Detecting Online Risks and Supportive Interaction in Instant Messenger Conversations using Czech Transformers

Ondřej Sotolář, Jaromír Plhák, Michal Tkaczyk, Michaela Lebedíková, and David Šmahel

IRTIS, Faculty of Informatics, Masaryk University,
Botanicka 68a, 602 00, Brno, Czech Republic
{xsotolar,xplhak}@fi.muni.cz, tkaczyk@fss.muni.cz,
{m.lebedikova,davs}@mail.muni.cz

Abstract. We present a comparison of state-of-the-art models for text classification of Online Risks and Supportive Interaction in anonymized Instant Messenger conversations held in Czech. We compare the transformer models Czert, RobeCzech, and FERNET-C5 with the Fasttext classifier as a baseline. For the comparison, we build a novel dataset with five subcategories for the Online Risks and five for the Supportive Interaction. We solve the balanced classification problem achieving 75.44 - 89.66 F1 score depending on the category. Our results show that the transformer models perform consistently better than the baseline.

Keywords: Online Risks · Supportive Interaction · Facebook Messenger · Text Classification

1 Introduction

Starting Natural Language Processing research in a new language domain brings uncertainty about how existing models and tools will perform in it. In such case, it is a good practice to compare several candidate models and select the best-performing ones to develop further.

In our case, the domain of interest is composed of anonymized Instant Messenger (IM) conversations of Czech adolescents conducted in Czech. Current research [1] is trying to examine the effect of smartphone use on the well-being of adolescents through analyzing data collected on-device. The IM conversations constitute a significant portion of this data, and the classification will allow for the measurement of smartphone use with high validity. It will provide insights into what the users actually do on their devices in IM conversations and what is the possible impact on their well-being.

So far, this specific domain has been under-researched in NLP. We try to establish the difficulty of classifying the IM messages (without context) into the respective sub-categories of the Online Risks and Supportive Interaction categories, described in 3.2. We perform a model comparison by fine-tuning four new Czech transformer models using the Fasttext classifier as the baseline.

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2021, pp. 19–28, 2021.

2 Related work

The work in the domain of text obtained from social networks is highly diverse, as it is an active area of research. Close in the language domain and study participants' age is the BlackBerry project [27] that examined adolescents' text messages. The authors of [26] classify social network messages of adolescents from various sources by ensembling various statistical machine learning models trained on word N-grams. Both of the mentioned works were carried out on English corpora. In Czech, sentiment analysis was carried out on a dataset of Czech Facebook posts in [9]. All of these works use methods pre-dating the widespread use of text embeddings.

Contemporary NLP classification methods leverage the strength of pre-trained deep language representation models, which have surpassed statistical approaches, such as used in [15] in various Text Classification (TC) tasks. A systematic review of the Neural Network (NN) architectures in [17] gave us guidance on the choice of the baseline model. We chose those transformer models that achieve SOTA for Czech, based on the comparison in [14]. Until very recently, the multilingual models SlavicBERT [2], mBERT [7], and XLM-RoBERTa-base [6] achieved SOTA results for Czech. They were recently surpassed in classification by BERT-based models Czert-B [22], FERNET-C5 [14], and RoBERTa based RobeCzech [25] that achieve comparable results with larger multilingual models, such as the XLM-RoBERTa-large [6]. Czert and RobeCzech are trained on a combination of Czech National Corpus [12], Czech Wikipedia dump, and Czech news crawl. FERNET is trained on C5, a new Common Crawl-based corpus. For completeness, we also measured the smaller ELECTRA [4] model, Small-E-Czech [21], trained on a Czech web crawl and search queries.

3 Methods

3.1 Language Domain

The domain of private IM conversation is much less explored than the domain of publicly available text gathered from social networks. Arguably, because such data are hard to obtain, they may contain sensitive information and thus need to be anonymized, which is challenging. To solve it, we used the methods described in [24]. Some features and issues in this domain are given by the fact that the communication is held in Czech, it is conducted in private, through IM communication tools, and it is communication between adolescents, their peers, and sometimes also caregivers, such as parents. The dialogues are commonly conducted in informal language. Their syntactic, stylistic, and grammatical quality is considerably lower than formal styles, such as the encyclopedic and journalistic styles, predominantly represented in the pre-trained models' training corpora. The difference from the informal but for-public intended text, such as status messages from social networks, forums, discussions, and chat room conversations, all of which also occur in the training sets of language models, remains un-quantified. Ultimately, when using such language models

on tasks in our particular domain, the data cannot be considered to be within-domain [8].

3.2 Annotated Corpus

We have created an annotated corpus of Facebook Messenger conversations of adolescents participating in our study (*N=17, 13-17 years old*). Out of all collected conversation records, we drew stratified batches of conversation samples of representative size to ensure variability of the phenomena under consideration in the annotated corpora. The total size of the annotated corpora for SI and OR, expressed in number of rows of text, is (*N=270,760, N=196,196*), also shown in Table 2 among other statistics.

The **Annotation categories**, i.e. Supportive Interactions (SI) and Online Risks (OR) were derived from relevant research and theory in the fields of psychology and communication [18,5,3,23,20]. Both categories refer to different, conceptually unrelated types of communicative behavior, and they differ from each other also in terms of their linguistic features. SI covers a range of communicative behaviors oriented at achieving the same intention, which is providing social support through interpersonal communication to another participant of conversation. Data falling under the OR category are defined by the mere fact of referring to a particular topic, i.e., different types of risks to adolescents' health and development, no matter whether at the interactional or ideational level of language [10], e.g., it could be instances of online aggression directed at conversation participants but also references to aggressive behavior conducted by someone else offline.

Since each of the two categories contained several sub-categories (see Table 1), the annotation was posed as a multi-label problem for each category[1]. Labels could be assigned to either a *single row* of a conversation or a *block of consecutive rows*. In order to create contextual units for the annotators to evaluate rows or blocks, they were delimited by the conversation turns of chat participants.

We used Cohen's κ [13] to measure the **IAA** because each category has been annotated by two annotators (see Table 2 for the achieved IAA). In the case of SI, positive examples were frequent enough, and we achieved a satisfactory level of IAA. It oscillated between batches between moderate (.41 to .60) and substantial (.61 to .80) and was constantly improving. For OR, the occurrences were rarer; thus, we abandoned the random sampling of batches. Instead, we first draw samples that scored the highest with preliminary classifiers trained on the previously annotated data, which improved the yield. The IAA oscillated between slight (.21-40) and moderate (.41 to .60) and improved inconsistently.

To sum up, for each category, we have obtained labels of different quality. Especially in the case of OR, the reliability of the data is not entirely satisfactory.

[1] While the annotation problem was indeed multi-label, due to various constraints, the annotators always assigned only the most probable label and indicated that there could be more labels on the particular line, leaving it unfinished. This effectively makes the problem multi-class.

Table 1: Description of categories.

Category	Description
Supportive Interactions	
Information Support	provide useful knowledge and information
Emotional Support	express intimacy, caring, liking, empathy, or sympathy
Social Companionship	convey a sense of belonging, inclusivity, will to spend time together in leisure, recreational activ.
Appraisal	express acceptance, respect, validation, esteem
Instrumental Support	offer practical help or resources, assistance in getting necessary tasks done
Online Risks	
Aggression, Harassment, Hate	use of or reference to offensive language and slander to cause harm
Mental Health Problems	reference to long-standing MH problems: suicide, self-harm, depression, eating disorders
Alcohol, Drugs	reference to experiences with alcohol and drugs
Weight Loss, Diets	discussions of weight-loss, working out and diets
Sexual Content	sexual or sexually suggestive discussions

Table 2: Statistics of the annotated corpus.

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

3.3 Training Dataset

The phenomena we are classifying are rare events (see column $P(cat)$ for the percentage of rows in Table 2). Solving the imbalance of a dataset that would respect the original distribution is not among the goals of this article; therefore, we built binarized balanced datasets. They are composed of all the positively

labeled data per respective class, complemented by an equivalent amount of semi-randomly chosen negatively labeled data, in both cases labeled by at least one annotator. The positive labels often span across multiple rows of a single participant. We concatenated such cases into blocks (column *blocks* in Table 2) of one or more consecutive rows with the same label, thus reducing the overall example count. The negative examples were selected randomly but with paying attention to the distribution of several features of the positive blocks (character count, line count, number of participants), in an effort to minimize the statistical bias introduced by the undersampling. Preprocessing consisted only of lowercasing and removal of examples shorter than five characters.

3.4 Baseline Model

We chose the Fasttext classifier [11] as our baseline model, which is based on a shallow feed-forward NN using word embeddings as inputs. It can achieve high accuracy on many TC benchmarks, especially on datasets with high syntactic variance, which is our case. We have used the automatic tuning feature to determine ideal hyperparameters. We have also measured the impact of using pre-trained embeddings.

3.5 Transformer Models

BERT [7], is a transformer model pre-trained on a large corpus in a self-supervised fashion, with the Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) objectives. In MLM, the model randomly masks a portion (15%) of the words in the input, then inputs the sequence in the model and learns to predict the masked words. This is different from recurrent neural networks or from auto-regressive models like GPT [19] which mask the future tokens. In NSP, the model, given two sequences, learns to predict if the second sequence follows the first one. This way, the model learns low-level, bi-directional representations of the target language from which we can create a classifier by a process called fine-tuning. The model outputs a special token [CLS] that encodes the final hidden state h of the BERT model after inputting the sequence. Finally, a softmax layer is added on top of the model to predict the probability of label l:

$$p(l|\mathbf{h}) = softmax(W\mathbf{h}), \tag{1}$$

where W are the new layer's parameters which are learned by minimizing the cross-entropy loss using the task-specific dataset.

There are several variants of BERT that alter some of its components to either improve it, shrink it, or achieve some other goal. RoBERTa [16], whose goal is to improve the absolute performance, differs from BERT in the masking process, tokenization, and pre-training. In BERT, the masking is performed only once at data preparation time: the model masks each sequence a fixed number of times. Therefore, at training time, the model will only use those previously generated variations. On the other hand, in RoBERTa, the masking is done

during training, each time a sequence is incorporated in a minibatch. As a result, the number of potentially different masked versions of each sequence is not bounded like in BERT. RoBERTa additionally uses a different style of BPE tokenization (same as GPT-2). While BERT highlights the merging of two subsequent tokens, RoBERTa's tokenizer instead highlights the start of a new token with a specific unicode character to avoid the use of whitespaces. Furthermore, RoBERTa removes the NSP task from pre-training. Thus, in theory, the RoBERTa model is more effectively regularized and can be trained for more epochs to achieve better results.

ELECTRA is a BERT-based architecture, whose goal is to shrink the network. Instead of using a masking token for the MLM, it provides plausible replacements sampled from a generator network. It offers solid performance while keeping the network several times smaller than BERT or RoBERTa.

4 Results

We summarize the experimental results in Table 3. They partially confirm the results of [14] by showing that the FERNET-C5 model performs among the two best models across our categories. However, in most experiments, RobeCzech could achieve comparable or better performance. The Czert model, being the first Czech transformer, is expectantly performing consistently worse than both the newer models. The performance of Small-E-Czech is also worse compared to the best models in all cases. On the other hand, the model is significantly smaller, and its training is faster. Surprisingly, the much simpler Fasttext model can approach the performance of the transformer models, provided that there is enough training data. On the categories with fewer training examples, the strength of transfer learning showed in the much larger gap in performance between the transformer models and Fasttext.

4.1 Hyperparameters

We optimized the hyperparameters globally for all transformer models and all categories at once. We based on the default values on [14] and used grid search only to slightly tweak them to fit our dataset and hardware. The results can be therefore easily compared with the previous works. The reported results use the following settings: (*batch size=128, peak learning rate=1e-5, warmup steps=1/3 no. total steps w. linear decay*). We trained for 20 epochs; however, with early stopping, which showed the ideal number of epochs be between 7-10, which confirms the 10 epoch setting used by [14].

For the Fasttext model, we used it's automatic hyperparameter optimization feature that resulted in (*dim=300, epoch=36, lr=0.05, lrUpdateRate=100, maxn=5, minn=2*) with other parameters left on default. Experiments with pretrained Fastext embeddings did not result in improvement.

Table 3: Results of binary classification. We report the cross-validated F1 score.

Category	Czert-B	RobeCzech	FERNET-C5	Small-E-Czech	Fasttext
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

4.2 Error Analysis

Many misclassified samples point to the obvious lack of context for each example. This causes the model to miss many finer points of the annotation manual, such as the instruction to assign a negative label to samples with a sarcastic connotation (sometimes expressed with an emoticon). However, including context would require a modification of the compared models, which is not among the goals of this article.

The analysis of high-certainty but misclassified predictions revealed that many samples rely on only one or two keywords, as shown in the model-view diagram of the *bertviz* tool [28] in Figure 1. If such keywords form a majority on one side of the binary classifier, it tends to classify all such samples into one class, some of them wrongly. Another reason for this class of error that we confirmed is that some of the misclassified samples are actually classified correctly, but the annotators disagreed on the label.

The analysis of low-certainty samples shows that these are, on average, considerably shorter than the high-certainty ones. They contain a number of one-word and text fragment samples which, in combination with the lack of context, does not provide the classifier enough input to perform well.

4.3 Discussion and Further Work

Our investigation yielded some interesting findings, such as the fact that the Fasttext model can rival the much larger transformers even without pretraining. While being a simpler model, the original implementation of the model is very efficient. That enables the search for hyperparameters to be several orders of magnitude faster than for the transformers models in the HuggingFace [29] library, which we used.

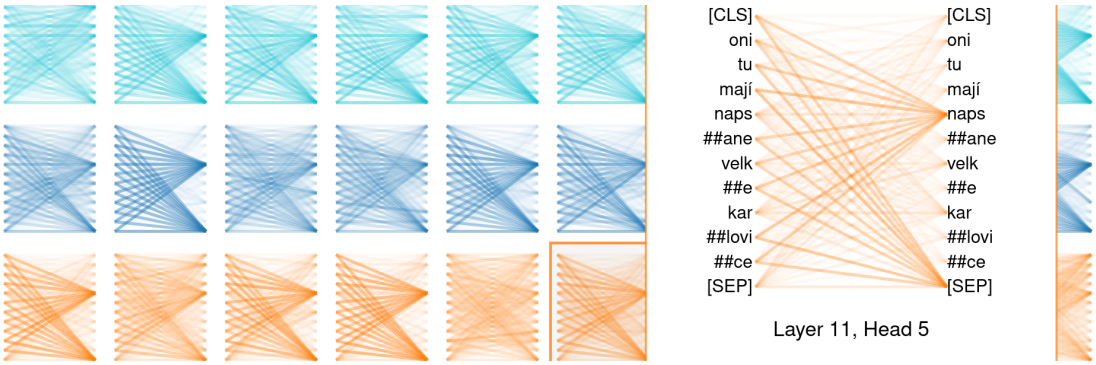

Fig. 1: Model-view of the last three layers of Czert for the high-certainty misclassified sequence, for the sub-category Information Support: *'oni tu mají napsane velke karlovice'*. The tokens on the left side of the bipartite graph provide attention to the right-side ones. We can see the repeating pattern on both the detail of the last layer's last head and the miniatures of other heads. The classifier is heavily biased towards the token *'naps'*, a part of the verb *'written'*, an expected keyword of this category.

Overall, we consider the results of this work to set solid baselines, to which new results can be compared. For further work, we suggest improving the regularization of the dataset by dropout or data augmentation, which could improve performance on the high-certainty misclassified samples by addressing the keyword issue. Additionally, further cleaning of the low-certainty samples could improve classification on this class of error. Furthermore, a more sophisticated hyperparameter search could improve the performance of the transformer models.

However, the obvious next step should be modeling the impact of the context of the messages. For example, [30] has shown that as far back as two months of previous dialogue can help improve the classification of new messages.

5 Conclusions

We have compared four new Czech transformer models on the task of text classification. We have shown that they provide a consistent improvement over the baseline Fasttext model and partially confirm the results from previous works, showing that the FERNET and RobeCzech models perform better than the Czert or Small-E-Czech models. In doing so, we prove that in the language domain of our dataset, i.e., short IM messages held in Czech, classification can be successfully performed even without the messages' context. We have built new annotated corpora for each of the sub-categories of Supportive Interactions and Online Risks categories, created datasets of them, and trained text classification models that have achieved 75.44 - 89.66 F1 score.

Acknowledgements. This work has received funding from the Czech Science Foundation, project no. 19-27828X.

References

1. Understanding the impact of technology on adolescent's well-being (FUTURE). https://irtis.muni.cz/research/projects/future, accessed: 2021-10-28
2. Arkhipov, M., Trofimova, M., Kuratov, Y., Sorokin, A.: Tuning multilingual transformers for language-specific named entity recognition. In: Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing. pp. 89–93 (2019)
3. Bonino, S., Cattelino, E., Ciairano, S.: Adolescents and risk: Behaviors, functions and protective factors. Springer Milan (2005), https://books.google.com.bn/books?id=FcFeNk_38-IC
4. Clark, K., Luong, M.T., Le, Q.V., Manning, C.D.: Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555 (2020)
5. Cohen, S., Underwood, L.G., Gottlieb, B.H.: Social Support Measurement and InterventionA Guide for Health and Social Scientists: A Guide for Health and Social Scientists. Oxford University Press, New York, NY (09 2015). https://doi.org/10.1093/med:psych/9780195126709.001.0001
6. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116 (2019)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
8. Elsahar, H., Gallé, M.: 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). pp. 2163–2173 (2019)
9. Habernal, I., Ptáček, T., Steinberger, J.: 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. pp. 65–74 (2013)
10. Halliday, M.A.K., Matthiessen, C.M.I.M.: An introduction to functional grammar / M.A.K. Halliday. Hodder Arnold London, 3rd ed. / rev. by christian m.i.m. matthiessen. edn. (2004)
11. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759 (2016)
12. Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Chlumská, L., Jelínek, T., Kováříková, D., Petkevic, V., Procházka, P., et al.: Syn2015: Representative corpus of contemporary written czech. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 2522–2528 (2016)
13. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1), 159–174 (1977)
14. Lehečka, J., Švec, J.: Comparison of czech transformers on text classification tasks. In: International Conference on Statistical Language and Speech Processing. pp. 27–37. Springer (2021)
15. Linkov, V., Smerk, P., Li, B., Smahel, D.: Personality perception in instant messenger communication in the czech republic and people's republic of china. Studia Psychologica **56**(4), 287 (2014)
16. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)

17. Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., Gao, J.: Deep learning–based text classification: A comprehensive review. ACM Computing Surveys (CSUR) **54**(3), 1–40 (2021)
18. Nick, E.A., Cole, D.A., Cho, S.J., Darcy K. Smith, T.G.C., Zelkowitz, R.: The online social support scale: Measure development and validation. Psychological assessment **30**(9), 1127–1143 (2018). https://doi.org/10.1037/pas0000558
19. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training (2018)
20. Remschmidt, H., Nurcombe, B., Belfer, M., Sartorius, N., Okasha, A.: The Mental Health of Children and Adolescents: An area of global neglect. World Psychiatric Association, Wiley (2007), `https://books.google.cz/books?id=bENaj6hBuQUC`
21. Seznam.cz: Small-e-czech. `https://github.com/seznam/small-e-czech` (2021)
22. Sido, J., Pražák, O., Přibáň, P., Pašek, J., Seják, M., Konopík, M.: Czert–czech bert-like model for language representation. arXiv preprint arXiv:2103.13031 (2021)
23. Smahel, D., Machackova, H., Mascheroni, G., Dedkova, L., Staksrud, E., Ólafsson, K., Livingstone, S., Hasebrink, U.: EU Kids Online 2020: survey results from 19 countries. EU Kids Online (2020)
24. Sotolář, O., Plhák, J., Šmahel, D.: Towards personal data anonymization for social messaging. In: International Conference on Text, Speech, and Dialogue. pp. 281–292. Springer (2021)
25. Straka, M., Náplava, J., Straková, J., Samuel, D.: Robeczech: Czech roberta, a monolingual contextualized language representation model. arXiv preprint arXiv:2105.11314 (2021)
26. Tuarob, S., Tucker, C.S., Salathe, M., Ram, N.: An ensemble heterogeneous classification methodology for discovering health-related knowledge in social media messages. Journal of Biomedical Informatics **49**, 255–268 (2014). https://doi.org/https://doi.org/10.1016/j.jbi.2014.03.005, `https://www.sciencedirect.com/science/article/pii/S1532046414000628`
27. Underwood, M.K., Ehrenreich, S.E., More, D., Solis, J.S., Brinkley, D.Y.: 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)
28. Vig, J.: Bertviz: A tool for visualizing multihead self-attention in the bert model. In: ICLR Workshop: Debugging Machine Learning Models (2019)
29. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al.: Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771 (2019)
30. Yang, D., Flek, L.: Towards user-centric text-to-text generation: A survey. In: Ekštein, K., Pártl, F., Konopík, M. (eds.) Text, Speech, and Dialogue. pp. 3–22. Springer International Publishing, Cham (2021)