Computational Investigation of Linguistic Markers in Discourse of Political Adversaries via Interpretation of Recurrent Neural Network

Poster · February 2018

DOI: 10.13140/RG.2.2.12282.59845

3 authors:

Bogdan Kirillov
Skolkovo Institute of Science and Technology

Aleksandra Petrovskaya
Nencki Institute of Experimental Biology

Anastasia Sergeeva
University of Luxembourg

Computational Investigation of Linguistic Markers in Discourse of Political Adversaries via Interpretation of Recurrent Neural Network

Bogdan Kirillov Aleksandra Petrovskaya Anastasia Sergeeva

Skolkovo Institute of Science and Technology

Higher School of Economics

ITMO University

Objectives

A community discourse can be analyzed through texts written by participants of the community since it is expressed in those texts in various ways. For large communities, sheer amount of texts generated limits the ability of human researcher to comprehend unique features of a discourse. But modern Machine Learning algorithms are able to process large amount of texts thus aiding the human researcher in investigation. In this study we offer:

1. a large corpus made from three types of text: writings of Russian pro-government and opposition activists and neutral texts without political coloring;
2. a modern word-level Recurrent Neural Network-based approach for unsupervised detection of discourse-specific linguistic markers.

Results

We have gathered a corpora of large politically-colored texts and using the corpora as training set have constructed a Deep Recurrent Neural Network-based classifying model that is able to distinguish among neutral and politically-colored (both pro-establishment and opposition) sentences with accuracy around 54% (for three-class classification the performance of random choice is 33.33...%, the gain is about 21%). This provides a computational evidence in favor of hypotheses "discourses embedded in texts of political adversaries diverge" and "discourses embedded in texts of political adversaries differ from ones embedded in neutral texts". In addition, we have made a successful attempt to interpret the activations of trained model whose results (in a way) converge to results of more common methods like intent analysis. Due to instability of learning and difficulties in investigation of the trained network's activation we plan to employ more complex methods of neural network interpretation in our future research.

Corpus statistics

Figure 1:Wordcloud for opposition corpus

Figure 2:Wordcloud for pro-establishment corpus

	Number of files	Number of sentences used for training	Mean length of a sentence (words)
Pro-establishment	200	13959 (all of them)	15.8
Opposition	165	12896 (all of them)	14.9
Neutral	649	12000 (chosen at random)	20.5

Intent analysis

	Pro-Establishment	Opposition	RNN-friendly?
Local irony	+	+	+
Thematic irony	-	+	-
Lexical diversity	-	+	+
Series of synonims	+	+	+
Georgaphy-related info	+	-	+
"Corruption"	-	+	+
Speech style	Formal	Casual	+
Flesch–Kincaid	Relatively difficult	Relatively difficult	Not applicable
Fog index	Higher education (15.78)	High school (9.3)	Not applicable
Tonality	+	Neutral, positive	+

"+" here denotes presence of a category in texts that belong to a class. Column "RNN-friendly?" is attributed to category of intent analysis, not to class of texts. + here means that RNN may theoretically detect this category.

Model

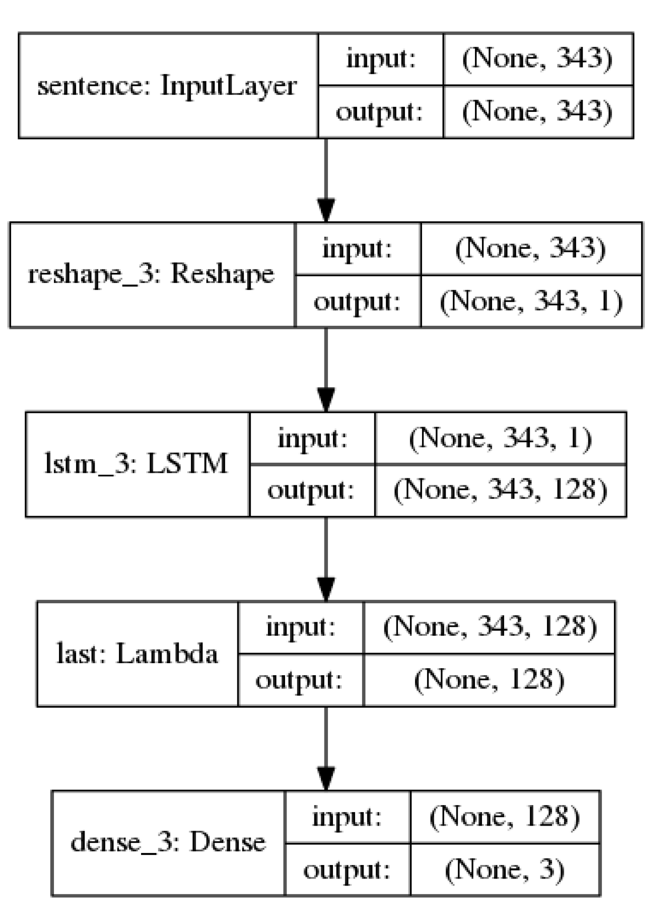

Figure 3:Chosen configuration of a classifyer

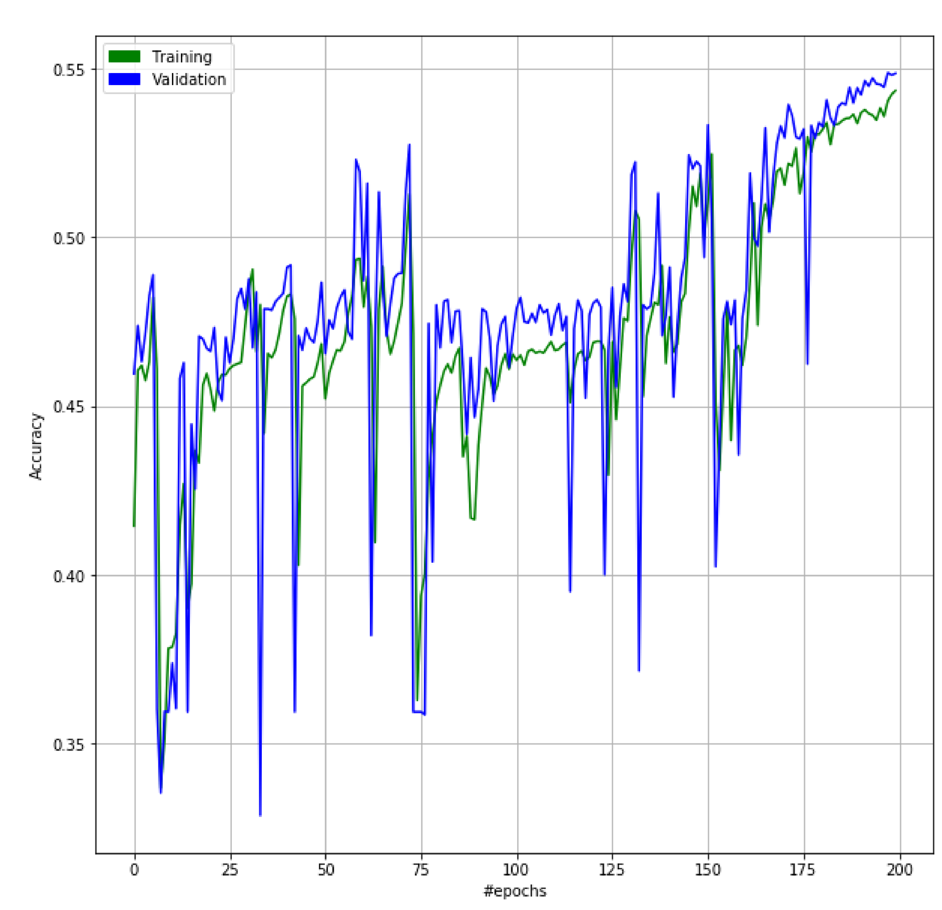

Figure 4:History of training

predicted	0	1	2
real			
0	81.215329	-9.417010	-9.948491
1	-24.864794	53.160394	-50.361113
2	-46.942803	-48.584185	63.140591

Figure 5:Quetelet indices (in %) for predicted and real labels

We investigate strength and weaknesses of the trained model by calculating Quetelet indices:

$$q(column|row) = \frac{P(column|row) - P(column)}{P(column)} \quad (1)$$

One can see that model predicts some classes better than others. For example, the real class being 0 (neutral sentence) raises probability of prediction the neutral label by 81%, but the same raise is 53% for 1 (opposition) and 63% for 2 (pro-establishment).

Workflow

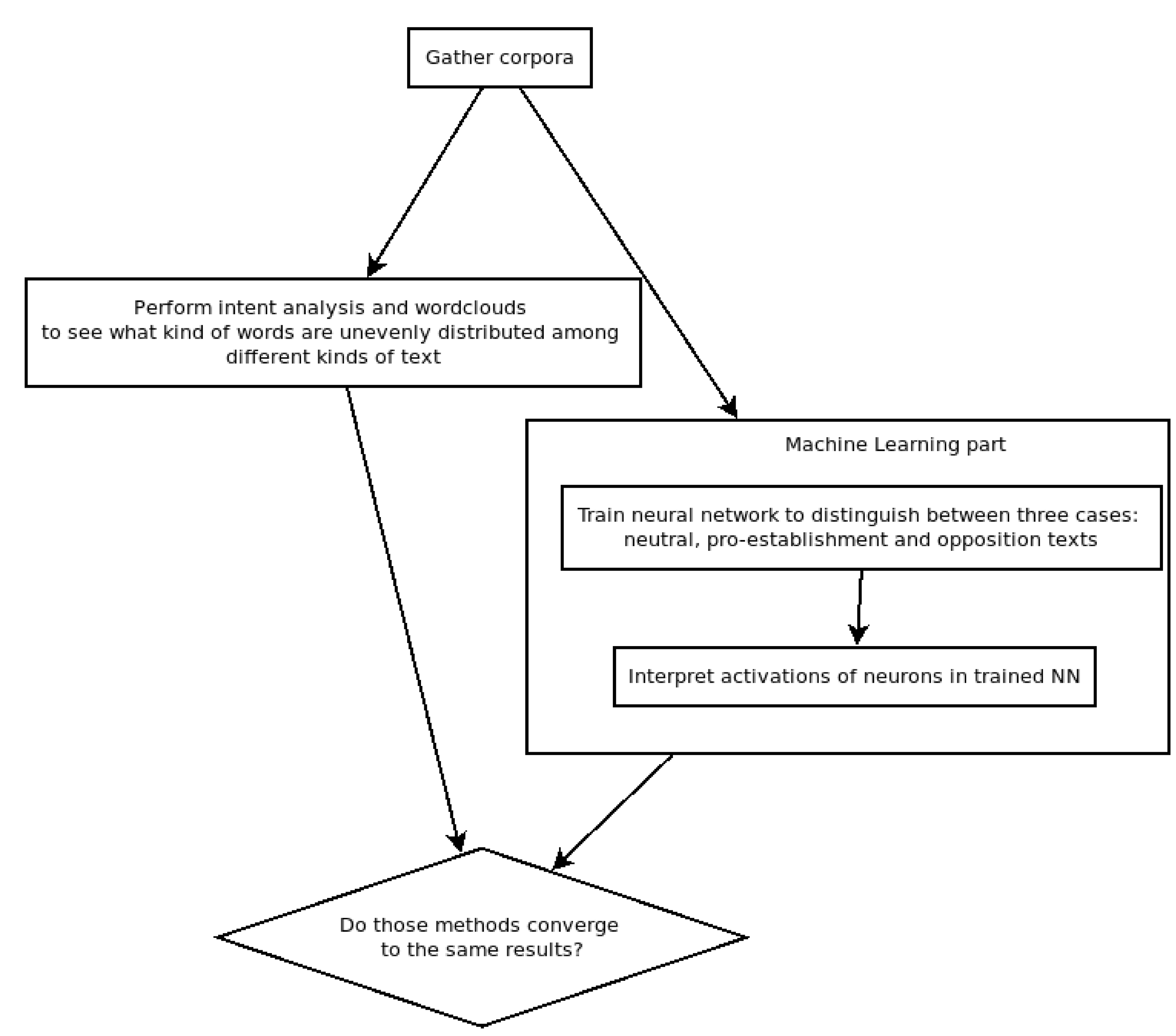

Figure 6:Workflow of the study

Interpretation

```
activations[8]#Neutral
```

литературный критик александр архангельский декабрь на встреча с президент россия владимир путин выступать с речь о давление на деятель культура и необходимость гуманизация культура и общество

```
activations[10]#Opposition
```

владимир путин поручать такой оборудый вание создавать разуметься в россия а министр никифоров взять под козырек

```
activations[3]#Pro-establishment
```

политика администрация президент рф владимир планировать проводить анализ регион и сделать вывод какой губернатор следовать увольнять а какой оставлять писать газета ру со ссылка на источник в кремль

Figure 7:Examples of sentences colored by activation

Future studies

1. Automatized search for relevant neurons - manual search is quite tedious and may not be even possible since neural network usually learn a distributed representation;
2. Attention-based model instead of plain LSTM - models with predefined notion of attention are more easy to interpret since in this case we only have to visualize the attention;
3. Evolutionary selection of hyperparameters - the parameters of our network were chosen using educated trial and error. This might be improved by algorithmic ways of hyperparameter optimization;
4. Usage of a different ML-task as basis for model interpretation - e.g. sentiment-based text segmentation.

References

1. Петровская А. В., Нецелевое расходование бюджетных средств на примере оппозиционного и проправительственного российских дискурсов // Языки знания, языки власти: вопросы исторической эволюции и региональной специфики, Бюллетень научных студенческих обществ ННГУ им. Н.И. Лобачевского. Выпуск 7 http://smu.unn.ru/files/n7/razdel4.pdf - 2016.
2. Karpathy A. The unreasonable effectiveness of recurrent neural networks //Andrej Karpathy blog. – 2015.
3. `https://github.com/amueller/word_cloud`

Contact Information

Bogdan Kirillov 8k1r1ll0v@gmail.com
Aleksandra Petrovskaya petrovskaya93@bk.ru
Anastasia Sergeeva an.se.sergeeva@gmail.com