RoST-DL: Romanian Short Text Classification

Integrating Deep Learning for NLP in Romanian Psychology

Ioan Cristian Schuszter
West University of Timisoara
chrisschuszter@gmail.com

Introduction

We propose a Deep Learning based system for classifying Romanian short sentences in the context of a psychotherapy of anxiety and depression study (**PsiTAD***).

Three datasets of answers tested:

- **emotions**
- **thoughts**
- **behaviors**

Goals and Methods Used

Goals

1. Initial analysis of the **fastText** embeddings on the Romanian datasets.
2. Implementation and comparison of 3 model architectures
3. Test the capabilities of transfer learning for this problem: embeddings trained on large corpus: small datasets(**~500** instances each)

Methods Used

- pre-trained Romanian embeddings using **fastText**
- PCA and averaging of all the sentences / class - inital exploration and data analysis
- Pandas, Keras, Tensorflow

Models:

- CNNs for sentence classification *[2]*
- GRUs
- BiLSTMs

Heavy usage of *dropout* is employed to avoid overfitting on the small datasets.

Data Analysis

Reasoning: good word embeddings should show some polarisation between the classes of the datasets.

Due to the n-grams for the word embeddings, even **OOV** words can be used with confidence (e.g. ***disconfort***, which should be ***discomfort***).

1. Top 5 most similar words for the target word (emotions)

Rank	*Input word*		
	depresie	**disconfort**	**tristete**
1	depresia	discomfort	tristetea
2	deprimare	inconfort	neliniste
3	anxietate	stres	tristetii
4	depresive	anxietate	dezamagire
5	neajutorare	iritare	deznadejde

Averaging all the sentences

2. Closest words to the averages of all sentences (emotions)

Rank	*Label*	
	1	**0**
1	nesiguranta	cred
2	neliniste	dar
3	incordare	simt
4	teama	spun
5	stresata	chiar

Ideal: show closeness to some domain-specific words in the positive examples, when computing the avg. sentence.

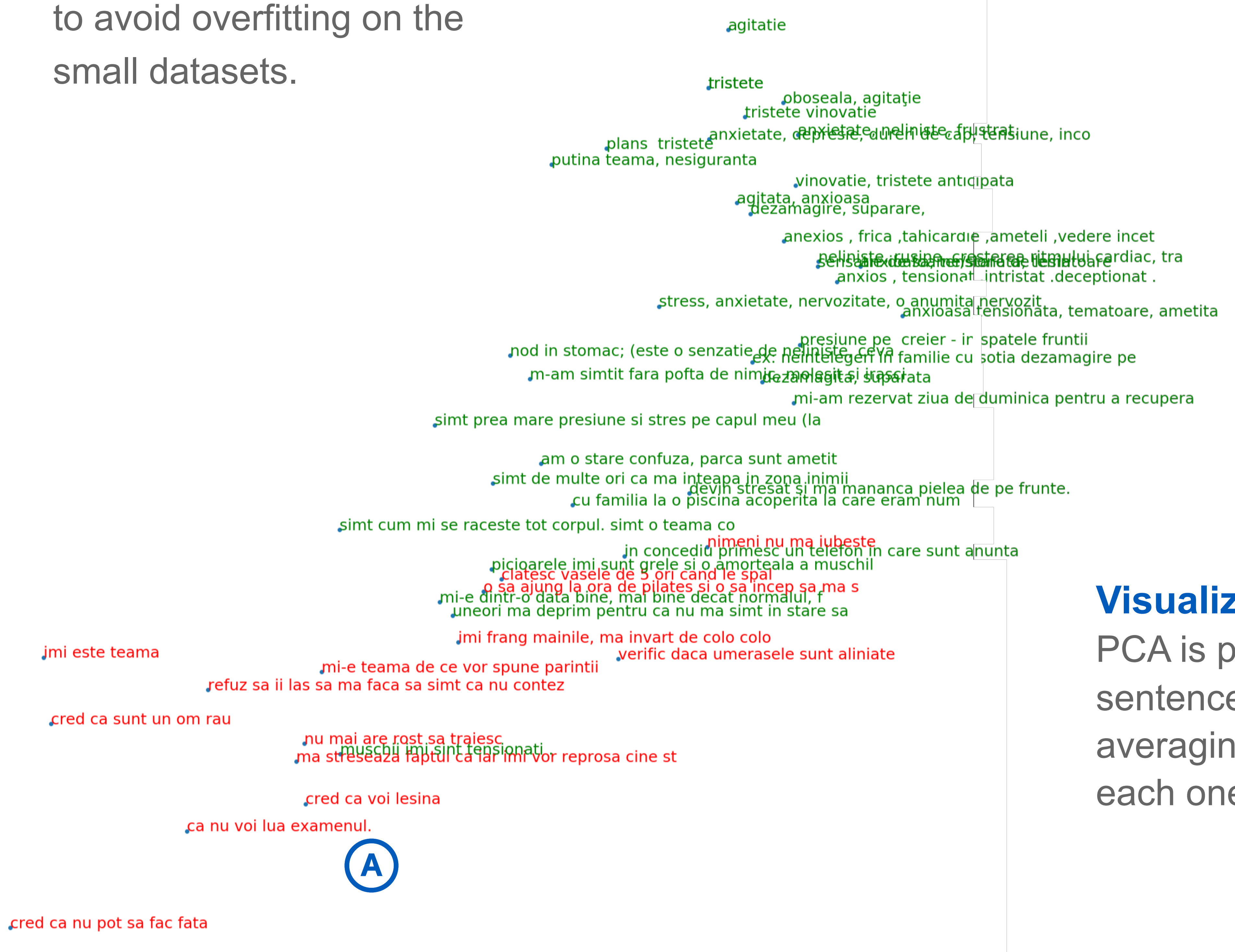
Implementation & Results

- **CNN**
  - varying filter sizes, 1D conv layers
  - dropout of 60%
  - ADAM optimizer. Params: 0.6, 0.99.
- **GRU**
  - two layers, the first creating a more abstract representation
  - 50% dropout
  - ADAM optimizer. Params: 0.9, 0.99.
- **BiLSTM**
  - Bidirectonal initial layer, 2nd layer uses the features produced by the 1st
  - 50% dropout and ADAM as above.

[Table 3] presents the classifier performance on all 3 datasets. Both **LSTMs** and **GRUs** have similar results, with **LSTMs slightly better** in all cases except "thoughts". **CNNs** provide more consistent results across runs, when compared to recurrent models.

Conclusions

- Some insights into the quality of **pre-trained word embeddings** for the Romanian language
- Thorough comparison between 3 real datasets from the Faculty of Psychology, using 3 different architectures
- Future roll-out as an automatic labelling system, easing the job of psychologists.



3. Classifier performances (30 runs each)

Dataset	*Metric*	*CNN*	*GRU*	*BiLSTM*
Emotions	**Std. Dev. Acc**	**1.97**	3.31	3.90
	Max. Acc.	88.09	92.85	**95.23**
	Mean. Acc.	83.9	86.66	**89.88**
Behavors	**Std. Dev. Acc**	**3.88**	5.24	6.42
	Max. Acc.	90.47	88.09	**92.85**
	Mean. Acc.	**86.19**	82.26	**81.71**
Thoughts	**Std. Dev. Acc**	**1.30**	2.05	2.52
	Max. Acc.	83.33	83.33	83.33
	Mean. Acc.	**80.95**	79.66	79.64

Visualizing

PCA is performed on the sentences (obtained from averaging the embeddings of each one) **[Figure A]**

References

1. Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606, 2016
2. Yoon Kim. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882, 2014.
3. Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15(1):1929–1958, 2014
4. Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. Learning word vectors for 157 languages. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018), 2018
5. West University of Timisoara, Faculty of Psychology. The Psychotherapy of anxiety and depression (psitad). * https://e-cbt.ro/program/psitad/.