MUNI FI

Detecting Online Risks and Supportive Interaction in Instant Messenger Conversations

using Czech Transformers

Ondřej Sotolář
xsotolar@fi.muni.cz
Faculty of Informatics, Masaryk University

December 10, 2021

Contents

Introduction

- New language domain => uncertain performance

Introduction

- New language domain => uncertain performance
- Domain of interest: IM conversation of adolescents in Czech
  - significant portion of data collectable on-device
  - insights into what the users actually do on their devices compared to what they believe, say, or reflect they do

Introduction

- New language domain => uncertain performance
- Domain of interest: IM conversation of adolescents in Czech
  - significant portion of data collectable on-device
  - insights into what the users actually do on their devices compared to what they believe, say, or reflect they do
- Task: Text Classification
  - Supportive Interaction
  - Online Risks

Language Domain: IM Conversations

- Anonymized [1],
- in Czech,
- in private,
- through IM communication tools,
- communication between adolescents.

Language Domain: IM Conversations

- Anonymized [1],
- in Czech,
- in private,
- through IM communication tools,
- communication between adolescents.

- Informal language: syntactic, stylistic, and grammatical quality lower than encyclopedic and journalistic styles in pre-trained models,
- Un-quantified difference to public text from social networks,

Language Domain: IM Conversations

- Anonymized [1],
- in Czech,
- in private,
- through IM communication tools,
- communication between adolescents.

- Informal language: syntactic, stylistic, and grammatical quality lower than encyclopedic and journalistic styles in pre-trained models,
- Un-quantified difference to public text from social networks,

=> not within-domain [2] of training corpora of Czech transformers.

Related Work

- Adolescents' social network data
  - BlackBerry project [3] that examined adolescents' text messages
  - [4] classify social network messages of adolescents from various sources,
  - sentiment analysis dataset of Czech Facebook posts in [5]
    - => all pre-date embeddings

Related Work

- Adolescents' social network data
  - BlackBerry project [3] that examined adolescents' text messages
  - [4] classify social network messages of adolescents from various sources,
  - sentiment analysis dataset of Czech Facebook posts in [5]
    - => all pre-date embeddings
- Text classification
  - systematic review of the Neural Network architectures for TC [6]
  - Czech transformer comparison in [7]

Corpus: anonymized Facebbok Msg. (N=17, 13-17 yo)

Category	# rows labeled	P(cat)	κ	# blocks
Supportive Interactions (N=270,760)				
Information Support	9967	5.08	0.685	5325
Emotional Support	9669	4.93	0.639	7284
Social Companionship	5317	2.71	0.599	4047
Appraisal	2338	1.19	0.65	1874
Instrumental Support	3331	1.7	0.604	2482
Online Risks (N=196,196)				
Aggression, Harassment, Hate	5382	1.99	0.47	3737
Mental Health Problems	3098	1.14	0.46	1605
Alcohol, Drugs	2288	1.17	0.609	1625
Weight Loss, Diets	91	0.03	-	46
Sexual Content	3563	1.32	0.485	2949

Compared Models

- Embeddings -> AVG -> softmax
  - Fasttext [8]

Compared Models

- Embeddings -> AVG -> softmax
  - Fasttext [8]
- BERT -> [CLS] -> softmax
  - Czert-B [9]
  - FERNET-C5 [7]

Compared Models

- Embeddings -> AVG -> softmax
  - Fasttext [8]
- BERT -> [CLS] -> softmax
  - Czert-B [9]
  - FERNET-C5 [7]
- RoBERTa -> [CLS] -> softmax
  - RobeCzech [10]
- ELECTRA -> [CLS] -> softmax
  - Small-E-Czech [11]

Results: F1

Category	Czert-B	FERNET-C5		Fasttext
		RobeCzech	Small-E-Czech	
Supportive Interactions				
Information Support	71.95	**75.44**	74.91	73.73	70.89
Emotional Support	74.63	76.67	**78.2**	72.94	74.05
Social Companionship	79.58	83.99	**84.74**	81.73	79.85
Appraisal	81.23	81.49	**85.87**	70.14	82.07
Instrumental Support	76.63	**82.12**	79.6	78.35	75.67
Online Risks				
Aggression, Harassment, Hate	84.41	**88.23**	**88.23**	83.24	83.24
Mental Health Problems	72.49	82.82	**85.11**	77.05	64.39
Alcohol, Drugs	87.17	**89.66**	87.6	81.40	63.22
Weight Loss, Diets	-	-	-	-	-
Sexual Content	70.62	74.33	**81.94**	67.72	63.16

Error Analysis: high-certainty errors

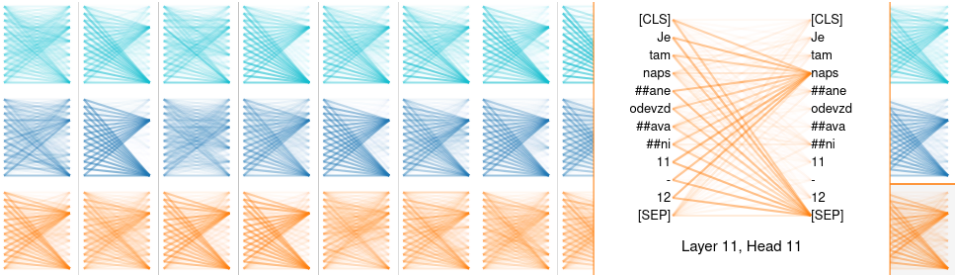

Figure: bert-viz Model-view of the last three layers of Czert for the high-certainty misclassified sequence, for the sub-category Information Support: *'Je tam napsane odevzdavani 11-12'*. The model is heavily biased towards the token *'naps'*, a part of the verb *'written'*, an expected keyword of this category.

Also, annotator disagreements.

Error Analysis: low-certainty errors

- Considerably shorter than the high-certainty ones
- One-word and text fragment samples
  - => In combination with the lack of context enough input

Discussion

- Fasttext: fast hyperparameter search, but overfits

Discussion

- Fasttext: fast hyperparameter search, but overfits
- Transformers: slow hyp. search, but generalizes well
  - all results should improve with more sophisticated hyp. search

Discussion

- Fasttext: fast hyperparameter search, but overfits
- Transformers: slow hyp. search, but generalizes well
  - all results should improve with more sophisticated hyp. search
- Solve high-certainty errors:
  - regularize (dropout, augmentation)
  - get more votes on annotator disagreements
  - finish the multi-label annotation

Discussion

- Fasttext: fast hyperparameter search, but overfits
- Transformers: slow hyp. search, but generalizes well
  - all results should improve with more sophisticated hyp. search
- Solve high-certainty errors:
  - regularize (dropout, augmentation)
  - get more votes on annotator disagreements
  - finish the multi-label annotation
- Solve low-certainty errors:
  - improve preprocessing
  - add context

Bibliography I

[1] Ondřej Sotolář, Jaromır Plhák, and David Šmahel. Towards personal data anonymization for social messaging. In *International Conference on Text, Speech, and Dialogue*, pages 281–292. Springer, 2021.

[2] Hady Elsahar and Matthias Gallé. To annotate or not? predicting performance drop under domain shift. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2163–2173, 2019.

[3] Marion K Underwood, Samuel E Ehrenreich, David More, Jerome S Solis, and Dawn Y Brinkley. The blackberry project: the hidden world of adolescents' text messaging and relations with internalizing symptoms. *Journal of Research on Adolescence*, 25(1):101–117, 2015.

[4] Suppawong Tuarob, Conrad S. Tucker, Marcel Salathe, and Nilam Ram. An ensemble heterogeneous classification methodology for discovering health-related knowledge in social media messages. *Journal of Biomedical Informatics*, 49:255–268, 2014. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2014.03.005. URL: https://www.sciencedirect.com/science/article/pii/S1532046414000628.

[5] Ivan Habernal, Tomáš Ptáček, and Josef Steinberger. Sentiment analysis in czech social media using supervised machine learning. In *Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 65–74, 2013.

[6] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. Deep learning–based text classification: a comprehensive review. *ACM Computing Surveys (CSUR)*, 54(3):1–40, 2021.

[7] Jan Lehečka and Jan Švec. Comparison of czech transformers on text classification tasks. In *International Conference on Statistical Language and Speech Processing*, pages 27–37. Springer, 2021.

Bibliography II

[8] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[9] Jakub Sido, Ondřej Pražák, Pavel Přibáň, Jan Pašek, Michal Seják, and Miloslav Konopık. Czert–czech bert-like model for language representation. *arXiv preprint arXiv:2103.13031*, 2021.

[10] Milan Straka, Jakub Náplava, Jana Straková, and David Samuel. Robeczech: czech roberta, a monolingual contextualized language representation model. *arXiv preprint arXiv:2105.11314*, 2021.

[11] Seznam.cz. Small-e-czech. `https://github.com/seznam/small-e-czech`, 2021.

MUNI
FACULTY
OF INFORMATICS