69.25%	69.25%
Ukraine	97.11%	89.31%	60.91%

out-of-domain:

Country	Is Native	Language Family	Native Language	Number of Chunks
Australia	47.88%	47.88%	47.88%	421020
Austria	76.70%	33.16%	25.32%	294280
Bosnia	82.34%	54.37%	15.01%	71010
Bulgaria	73.94%	31.90%	5.75%	119990
Croatia	74.79%	38.99%	8.77%	143900
Czech Republic	75.88%	37.69%	8.01%	175320
Denmark	80.13%	33.42%	21.56%	681270
Estonia	73.47%	34.88%	7.17%	110720
Finland	73.37%	18.12%	12.14%	575030
France	73.58%	24.00%	13.13%	560100
Germany	77.44%	34.09%	26.01%	1587250
Greece	73.65%	18.10%	9.14%	205540
Hungary	76.29%	17.25%	9.86%	144520
Ireland	52.17%	52.17%	52.17%	919200
Italy	74.12%	24.15%	13.53%	246910
Latvia	86.07%	58.77%	17.19%	90890
Lithuania	76.59%	40.99%	10.33%	135670
Netherlands	71.21%	28.24%	9.81%	1246100
Norway	68.17%	29.72%	10.79%	413880
Poland	76.15%	42.31%	12.33%	437470
Portugal	75.18%	22.66%	10.45%	311950
Romania	76.39%	18.42%	6.74%	292770
Russia	74.12%	42.46%	13.34%	162160
Serbia	74.36%	38.85%	7.91%	105890
Slovakia	86.14%	59.08%	26.25%	110050
Slovenia	76.87%	36.48%	10.72%	73020
Spain	69.94%	22.91%	11.78%	330950
Sweden	70.08%	27.04%	8.23%	772480
Turkey	76.78%	28.72%	21.14%	177940
UK	52.70%	52.70%	52.70%	3418210
US	43.86%	43.86%	43.86%	5436530
Ukraine	87.49%	63.79%	22.07%	120050
Total	60.94%	39.40%	31.66%	19892070

Word2Vec

This word embedding model was pre-trained on the Google News dataset, about 100B words. The model contains 3M unique words, each with a 300 degrees vector.

in-domain:

Country	Is Native	Language Family	Native Language
All	77.60%	48.61%	29.14%
Australia	66.35%	66.35%	66.35%
Austria	75.59%	39.79%	34.36%
Bosnia	88.09%	70.95%	41.36%
Bulgaria	77.24%	39.55%	10.81%
Croatia	78.72%	44.83%	8.03%
Czech Republic	79.45%	44.71%	15.03%
Denmark	78.88%	42.19%	25.27%
Estonia	80.65%	51.44%	15.84%
Finland	74.38%	18.05%	12.66%
France	71.70%	39.07%	28.68%
Germany	75.59%	39.79%	34.36%
Greece	77.69%	38.60%	28.48%
Hungary	80.67%	23.10%	17.16%
Ireland	66.35%	66.35%	66.35%
Italy	79.35%	43.73%	33.63%
Latvia	86.86%	65.05%	22.11%
Lithuania	83.35%	57.81%	17.24%
Netherlands	66.04%	29.19%	9.17%
Norway	69.59%	35.88%	13.41%
Poland	81.14%	53.65%	22.96%
Portugal	71.68%	37.44%	20.85%
Romania	70.93%	18.44%	8.32%
Russia	86.13%	70.28%	32.56%
Serbia	83.23%	53.73%	24.52%
Slovakia	86.88%	60.79%	28.82%
Slovenia	75.86%	37.34%	5.80%
Spain	82.68%	57.12%	46.69%
Sweden	74.28%	37.67%	19.19%
Turkey	88.82%	60.49%	54.16%
UK	66.35%	66.35%	66.35%
US	66.35%	66.35%	66.35%
Ukraine	92.23%	79.57%	35.68%

out-of-domain:

Country	Is Native	Language Family	Native Language	Number of Chunks
Australia	56.18%	56.18%	56.18%	421020
Austria	74.91%	28.94%	20.61%	294280
Bosnia	79.34%	46.37%	11.04%	71010
Bulgaria	72.28%	30.96%	1.30%	119990
Croatia	73.53%	36.26%	5.20%	143900
Czech Republic	73.79%	34.28%	1.53%	175320
Denmark	73.38%	26.73%	10.21%	681270
Estonia	69.33%	32.72%	1.66%	110720
Finland	71.97%	12.28%	6.92%	575030
France	74.59%	25.57%	15.10%	560100
Germany	74.40%	30.10%	22.40%	1587250
Greece	73.37%	14.82%	6.32%	205540
Hungary	75.10%	11.60%	2.46%	144520
Ireland	58.25%	58.25%	58.25%	919200
Italy	75.29%	27.84%	15.16%	246910
Latvia	80.67%	45.91%	5.66%	90890
Lithuania	75.85%	37.32%	6.11%	135670
Netherlands	69.35%	23.72%	3.50%	1246100
Norway	68.37%	23.44%	6.07%	413880
Poland	75.19%	40.08%	8.49%	437470
Portugal	76.65%	26.34%	9.86%	311950
Romania	75.11%	16.43%	1.33%	292770
Russia	72.71%	43.18%	10.89%	162160
Serbia	73.98%	38.13%	6.28%	105890
Slovakia	81.66%	46.37%	12.11%	110050
Slovenia	73.96%	32.79%	6.98%	73020
Spain	70.25%	23.72%	15.16%	330950
Sweden	69.77%	24.65%	5.37%	772480
Turkey	79.56%	29.82%	25.37%	177940
UK	59.74%	59.74%	59.74%	3418210
US	47.62%	47.62%	47.62%	5436530
Ukraine	80.46%	54.64%	16.27%	120050
Total	62.82%	40.51%	32.38%	19892070

fastText

This word embedding model was pre-trained on a Common Crawl, about 600B words. The model contains 2M unique words, each with a 300 degrees vector.
in-domain:

Country	Is Native	Language Family	Native Language
All	81.93%	57.26%	40.08%
Australia	67.66%	67.66%	67.66%
Austria	80.86%	51.13%	45.59%
Bosnia	92.88%	79.88%	49.21%
Bulgaria	82.03%	50.26%	27.99%
Croatia	84.65%	57.61%	22.72%
Czech Republic	84.00%	51.16%	27.24%
Denmark	87.89%	58.99%	41.93%
Estonia	85.13%	61.89%	36.27%
Finland	78.09%	30.06%	25.11%
France	75.78%	43.41%	33.02%
Germany	80.86%	51.13%	45.59%
Greece	81.40%	42.94%	33.79%
Hungary	86.06%	37.95%	33.65%
Ireland	67.66%	67.66%	67.66%
Italy	82.17%	49.25%	40.30%
Latvia	92.13%	74.95%	42.15%
Lithuania	86.49%	66.49%	36.82%
Netherlands	72.07%	42.33%	23.06%
Norway	77.61%	52.54%	31.40%
Poland	84.54%	58.19%	28.56%
Portugal	78.60%	46.11%	32.64%
Romania	79.66%	33.71%	25.62%
Russia	86.86%	72.72%	41.52%
Serbia	88.01%	65.54%	32.94%
Slovakia	91.62%	67.02%	38.44%
Slovenia	81.05%	50.97%	23.02%
Spain	85.60%	64.06%	53.57%
Sweden	79.01%	47.34%	22.68%
Turkey	91.40%	70.39%	63.98%
UK	67.66%	67.66%	67.66%
US	67.66%	67.66%	67.66%
Ukraine	94.56%	83.83%	53.08%

out-of-domain:

Country	Is Native	Language Family	Native Language	Number of Chunks
Australia	57.09%	57.09%	57.09%	421020
Austria	79.56%	40.11%	33.24%	294280
Bosnia	85.35%	57.03%	15.78%	71010
Bulgaria	76.62%	36.13%	3.30%	119990
Croatia	77.49%	43.82%	9.70%	143900
Czech Republic	77.23%	37.50%	3.87%	175320
Denmark	81.96%	37.90%	17.00%	681270
Estonia	73.07%	37.06%	5.06%	110720
Finland	74.74%	13.36%	9.07%	575030
France	77.66%	24.66%	15.35%	560100
Germany	79.44%	41.00%	33.88%	1587250
Greece	76.77%	14.23%	8.33%	205540
Hungary	79.13%	11.91%	6.34%	144520
Ireland	60.62%	60.62%	60.62%	919200
Italy	77.20%	27.26%	17.42%	246910
Latvia	87.49%	57.40%	12.21%	90890
Lithuania	77.58%	42.11%	10.96%	135670
Netherlands	72.76%	34.23%	7.98%	1246100
Norway	72.16%	33.72%	8.93%	413880
Poland	77.99%	44.02%	7.90%	437470
Portugal	81.66%	26.96%	15.58%	311950
Romania	79.30%	18.04%	11.28%	292770
Russia	72.52%	46.48%	13.55%	162160
Serbia	79.15%	45.25%	10.39%	105890
Slovakia	86.71%	52.82%	14.63%	110050
Slovenia	78.75%	38.60%	12.00%	73020
Spain	71.95%	23.73%	15.79%	330950
Sweden	73.37%	34.97%	9.53%	772480
Turkey	81.60%	31.41%	27.00%	177940
UK	63.24%	63.24%	63.24%	3418210
US	50.59%	50.59%	50.59%	5436530
Ukraine	86.07%	63.17%	28.72%	120050
Total	66.34%	45.31%	36.47%	19892070

GloVe

This word embedding model was pre-trained on a Common Crawl, about 840B words. The model contains 2.2M unique words, each with a 300 degrees vector.
in-domain:

Country	Is Native	Language Family	Native Language
All	79.21%	52.84%	35.12%
Australia	68.92%	68.92%	68.92%
Austria	77.53%	45.78%	42.26%
Bosnia	91.99%	78.56%	47.95%
Bulgaria	81.05%	45.40%	19.55%
Croatia	81.83%	55.09%	18.76%
Czech Republic	80.93%	43.96%	16.92%
Denmark	81.05%	48.92%	36.71%
Estonia	82.76%	60.12%	30.41%
Finland	73.91%	21.38%	15.35%
France	70.89%	36.94%	28.50%
Germany	77.53%	45.78%	42.26%
Greece	81.83%	44.22%	33.65%
Hungary	83.73%	27.34%	22.76%
Ireland	68.92%	68.92%	68.92%
Italy	76.02%	42.64%	33.77%
Latvia	91.08%	75.11%	38.05%
Lithuania	84.52%	63.06%	32.68%
Netherlands	64.48%	32.90%	12.45%
Norway	69.68%	39.92%	22.05%
Poland	81.74%	53.71%	20.12%
Portugal	69.74%	34.38%	19.57%
Romania	76.25%	24.38%	16.17%
Russia	87.24%	73.18%	41.42%
Serbia	86.94%	63.41%	31.64%
Slovakia	91.16%	67.79%	38.74%
Slovenia	78.50%	45.96%	14.83%
Spain	81.08%	57.79%	50.06%
Sweden	70.63%	35.56%	9.86%
Turkey	89.63%	66.39%	60.02%
UK	68.92%	68.92%	68.92%
US	68.92%	68.92%	68.92%
Ukraine	95.35%	85.50%	51.46%

out-of-domain:

Country	Is Native	Language Family	Native Language	Number of Chunks
Australia	58.57%	58.57%	58.57%	421020
Austria	74.52%	32.71%	30.38%	294280
Bosnia	79.72%	54.08%	16.88%	71010
Bulgaria	70.38%	33.66%	1.64%	119990
Croatia	70.27%	41.93%	7.04%	143900
Czech Republic	70.73%	36.62%	3.34%	175320
Denmark	76.06%	29.03%	16.34%	681270
Estonia	68.02%	35.35%	4.16%	110720
Finland	67.88%	11.49%	5.35%	575030
France	70.03%	22.38%	15.80%	560100
Germany	75.38%	36.24%	33.98%	1587250
Greece	72.25%	15.64%	8.56%	205540
Hungary	71.89%	11.28%	3.37%	144520
Ireland	62.95%	62.95%	62.95%	919200
Italy	73.01%	25.11%	18.01%	246910
Latvia	83.17%	59.27%	12.47%	90890
Lithuania	72.66%	40.81%	6.47%	135670
Netherlands	65.25%	26.72%	4.02%	1246100
Norway	64.93%	24.78%	3.46%	413880
Poland	72.87%	42.33%	6.06%	437470
Portugal	72.74%	19.93%	6.49%	311950
Romania	74.35%	15.35%	8.10%	292770
Russia	69.34%	45.79%	14.72%	162160
Serbia	71.05%	41.99%	10.46%	105890
Slovakia	83.01%	55.18%	18.91%	110050
Slovenia	69.99%	36.18%	5.61%	73020
Spain	66.61%	19.74%	13.65%	330950
Sweden	64.69%	23.91%	2.59%	772480
Turkey	78.65%	35.23%	30.47%	177940
UK	65.14%	65.14%	65.14%	3418210
US	51.55%	51.55%	51.55%	5436530
Ukraine	85.99%	64.38%	25.78%	120050
Total	64.13%	43.74%	36.09%	19892070

Doc2Vec

The following results were acquired using gensim's doc2vec implementation The Doc2Vec model was trained by us on all the text data of the in-domain chunks Each sentence was fed into the model for training, no extra pre-processing was done The model contains a 500 degrees vector for each sentence This method obvious drawback is the model was fed sentences for training, but paragraphs of unrelated sentences for inference The total number of words that were fed to the model was 139134960 and a total of 967823 unique words.

in-domain:

Country	Is Native	Language Family	Native Language
All	83.85%	60.98%	46.07%
Australia	66.53%	66.53%	66.53%
Austria	81.85%	48.84%	42.90%
Bosnia	96.35%	87.57%	60.39%
Bulgaria	86.15%	55.05%	34.26%
Croatia	87.12%	61.50%	27.44%
Czech Republic	85.29%	54.22%	31.18%
Denmark	92.11%	62.72%	56.19%
Estonia	87.89%	66.41%	41.52%
Finland	80.37%	33.59%	29.47%
France	76.82%	45.50%	33.00%
Germany	81.85%	48.84%	42.90%
Greece	83.89%	47.26%	37.97%
Hungary	89.68%	45.40%	41.74%
Ireland	66.53%	66.53%	66.53%
Italy	83.53%	51.87%	42.15%
Latvia	97.61%	86.35%	59.17%
Lithuania	89.13%	70.47%	43.59%
Netherlands	71.99%	41.34%	26.63%
Norway	77.22%	53.16%	38.48%
Poland	87.87%	65.78%	39.57%
Portugal	80.63%	52.43%	38.64%
Romania	81.50%	39.47%	31.50%
Russia	89.72%	75.84%	49.92%
Serbia	90.87%	68.64%	40.00%
Slovakia	96.80%	83.96%	62.64%
Slovenia	83.51%	56.00%	33.31%
Spain	86.84%	69.63%	58.36%
Sweden	79.63%	48.44%	32.07%
Turkey	92.56%	73.73%	67.63%
UK	66.53%	66.53%	66.53%
US	66.53%	66.53%	66.53%
Ukraine	98.32%	91.24%	65.35%

out-of-domain:

Country	Is Native	Language Family	Native Language	Number of Chunks
Australia	54.61%	54.61%	54.61%	421020
Austria	75.86%	34.17%	26.20%	294280
Bosnia	89.66%	63.70%	17.65%	71010
Bulgaria	71.35%	30.02%	5.36%	119990
Croatia	73.91%	38.52%	9.84%	143900
Czech Republic	73.59%	35.38%	6.23%	175320
Denmark	90.13%	41.64%	34.73%	681270
Estonia	72.51%	33.77%	7.53%	110720
Finland	72.10%	17.19%	12.21%	575030
France	72.51%	25.79%	15.94%	560100
Germany	75.92%	35.44%	27.40%	1587250
Greece	72.32%	17.31%	9.51%	205540
Hungary	75.24%	16.28%	9.66%	144520
Ireland	60.02%	60.02%	60.02%	919200
Italy	73.55%	25.96%	13.84%	246910
Latvia	93.27%	68.88%	23.43%	90890
Lithuania	76.68%	39.76%	11.22%	135670
Netherlands	67.23%	29.01%	10.01%	1246100
Norway	68.87%	31.04%	12.39%	413880
Poland	75.31%	40.69%	13.76%	437470
Portugal	72.63%	25.87%	12.27%	311950
Romania	74.87%	18.72%	7.08%	292770
Russia	72.32%	42.50%	17.13%	162160
Serbia	73.95%	38.24%	7.97%	105890
Slovakia	75.92%	35.31%	12.44%	110050
Slovenia	75.92%	35.31%	12.44%	73020
Spain	68.52%	26.88%	14.17%	330950
Sweden	68.92%	29.89%	9.62%	772480
Turkey	77.86%	35.09%	28.31%	177940
UK	61.74%	61.74%	61.74%	3418210
US	57.80%	57.80%	57.80%	5436530
Ukraine	94.20%	71.54%	31.44%	120050
Total	66.52%	45.97%	38.52%	19892070

5 Results Analysis

In identifying native language English speaking countries were better identified than all other countries and Ukrainian is the most easily recognized native language.

However, English was the least identifiable language for the binary task. The English speaking countries native language average pulled the average accuracy up by almost 20% and in the binary identification task reduced the average accuracy by almost 10%.

In identifying families it was found that the Slavic family could be identified better then all other families, this corresponds to the fact that the Slavic languages is not similar to the English language. The least easily identifiable family is Latin possibly because English's root in the Latin family.

Doc2Vec produced the best results overall possibly because the training corpus of doc2vec model was the in-domain text chunks making it more relevant to our texts. The doc2vec training suffers from low memory capacity, and also suffers from a much smaller corpus size. Word embedding models produced better results then the baseline however word embeddings models dimensionallity reduction function requires improvements. The Doc2Vec and word embeddings produced better results then the baseline with much less features (300/500 vs 3500), making them better for low memory systems.

The in-domain results does not give a good indication for the out-of-domain results. The overall order of out-of-domain results is similar across models. it is possible that the different countries indicators for native language are shared across collected features.

Overall we saw improvements in using word embeddings as features.

Word embeddings method are faster and consume less memory during runtime and on disk.

Comparing to the original paper, None of the methods used come close to some other features such as social network features.

6 Improvements

Explore different technics to produce constant length feature vector from a variable length sentence-word vector, A possibility PCA variant on the word embeddings vectors.

Attempt using Neural-network based word-embedding models such as Bert and ELMo. Training Doc2Vec model with a high memory machine or on a better corpus to produce a better model.

7 Difficulties

the first difficulty we encountered was the inconsistent size of the available text for each country, to solve it we down-sampled the in-domain chunks of each country to be of a consistent size, this allowed comparable results between countries.

The second difficulty was accuracy discrepancy when using multi-model identifiers (one for each task). the "is native" and the "language family" were producing worse results for some countries compared to the "native language" model. to circumvent this issue we used only the model that identifies the native language and then extrapolated the language's family and if the speaker was a native English speaker.

Next since the in-domain results were ten-fold cross validation we shuffled the order of the chunks to ensure all "folds" have equal number of chunks from each country, Since all English speaker countries, speak English, we couldn't identify which chunks belongs to which country without significant overhead which we elected not to implement since we considered the in-domain results of secondary priority to the out-of-domain results. After solving the mentioned issues we encountered technical difficulties such as hitting the RAM limit on our machine, and the long run-time of the program.

The run-time of the program took between 16 hours for the word embedding models to 36 hours for the baseline results, and over 5 days for the Doc2Vec model. The Doc2Vec model implementation we used does not support loading the model to RAM, instead electing to use a memory mapped file, which slows our program.

To tackle the RAM issues we elected to persist the features matrices to disk, and loading them as needed. this lead us to implement a system which allows us to either calculate the features or use an existing matrix we have computed previously, this by itself was hard since different matrices might have been extracted using different vocabularies, for this reason we also saved the vocabularies, and the classifiers to be able to reproduce exactly the same results. The last issue we had was with the Doc2Vec model, we couldn't find a pre-trained Doc2Vec model online, so we resorted to training the model by ourselves. This training proved to be very memory consuming, initially we tried training the model on all of our text chunks, but a pre-train estimate has assessed that

at least 480GB of RAM where needed to train such a model. instead we only used the in-domain text chunks, which resulted in a 30GB of needed RAM. finally the Doc2Vec model was extremely slow to use, due to being used as a memory-mapped file. we tried searching for a way to load all of it into RAM but couldn't find one.

References

[1] Gili Goldin, Ella Rabinovich and Shuly Wintner. Native Language Identification with User Generated Content. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018), pages 3591-3601, Brussels, Belgium, November 2018.

[2] Quoc Le, Tomas Mikolov Distributed Representations of Sentences and Documents

Github: `https://github.com/makolath/NLP.HU`
fasttex: `https://fasttext.cc/docs/en/english-vectors.html`
GloVe: `https://nlp.stanford.edu/projects/glove/`
word2vec: `https://code.google.com/archive/p/word2vec/`
Doc2Vec implementation: `https://radimrehurek.com/gensim/models/doc2vec.html`