Classification of Adolescents' Risky Behavior in Instant Messaging Conversations

Jaromír Plhák **Ondřej Sotolář** **Michaela Lebedíková** **David Šmahel**
Faculty of Informatics, Masaryk University, Brno, Czech Republic

Abstract

Previous research on detecting risky online behavior has been rather scattered, typically identifying single risks in online samples. To our knowledge, the presented research is the first that presents a process of building models that can efficiently detect the following four online risky behavior: (1) aggression, harassment, hate; (2) mental health; (3) use of alcohol, and drugs; and (4) sexting. Furthermore, the corpora in this research are unique because of the usage of private instant messaging conversations in the Czech language provided by adolescents. The combination of publicly unavailable and unique data with high-quality annotations of specific psychological phenomena allowed us for precise detection using transformer machine learning models that can handle sequential data and involve the context of utterances. The impact of the context length and text augmentation on model efficiency is discussed in detail. The final model provides promising results with an acceptable F1 score. Therefore, we believe that the model could be used in various applications, e.g., parental applications, chatbots, or services provided by Internet providers. Future research could investigate the usage of the model in other languages.

1 INTRODUCTION

Instant messaging (IM) is a type of online communication that allows for the synchronous exchange of text, images, voice, and videos between two or more people (Huang and Leung, 2009) using applications and platforms such as Messenger or WhatsApp. This type of communication is prevalent among adolescents (Benotsch et al., 2012), who far outnumber adults in their use of IM (Valkenburg and Peter, 2011). Across Europe, more than half of adolescents use their smartphones daily or several times a day. The ubiquity of smartphones and other devices allows them to be continuously online. The use of the internet to communicate with friends and family ranges from 14% to 77% across countries (Smahel et al., 2020). IM, for example, allows adolescents to practice social skills, increasing their ability to form offline relationships (Koutamanis et al., 2013), explore their identity, and find information (Valkenburg and Peter, 2011). On the other hand, IM entails a variety of risks, such as cyber-aggression (Álvarez-García et al., 2018) or online solicitation (Valkenburg and Peter, 2011).

Research shows, for example, that more than 20% of European adolescents experienced victimization associated with aggression and cyberhate, and up to 39% received sexual messages in the past year (Smahel et al., 2020). Regarding offline risks, 18.8% of American adolescents seriously considered suicide, and 24.1% tried cigarettes at some point in their life (Underwood et al., 2020).

In spite of adolescent experiences with risks and the large volume of conversations adolescents engage in online, the occurrence of risks in their messages remains relatively low. This makes it difficult and inefficient to use conventional social science methods, such as content analysis, to analyze adolescent messages. Machine learning is one way to overcome this problem. Based on the current research on both online and offline risks for adolescents (Smahel et al., 2020; Underwood et al., 2020), we employ machine learning for detecting the most common risks: aggression and cyberhate, discussion of mental health issues, use of alcohol and drugs, and sexting in adolescent IM conversations in the Czech language. To the best of the authors' knowledge, we are the first to employ a machine learning approach in this area on real IM conversations provided by adolescents.

It was shown in previous research that the language style in this domain is substantially different from that of other publicly available corpora, and thus, models trained on texts from a different domain achieve a substantially lower performance (Sotolář, Plhák, Tkaczyk, et al., 2021). Corpora

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. Copyright 2023 by the author(s).

usually do not contain private IM data as they are hard to obtain, e.g., using web scrapping, and they usually contain sensitive information that needs to be anonymized. Therefore, we have created novel corpora from data from IM conversations provided by adolescents. Even though the corpora cannot be made publicly available due to the privacy of our participants, such unique data allows us to produce precise models to predict online risk occurrences.

To provide efficient online risk detection models, these corpora have to be annotated by high-quality annotations. We employ media and communications researchers as supervisors to map state-of-the-art research in this area, create annotation manuals, and train annotators that process the corpora of anonymized conversations using our newly developed annotation tool. Even though the annotated conversations were mainly in the Czech language, the annotation manuals can also be reused for adolescents' conversations in other languages.

Subsequently, we utilized transformer machine learning models that can handle sequential data and involve the context of utterances in conversations. Both monolingual and multilingual pre-trained models were employed in our experiments to improve precision and recall. The resulting models efficiently detect online risks that can be used, for example, in parental applications or applications that supervise users' well-being.

2 RELATED WORK

Detection of online risks in IM conversations of adolescents is a multi-disciplinary task, combining well-defined psychological tasks with NLP, and many methods can be found in the existing solutions. However, the combination of our source data type and the detected phenomena is unique, and the language domain of informal Czech dialogues conducted in private is specific.

Relevant methods reflect the general state of the NLP field - that is, the preference of *deep learning* models, particularly with *transfer learning* over classical ML methods relying on feature selection and extraction such as Dahiya, Mohta, and Jain (2020) and Wahyono et al. (2021).

This claim is best supported by both systematic reviews of work on related tasks such as *emotion detection* (Acheampong, Wenyu, and Nunoo-Mensah, 2020) and the results of competing models on shared tasks as used in the SemEval-2019: EmoContext (Chatterjee et al., 2019). In this task, the provided dataset consisted of dialogues, and the goal was to classify three emotion classes (angry, sad, happy) using two previous utterances. All top-performing systems used various embeddings as the text representation, with some combining feature-based models with neural models. However, various neural architectures, often leveraging transfer learning, formed the majority of the top solutions.

Another close task regarding the type of context, i.e., a temporal sequence of proceeding utterances, is the task of *dialogue act recognition*. The current methods for this task also utilize embeddings to express the utterances and build additional layers of neural networks above it, such as in Khanpour, Guntakandla, and Nielsen (2016) with the newer works leveraging transfer learning by using pre-trained language models (Martínek, Král, et al., 2019; Martínek, Cerisara, et al., 2021).

In *short text classification*, a more general related task, context utilization is vital. Here, the context is often not a temporal sequence, and it is not required to be textual. However, we can find instances where the context is similar. For instance, J. Y. Lee and Dernoncourt (2016) and Chen et al. (2019) present experiments with the number and shape of hierarchies of sequential layers of LSTMs and CNNs to capture the context as well as the meaning of every single utterance.

For suspected long-term dependencies in dialogues, such as depression detection, using the intra-participant differences has proved beneficial (Flek, 2020). When a whole dialogue is labeled with a single label, the metric of the earliness of detection can be used (Vogt, Leser, and Akbik, 2021).

2.1 Detecting Online Risks using ML Methods

This paper deals with the detection of the following online risks: (1) aggression, harassment, hate, (2) mental health problems, (3) alcohol, drugs, and (4) sexual content. Previous studies utilize various approaches and datasets, typically from online social networks. However, no previous studies have utilized authentic private IM conversations from adolescents, which is a strength of our study, as we can examine how adolescents discuss these online risks and adapt detection to their specific vocabulary.

Classification of cyberbullying is a novel task (Rosa, Pereira, et al., 2019). First, a recent review found that studies typically classify cyberbullying in its widest sense, including related constructs, such as aggression, profanities, or racism detection (Rosa, Matos, et al., 2018; Zhao, Zhou, and Mao, 2016). Second, studies train their algorithms on different datasets, such as Twitter (Al-garadi, Varathan, and Ravana, 2016) or YouTube (Dadvar, Trieschnigg, and Jong, 2013). The samples are usually obtained through APIs or website scraping, thus making the results of studies incomparable (Rosa, Pereira, et al., 2019).

Regarding machine learning applications in the area of mental health, studies are mostly aimed at detection and diagnosis (Shatte, Hutchinson, and Teague, 2019). Many studies are detecting mental health problems in online social networks (Rahman et al., 2020) which are considered most important when it comes to addressing an individual's mental health issues efficiently (De Choudhury,

Table 1: Overview of Number of Utterances and Inter-Annotator Agreement (Cohen's κ)

ONLINE RISK	ANNOTATED BY AT LEAST ONE ANNOT.	κ	REVIEWED BY SUPERVISOR	GOLD STANDARD
(1) Aggression, harassment, hate	5393 (1.979%)	.470	3898	3178
(2) Mental health problems	3101 (1.138%)	.460	1729	2236
(3) Alcohol, drugs	2301 (0.845%)	.609	1294	1990
(4) Sexual content	3550 (1.303%)	.485	2118	3116

2013). Studies in the area of mental health are broad, such as detecting suicidal tendencies on Twitter (O'Dea et al., 2015), detecting users with depression on Facebook (Wongkoblap, Vadillo, and Curcin, 2018) or detecting levels of stress based on interactions between users on Weibo (Lin et al., 2017). However, many challenges in this domain prevail, such as conceptualizing mental health too broadly and failing to acknowledge the multitude of mental health problems, sparsity of data, or multilingualism of corpora from social networks (Rahman et al., 2020). In our project, mental health is conceptualized as a long-term problem that clearly affects the participants in a conversation and includes known psychological illnesses (such as eating disorders, anxiety, and depression). Our goal was to detect only explicit mentions of the problems, not diagnose the participants based on symptoms. Therefore, we were able to capture discussions of mental health accurately, and thus we were able to grasp the multitude of mental health problems without being too broad.

In the case of tobacco, drug, and alcohol use detection, studies took advantage of social media data. For example, one study involved corpora of tweets selected based on smoking-related slang and found that tobacco and related drugs are often discussed online (Pant et al., 2019). A study related to drug use focused on detecting drug dealers on Instagram (Li et al., 2019). Regarding alcohol use, studies are mainly based on detecting alcohol use and misuse in clinical texts (Alzoubi et al., 2018; Afshar et al., 2019; Afzali et al., 2019). However, detecting alcohol use in tweets is also highly feasible (Aphinyanaphongs et al., 2014).

Finally, in the area of sexting, most studies are focused on intimate violence, sexual offenders, and cybergrooming. Several studies used the Perverted Justice dataset with convicted sex offenders and volunteers posing as teenagers as a base for machine learning training (Razi, 2020). A study from South Korea used big data, combining sources such as local news websites, social networks, or bulletin boards to examine trends and patterns of youth sexting. However, data were selected only by using keywords, such as ”sexting,” ”porn sharing,” or ”adult video distribution.” The authors found that adolescents sext in order to gain attention from peers and that file-sharing is more common than image distribution (J. Song, T. M. Song, and J. R. Lee, 2018). The main shortcoming of studies in this area is that they overlook the potential of sexting to be a positive activity and thus fail to detect sexting among adolescents that is positive and in line with their sexual development.

3 METHODS

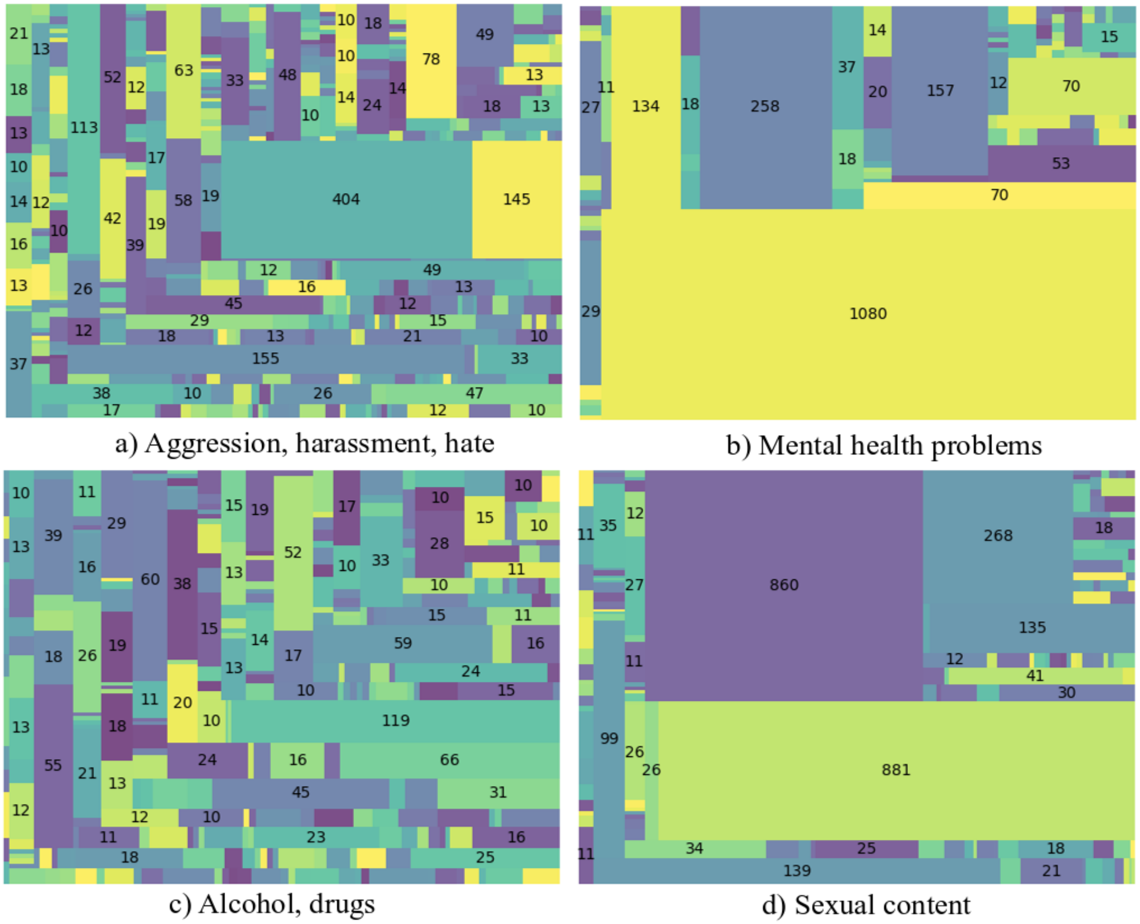
3.1 Corpora

The original corpora were created from the files of 22 users (13-17 years old) that were manually exported from the Messenger communication tool developed by Meta Platforms. Due to the sensitivity of the data, we strictly adhere to the legal and ethical recommendations of the Research Ethics Committee. The data are suitably anonymized for the researchers and annotators. However, there is a possibility to identify the participants by inference (e.g., by a person with good local knowledge). Therefore, the researchers and annotators signed a non-disclosure agreement. To further protect our participants, we cannot make the data public. On the other hand, experiments with this kind of data are beneficial and unique, as they present language as it naturally occurs between participants and is mostly not redacted (for example, as compared to public discussion forums). Our results can be compared with current models that are usually trained on publically available data that do not correspond to private conversations.

Table 2: The Number of Users in Annotated Conversations And Statistics About Utterances. Authors Have at Least One Line Annotated by the Annotator; Participating Users Are Those Who Were Involved in the Conversations Where the Online Risk Appeared

ONLINE RISK	AUTHORS	PARTICIPANTS
(1) Aggression	403	860
(2) Mental health	89	272
(3) Alcohol, drugs	301	728
(4) Sexual content	144	420

Each separate file consisting of the dialogue between the participant and another person (or persons in case of a group chat) in the Czech language was then divided into smaller parts called conversations. A conversation ends when a person does not write any utterance for at least 60 minutes. The total number of such conversations is



a) Aggression, harassment, hate

b) Mental health problems

c) Alcohol, drugs

d) Sexual content

Figure 1: The Number of Positively Annotated Utterances Authored by One Person for Each Online Risk. One Rectangle Represents a Person, and the Relative Size of the Rectangle and the Number Within Each Rectangle Represents the Number of Utterances

90,422. The total number of uploaded textual utterances was 1,260,492. They were authored by 2165 different people from 2015-09-14 to 2020-12-14.

All text messages were anonymized and access-protected on multiple levels. Prior to uploading to a private server through a custom desktop application, the data were anonymized with the tool described in (Sotolář, Plhák, and Šmahel, 2021). Moreover, multimedia messages were replaced by appropriate tokens (like <photo>, <gif>, <audio_file>, <sticker>) to preserve anonymity and also the flow of the dialogue (e.g., to know that the communication partner reacts to an uploaded photo). The segmentation also strengthens the anonymity of the data (when presented to annotators randomly) because it broke familiarity with both the dialogue authors’ style and discussed topics. The annotators provided the last level of anonymization as they were members who passed intensive training in data confidentiality. If they found or suspected a rare case of re-identification or attribute disclosure, they reported the case, which was mitigated by manual annotation.

3.1.1 Annotation

A custom web annotation tool was developed to facilitate the annotation process. It allows annotators to tag each line of conversation with one appropriate category of the online risks. Moreover, one additional tag could be added: a question mark (the annotator is unsure about the category) or +T (the annotator assumes that another online risk category can be used). Annotators were also allowed to load previous and subsequent conversations to assess context.

Based on current research on the risky online behavior of adolescents, we developed an annotation manual for all four online risks. In the first phase, we trained two annotators for two months and incrementally refined the manual according to the specific style of IM conversations. Consequently, the annotators started to code randomly selected parts of the corpora. Criteria were the strict interpretation of annotation guidelines, excluding everything that is only implied but not clearly identifiable. There should be an explicit mention of the given phenomenon.

However, the risks in our corpora have sparse occurrence; therefore, we decided to prepare a preliminary classifier to identify conversations with a higher chance of containing utterances with a given online risk. Altogether, the total number of 35,000 conversations with 272,465 utterances were processed, and the number of utterances annotated by at least one annotator (even if they marked it with a question mark as an additional tag) was in total 14,345 (5.26%). The exact numbers for each online risk can be seen in the first column of Table 1.

3.1.2 Gold Standard Generation

The corpus was reviewed by a supervisor (social science researcher focusing on researching adolescents' well-being and technology use) in order to create the gold standard dataset. Utterances annotated by both annotators (without question mark as an additional tag) were included in the gold standard dataset without further processing.

The supervisor made final decisions when the utterance was annotated:

- by exactly one annotator (with or without a question mark as an additional tag),
- by both annotators with at least one question mark as an additional tag,
- with +T as an additional tag.

3.1.3 Data Variability

The analysis of the datasets revealed that individual authors had contributed different amounts of text to the gold standard. This data variability over the users is shown in Figure 1.

Both mental health problems and sexual content have three dominant users that produced over half of the annotated utterances. On the other hand, (1) aggression, harassment, hate, and (3) alcohol, drugs have a more uniform distribution of utterances among authors. The exact number of unique users involved in conversations is shown in Table 2.

We hypothesized that this discrepancy negatively influences the representativeness of the samples, which will cause the models not to generalize well. For this reason, we have excluded categories (2) mental health problems and (4) sexual content from experimentation. Nevertheless, we show the classification results for the sake of completeness.

3.2 Models

For the classification, we used models based on the Transformer architecture (Vaswani et al., 2017) that were pretrained on various large language corpora. We added a final softmax classification layer and fine-tuned all models on our datasets. The candidate model selection was based on literature (Sotolář, Plhák, and Šmahel, 2021; Straka et al., 2021) by picking the models that were top-performers on tasks related to ours. We included both mono and multilingual models of various sizes in the comparison.

4 EXPERIMENTS

4.1 Dataset Generation

The datasets consist of examples created by concatenating adjacent utterances in the conversation. The example's label is determined by the label of the target utterance, with the previous text being considered as context. Previous work indicates that such contexts, no matter the speaker, significantly improves the classification performance (Lugini and Litman, 2021). Its authors have also unsuccessfully tried to apply the attention mechanism to learn the optimal context length and position, resulting in the suggestion to treat it as a hyperparameter. The results of our experiments with the length of the context are presented in Section 4.2.

To boost the statistical significance, we have used 5-fold cross-validation. We have additionally split the test parts in a 1:3 ratio to use the smaller one as a development partition for model selection and the remainder for measurements only.

Table 3: Mean, Median, and Standard Deviation of Characters Count in Utterances That Were in the Gold Standard Dataset and Class Ratio (Number of Positive Utterances to Negative Utterances)

RISK	MEAN	MEDIAN	STDEV	RATIO
(1)	44.50	23	136.08	1:85
(2)	102.77	68	114.69	1:121
(3)	40.21	28	59.14	1:135
(4)	35.53	25	62.01	1:86

4.1.1 Category Distribution and Imbalance

The analysis of the annotations showed that the overlap of categories (i.e., particular online risks) was negligible. Only 96 utterances (out of 10,422) have been positively annotated by more than one online risk category. Therefore, we decided to create separate datasets for each category for binary classification, with which it is easier to experiment. In each dataset, the distribution is highly biased towards the negative class (the ratio is shown in Table 3). We experimented with two approaches to address the imbalance: weighting the loss function and augmenting the training data by adding paraphrases of the minority class examples. The augmentation method is described in Section 4.4 and the findings are summarized in Table 4.

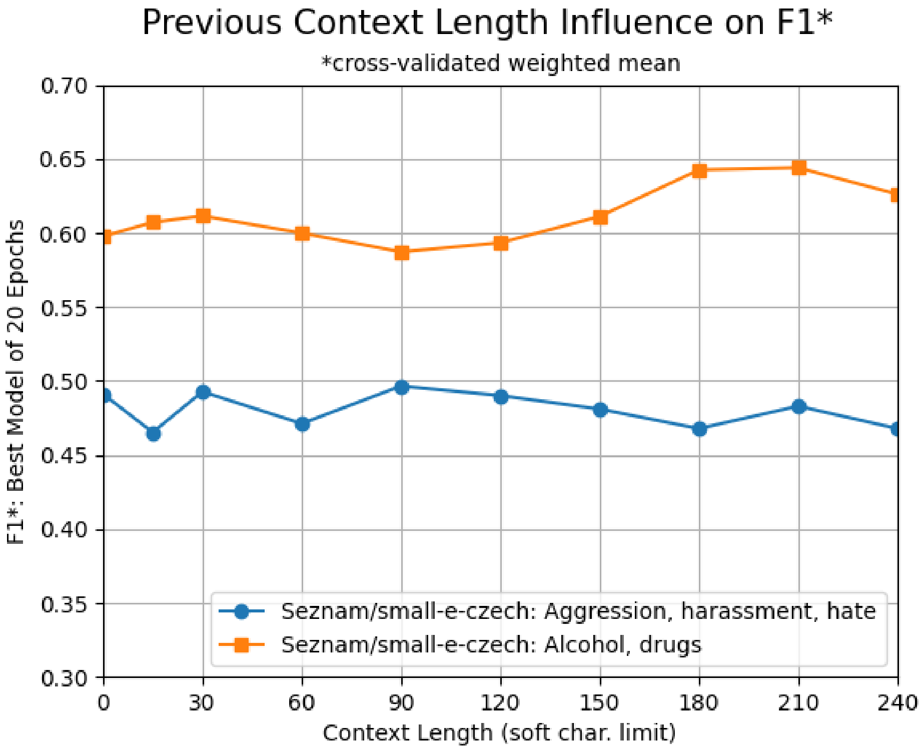


Figure 2: Experiments With Context Length

4.2 Including Context

We experimented with the length of the utterance's previous context to determine how different settings affect the classification. The length has a soft limit measured in characters. Since the natural conversation unit is an utterance, we opted not to split them. If the character limit occurred within an utterance, it was prepended as a whole. Therefore, context length could exceed the limit. We set the increment length to 30 characters based on the average and median length of utterances (24.74 and 15 characters).

The results are shown in Figure 2. Our findings show a significant difference between the categories: for category (1) aggression, harassment, hate, the addition of context has improved the result only slightly, while for category (3) alcohol and drugs, the improvement is significant. We set the best context length of the category (1) to 90 and 210 for category (3) for further experiments.

4.3 Model Selection

We compared the performance of three Transformer models:

- the small-size monolingual Small-E-Czech,
- the full-size monolingual Robeczech-base,
- the full-size XLM-Roberta-base model.

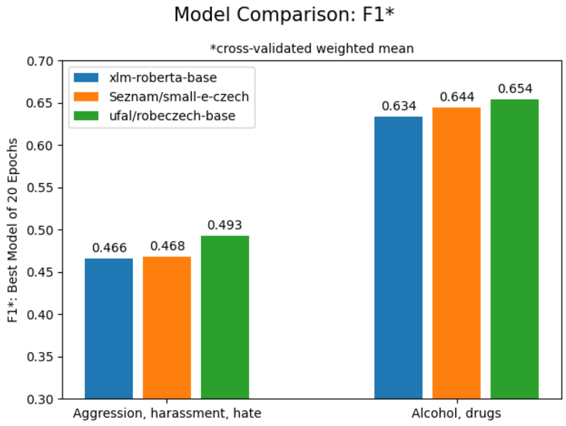


Figure 3: Model Comparison for Categories: (1) Aggression, Harassment, Hate, and (3) Alcohol, Drugs

The results shown in Figure 3 indicate that the monolingual model outperforms the multilingual one of comparable size. The larger Robeczech model outperforms the smaller Small-E-Czech, which is consistent with the literature (Devlin et al., 2018). Interestingly, Small-E-Czech performs with the same results as the approximately 2.5 times larger XLM-Roberta, making it a favorite for experimentation due to the shorter compute time and smaller memory footprint.

Table 4: Final Results (_*for Mental health problems and Sexual content, classification results were achieved on non-representative data, and therefore, experiments with augmented data were not conducted_)

ONLINE RISK	MODEL	CONTEXT LENGTH	F1 WITHOUT / F1 WITH AUGMENTATION
(1) Aggression, harassment, hate	xlm-roberta-base	90	.466 / .465
	small-e-czech	90	.468 / .455
	robeczech-base	90	**.493** / .490
(2) Mental health problems*	_small-e-czech_	_90_	_.66 / —_
(3) Alcohol, drugs	xlm-roberta-base	210	.634 / .629
	small-e-czech	210	.644 / .647
	robeczech-base	210	**.654** / .652
(4) Sexual content*	_small-e-czech_	_90_	_.828 / —_

4.4 Data Augmentation

We augmented the training set by adding paraphrases of positive examples. This can improve the model's robustness, or even absolute performance (Fadaee, Bisazza, and Monz, 2017). The paraphrases are generated by a back translation, which was shown to outperform (Xu et al., 2020) word-level methods such as the EDA (Wei and Zou, 2019).

In our case, positive examples from the training set were augmented three times using the OPUS-MT model (Tiedemann and Thottingal, 2020). Utterances were translated:

- to English, then to German, and back to the Czech language,
- to English and back to the Czech language,
- to German and back to the Czech language.

For both online risk categories we experimented with, we compared the F1 score achieved with and without augmentation with the same model settings. The results, see Table 4, show that augmentation did not have a substantial impact on the F1 score.

4.5 Error Analysis

We performed the analysis of predictions with Layer Integrated Gradients (Sundararajan, Taly, and Yan, 2017). We considered analyzing the attention directly, but the sequences proved to be too long, which caused the attention weights to be too scattered.

The analysis led to interesting results: firstly, it showed that the linguistic features that define (1) aggression, harassment, hate, and (3) alcohol, drugs are different. The wrongly predicted examples in (1) tend to be sarcastic or a part of inter-group language style, which might use aggressive vocabulary when, in fact, it is not. On the other hand, (2) seems to be well determined by keywords and phrases. The erroneous predictions tend to be caused by attending tokens positioned within the classified utterances' context (shown in Figure 5), which the model is not supposed to use for classification directly due to the way the training examples are constructed.

4.6 Results

The results of our experiments are summarized in Table 4. The best F1 score varies between .493 and .828 for examined online risks. Confusion matrices are shown in Figure 4.

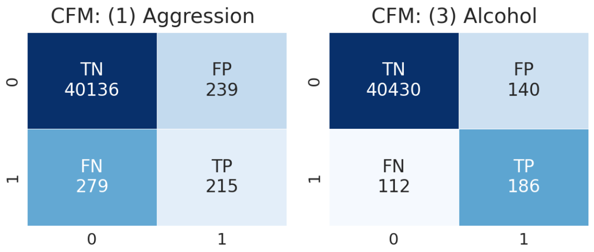
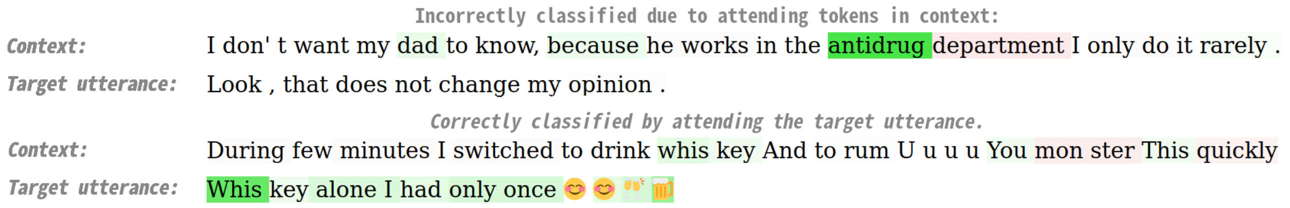


Figure 4: Confusion Matrices for Categories: (1) Aggression, Harassment, Hate, and (3) Alcohol, Drugs

All examined online risks except category (3) alcohol, drugs are hard to classify even by human annotators, as shown with the inter-annotator agreement measured with Cohen's κ in Table 1. The classification results are consistent with the achieved κ for the given category: *moderate* for (1) aggression, harassment, hate, and *good* for (3) alcohol, drugs. Through error analysis and discussions with the annotators, we have concluded that the data in category (1) comprises dialogues on a large variety of topics, while the topics are more similar for category (3). This effect is amplified by the category (3) definition in the annotation manual, which is more focused than (1).

Our experiments with data augmentation using back-translation have only resulted in minimal improvements, as shown in Table 4.



Incorrectly classified due to attending tokens in context:

Context: I don' t want my dad to know, because he works in the antidrug department I only do it rarely .

Target utterance: Look , that does not change my opinion .

Correctly classified by attending the target utterance.

Context: During few minutes I switched to drink whis key And to rum U u u u You mon ster This quickly

Target utterance: Whis key alone I had only once 😊 😊 🍻 🍺

Figure 5: Layer Integrated Gradients Attributions of the Classification Output With Respect to Input Features. Highlighted Text Contributes to the Prediction Proportionally to Its Shade, Green to the Positive and Red to the Negative Classes

The practical usability of our models varies by category. We presume the best model for category (3) alcohol, drugs could improve results with additional thresholding. Its negative predictive value is greater than 0.997, which would result in a low number of false alarms, which is crucial given the data distribution. To a lesser extent, this also applies to the model for category (1) aggression, harassment, hate with a negative predictive value greater than 0.993.

5 LIMITATIONS

The F1 score in category (4) sexual content was .82, which is very high considering the inter-annotator agreement κ of 0.485. However, this category and category (2) mental health problems are heavily affected by the combination of high bias and low data variance because most of the data were produced by only a few users (see Figure 1), and the model can learn to adapt to their language style.

Furthermore, detecting adolescents' online risky behavior is a particular problem, and therefore, it is hard to use annotated corpora in other tasks. Moreover, our work is focused on conversations in the Czech language with a small number of native speakers (10.7 million).

The main limitation is reproducibility because the corpora of private instant messaging conversations cannot be made public due to the protection of the privacy of our participants and everyone who took part in the chats. On the other hand, experiments with this kind of data are beneficial and unique, as they present language as it naturally occurs between participants and is mostly not redacted (for example, as compared to public discussion forums). Moreover, provided models as well as the training script are published (Sotolář and Plhák, 2023).

6 CONCLUSIONS

This work provided innovative classification models for adolescents' risky online behavior trained on unique datasets of real instant messaging conversations. We build new models for categories of four risk categories: (1) aggression, harassment, hate; (2) mental health problems; (3) alcohol, drugs; and (4) sexual content. We have produced annotation manuals and software tools and created new annotated corpora.

We have analyzed data within the corpora and experimented with different models, context settings, and class-imbalance solutions to produce predictive models. Using the settings optima, we have trained models for two categories, achieving an acceptable F1 score of .493 for category (1) aggression, harassment, hate, and solid .654 for category (3) aggression, harassment, and hate. Our results indicate substantial differences in the classification of different online risks. There are also differences in the prevalence of online risks within the corpora, impacting the classification. Two categories were authored by only a few adolescents (mental health problems and sexual content), resulting in datasets that are not representative samples. Future research would require more diverse corpora for training more reliable classification models for these categories.

Our classification models are practically usable in many applications like parental control applications, chat-bots, real-time services provided by social networking sites, or any other services working with the dialogues of our target age group. Furthermore, our annotation manual, tools, and results of experiments with data generation, model selection, and other methods are beneficial for solving related tasks with conversations in other languages.

7 FUTURE WORK

For future work, we suggest experimenting with other machine translation models to improve the paraphrase generation because multiple authors have shown that data augmentation can help solve the class imbalance. Experiments can be conducted using the state-of-the-art model mT5 (Xue et al., 2020) or use different approaches to generate paraphrases for solving class imbalance such as Wei and Zou (2019) suggest.

Other experiments with context can involve changing character-delimited context length, using both previous and subsequent context, using only the utterances of a specific user, etc.

A separate hyperparameter search for the individual models, possibly automated, might also help improve the models' results.

Future research should also consider the potential effects of dividing the categories into more specific subcategories, e.g., category (1) aggression, harassment, hate into three separate ones. This approach could lead to a better definition of problems, higher inter-annotator agreement, and possibly more efficient classification even with a limited number of positive utterances.

Acknowledgements

This work has received funding from the Czech Science Foundation (project number 19-27828X).

We would like to extend our sincere thanks to Michał Tkaczyk for the professional help with specifying label definitions. We also wish to thank the annotators for their excellent work.

References

Acheampong, Francisca Adoma, Chen Wenyu, and Henry Nunoo-Mensah (2020). “Text-based emotion detection: Advances, challenges, and opportunities”. In: *Engineering Reports* 2.7, e12189.

Afshar, Majid et al. (2019). “Natural language processing and machine learning to identify alcohol misuse from the electronic health record in trauma patients: development and internal validation”. In: *Journal of the American Medical Informatics Association* 26.3, pp. 254–261. ISSN: 1527-974X. DOI: 10.1093/jamia/ocy166.

Afzali, Mohammad H. et al. (2019). “Machine-learning prediction of adolescent alcohol use: a cross-study, cross-cultural validation”. In: *Addiction* 114.4, pp. 662–671. DOI: https://doi.org/10.1111/add.14504.

Álvarez-García, David et al. (2018). “Individual, Family, and Community Predictors of Cyber-aggression among Adolescents”. In: *The European Journal of Psychology Applied to Legal Context* 2018avonline. DOI: 10.5093/ejpalc2018a8.

Alzoubi, Hadeel et al. (2018). “An Automated System for Identifying Alcohol Use Status from Clinical Text”. In: *2018 International Conference on Computing, Electronics Communications Engineering (iCCECE)*, pp. 41–46. DOI: 10.1109/iCCECOME.2018.8658578.

Aphinyanaphongs, Yin et al. (2014). “Text classification for automatic detection of alcohol use-related tweets: A feasibility study”. In: *Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014)*, pp. 93–97. DOI: 10.1109/IRI.2014.7051877.

Benotsch, Eric et al. (2012). “Sexting, Substance Use, and Sexual Risk Behavior in Young Adults”. In: *The Journal of adolescent health : official publication of the Society for Adolescent Medicine* 52. DOI: 10.1016/j.jadohealth.2012.06.011.

Chatterjee, Ankush et al. (2019). “SemEval-2019 task 3: EmoContext contextual emotion detection in text”. In: *Proceedings of the 13th international workshop on semantic evaluation*, pp. 39–48.

Chen, Jindong et al. (2019). “Deep short text classification with knowledge powered attention”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33, pp. 6252–6259.

Dadvar, Maral, Rudolf Berend Trieschnigg, and Franciska MG de Jong (2013). “Expert knowledge for automatic detection of bullies in social networks”. In: *25th Benelux Conference on Artificial Intelligence, BNAIC 2013*. Delft University of Technology, pp. 57–64.

Dahiya, Sonika, Astha Mohta, and Atishay Jain (2020). “Text classification based behavioural analysis of whatsapp chats”. In: *2020 5th international conference on communication and electronics systems (ICCES)*. IEEE, pp. 717–724.

De Choudhury, Munmun (2013). “Role of Social Media in Tackling Challenges in Mental Health”. In: *Proceedings of the 2nd International Workshop on Socially-Aware Multimedia*. SAM ’13. Barcelona, Spain: Association for Computing Machinery, pp. 49–52. ISBN: 9781450323949. DOI: 10.1145/2509916.2509921.

Devlin, Jacob et al. (2018). “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805*.

Fadaee, Marzieh, Arianna Bisazza, and Christof Monz (2017). “Data augmentation for low-resource neural machine translation”. In: *arXiv preprint arXiv:1705.00440*.

Flek, Lucie (2020). “Returning the N to NLP: Towards contextually personalized classification models”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7828–7838.

Al-garadi, Mohammed Ali, Kasturi Dewi Varathan, and Sri Devi Ravana (2016). “Cybercrime Detection in Online Communications”. In: *Comput. Hum. Behav.* 63.C, pp. 433–443. ISSN: 0747-5632. DOI: 10.1016/j.chb.2016.05.051.

Huang, Hanyun and Louis Leung (2009). “Instant messaging addiction among teenagers in China: Shyness, alienation, and academic performance decrement”. In: *CyberPsychology & Behavior* 12.6, pp. 675–679.

Khanpour, Hamed, Nishitha Guntakandla, and Rodney Nielsen (2016). “Dialogue act classification in domain-independent conversations using a deep recurrent neural network”. In: *Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers*, pp. 2012–2021.

Koutamanis, Maria et al. (2013). “Practice makes perfect: The longitudinal effect of adolescents’ instant messaging on their ability to initiate offline friendships”. In: *Com-*

puters in Human Behavior 29.6, pp. 2265–2272. ISSN: 0747-5632. DOI: https://doi.org/10.1016/j.chb.2013.04.033.

Lee, Ji Young and Franck Dernoncourt (2016). "Sequential short-text classification with recurrent and convolutional neural networks". In: *arXiv preprint arXiv:1603.03827*.

Li, Jiawei et al. (2019). "A Machine Learning Approach for the Detection and Characterization of Illicit Drug Dealers on Instagram: Model Evaluation Study". In: *J Med Internet Res* 21.6, e13803. ISSN: 1438-8871. DOI: 10.2196/13803.

Lin, Huijie et al. (2017). "Detecting Stress Based on Social Interactions in Social Networks". In: *IEEE Transactions on Knowledge and Data Engineering* 29, pp. 1820–1833.

Lugini, Luca and Diane Litman (2021). "Contextual argument component classification for class discussions". In: *arXiv e-prints*, arXiv–2102.

Martínek, Jiří, Christophe Cerisara, et al. (2021). "Cross-Lingual Approaches for Task-Specific Dialogue Act Recognition". In: *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer, pp. 232–242.

Martínek, Jiří, Pavel Král, et al. (2019). "Multi-lingual dialogue act recognition with deep learning methods". In: *arXiv preprint arXiv:1904.05606*.

O'Dea, Bridianne et al. (2015). "Detecting suicidality on Twitter". In: *Internet Interventions* 2.2, pp. 183–188. ISSN: 2214-7829. DOI: https://doi.org/10.1016/j.invent.2015.03.005.

Pant, Kartikey et al. (2019). "SmokEng: Towards Fine-grained Classification of Tobacco-related Social Media Text". In: *CoRR* abs/1910.05598.

Rahman, Rohizah Abd et al. (2020). "Application of Machine Learning Methods in Mental Health Detection: A Systematic Review". In: *IEEE Access* 8, pp. 183952–183964. DOI: 10.1109/ACCESS.2020.3029154.

Razi, Afsaneh (2020). "Deploying Human-Centered Machine Learning to Improve Adolescent Online Sexual Risk Detection Algorithms". In: *Companion of the 2020 ACM International Conference on Supporting Group Work*. GROUP '20. Sanibel Island, Florida, USA: Association for Computing Machinery, pp. 157–161. ISBN: 9781450367677. DOI: 10.1145/3323994.3372138.

Rosa, Hugo, David Matos, et al. (2018). "A "Deeper" Look at Detecting Cyberbullying in Social Networks". In: *2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. DOI: 10.1109/IJCNN.2018.8489211.

Rosa, Hugo, N. Pereira, et al. (2019). "Automatic cyberbullying detection: A systematic review". In: *Computers in Human Behavior* 93, pp. 333–345. ISSN: 0747-5632. DOI: https://doi.org/10.1016/j.chb.2018.12.021.

Shatte, Adrian B. R., Delyse M. Hutchinson, and Samantha J. Teague (2019). "Machine learning in mental health: a scoping review of methods and applications". In: *Psychological Medicine* 49.9, pp. 1426–1448. DOI: 10.1017/S0033291719000151.

Smahel, David et al. (2020). *EU Kids Online 2020: Survey results from 19 countries*. Tech. rep. EU Kids Online network. DOI: 10.21953/lse.47fdeqj01ofo.

Song, Juyoung, Tae Min Song, and Jin Ree Lee (2018). "Stay alert: Forecasting the risks of sexting in Korea using social big data". In: *Computers in Human Behavior* 81, pp. 294–302. ISSN: 0747-5632. DOI: https://doi.org/10.1016/j.chb.2017.12.035.

Sotolář, Ondřej and Jaromír Plhák (2023). *Huggingface Hub Repository with Models and Training Script*. URL: https://huggingface.co/xsotolar.

Sotolář, Ondřej, Jaromír Plhák, and David Šmahel (2021). "Towards Personal Data Anonymization for Social Messaging". In: *International Conference on Text, Speech, and Dialogue*. Springer, pp. 281–292.

Sotolář, Ondřej, Jaromír Plhák, Michal Tkaczyk, et al. (2021). "Detecting Online Risks and Supportive Interaction in Instant Messenger Conversations using Czech Transformers". In: *RASLAN 2021 Recent Advances in Slavonic Natural Language Processing*, p. 19.

Straka, Milan et al. (2021). "RobeCzech: Czech RoBERTa, a monolingual contextualized language representation model". In: *International Conference on Text, Speech, and Dialogue*. Springer, pp. 197–209.

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan (2017). "Axiomatic Attribution for Deep Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 3319–3328.

Tiedemann, Jörg and Santhosh Thottingal (2020). "OPUS-MT – Building open translation services for the World". In: *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. Lisboa, Portugal: European Association for Machine Translation, pp. 479–480.

Underwood, J Michael et al. (2020). "Overview and methods for the youth risk behavior surveillance system—United States, 2019". In: *MMWR supplements* 69.1, p. 1.

Valkenburg, Patti and Jochen Peter (2011). "Online Communication Among Adolescents: An Integrated Model of Its Attraction, Opportunities, and Risks". In: *The Journal of adolescent health : official publication of the Society for Adolescent Medicine* 48, pp. 121–7. DOI: 10.1016/j.jadohealth.2010.08.020.

Vaswani, Ashish et al. (2017). "Attention is all you need". In: *Advances in neural information processing systems* 30.

Vogt, Matthias, Ulf Leser, and Alan Akbik (2021). "Early Detection of Sexual Predators in Chats". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4985–4999.

Wahyono, Irawan Dwi et al. (2021). "Text Mining in Chat Room of Online Learning for Detection Emotion using Artificial Intelligence". In: *2021 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE)*. IEEE, pp. 63–67.

Wei, Jason and Kai Zou (2019). "Eda: Easy data augmentation techniques for boosting performance on text classification tasks". In: *arXiv pre-print arXiv:1901.11196*.

Wongkoblap, Akkapon, Miguel A. Vadillo, and Vasa Curcin (2018). "A multilevel predictive model for detecting social network users with depression". English. In: *Proceedings - 2018 IEEE International Conference on Healthcare Informatics, ICHI 2018*. 6th IEEE International Conference on Healthcare Informatics, ICHI 2018 ; Conference date: 04-06-2018 Through 07-06-2018. United States: Institute of Electrical and Electronics Engineers Inc., pp. 130–135. DOI: 10.1109/ICHI.2018.00022.

Xu, Binxia et al. (2020). "Data Augmentation for Multiclass Utterance Classification–A Systematic Study". In: *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 5494–5506.

Xue, Linting et al. (2020). *mT5: A massively multilingual pre-trained text-to-text transformer*. DOI: 10.48550/ARXIV.2010.11934.

Zhao, Rui, Anna Zhou, and Kezhi Mao (2016). "Automatic Detection of Cyberbullying on Social Networks Based on Bullying Features". In: *Proceedings of the 17th International Conference on Distributed Computing and Networking*. ICDCN '16. Singapore, Singapore: Association for Computing Machinery. ISBN: 9781450340328. DOI: 10.1145/2833312.2849567.

A ANNOTATION MANUAL: ONLINE RISKS

A.1 General Guidelines

- When the conversation does not include sufficient context for the online risk, and we are not sure whether to annotate, we do not annotate.

- We also annotate negative responses if they fit within the guidelines (e.g., “wanna go for a smoke?” “nope, i don't smoke”)

A.2 Responses and Long Conversations

We code responses when they are related to the person and explicitly tied to the risk.

Examples:

- “I am so anxious that I cannot attend school. I need to see my doctor soon.” [MENTAL HEALTH]

- “I am sure it will be better, your doctor can prescribe you stronger meds and soon the anxiety will be gone” [MENTAL HEALTH] vs. “Every problem has a solution, don't worry.” [NO TAG]

A.3 Aggression, Violence, Harassment, Hate Speech and Conversations with Elements of Aggression

Does the utterance include:

- exposure to vulgar / aggressive content,[1]

- aggressive contents / insults / threats / defamation,[2]

- referring to or incitement to violence / aggressive behavior (cyberaggression, harassment, violence).

- hate speech, xenophobia, racism, discrimination against a nation / ethnicity / color of skin / religion / sexuality / weight

In addition:

- We do not annotate vulgarisms that are not directed (e.g., just saying vulgarisms).

- We annotate vulgarisms even if they are in a friendly context (e.g., calling girlfriends bitches).

- We also include aggression towards groups and public figures (politicians, celebrities).

- Implicit racism is also coded, as well as aggression related to dehumanization and slander.

Examples:

- **violent/aggressive behavior:** “. . . so when I came back from work she was already there. . . and then I hit her and we fought . . . ”

- **racism:** “Yeah, that neighborhood is so full of gypsies that it's dark even during the day there.“

[1] Only when directed towards concrete people or groups.

[2] See footnote 1.

A.4 Mental Health Problems and Self-Harm

Does the utterance include:

- referring to / complaining about / experiencing long-term mental health problems like depression, anxiety, phobias, paranoia, insomnia, eating disorders (anorexia, bulimia, binge eating), self-harming, suicidal ideation,
- referring to therapy, medication, a psychiatric hospital,
- describing experiences with eating disorders, instructions for not eating, and drastic weight loss,
- sending pro-ana contents.

In addition:

- We do not annotate short-term feelings or affects (bad mood, sadness). It must be evident from the conversations that the symptoms are long-term.
- Our goal is to detect the discussion of mental health problems, not diagnose and/or assume the problems of our participants - therefore, we code explicit mentions of mental health problems and not the general discussions of bad moods, etc.

Examples:

- **self-harm/suicidal ideation:** "Ana… She was really mean. I had a panic-self-hate attack with like… umm… not positive consequences… and I wanted to overdose."
- **therapy:** "My psychiatrist is so stupid, I need to find a new one, he kicked me out and doesn't wanna prescribe me any meds!"

A.5 Alcohol and Drugs

Does the utterance include:

- referring to one's own or someone else's experience with alcohol or drugs (cigarettes, nicotine, tobacco, hookah pipes, marijuana, abusing medications …),
- making plans to drink alcohol or take drugs,
- seeking drugs,
- supporting/justifying alcohol and drugs,
- talking about the intention to try/use alcohol or drugs.

Examples:

- **talking about the experience with drugs:** "Man, I'm so effin stoned :-D!"
- **making plans to drink alcohol:** "Are 3 bottles of wine enough? :"-("

A.6 Sexual Content

Does the utterance include:

- links to / talking about pornography,
- discussing sexual experiences and one's sex life,

- flirting with sexually explicit content,
- sexting,
- sexual innuendo (even if meant as a joke between friends),
- soliciting nudes or sexual information.

Examples:

- **discussing one's sex life:** "You did not answer me last time... what's it like during sex from a boy's point of view? :P :D"
- **sexual innuendo (even as a joke):** "oh so you're at home? man, you can jerk off all day haha!"

B USING THE MODELS

```python
from transformers import AutoTokenizer, RobertaForSequenceClassification

TEXT = "Budou stacit 3 lahve vina?"
# ^ "Are 3 bottles of wine enough?"
MAX_LEN = 210

# initialize tokenizer
# (avoid buggy Rust implementation)
tokenizer = AutoTokenizer.from_pretrained("xsotolar/split-1_17_210_id-078-tokenizer", use_fast=False, truncation_side="left")

# HF version check
assert tokenizer.truncation_side == "left"

# initialize model
model = RobertaForSequenceClassification.from_pretrained("xsotolar/split-1_17_210_id-078")

# tokenize a given text
# cut off at MAX_LEN characters
inputs = tokenizer(TEXT, padding=True, truncation=True, max_length=MAX_LEN, return_tensors="pt").to("cuda")

# classify & return probabilities of
# [negative, positive] class
outputs = model(**inputs)
result = outputs[0].softmax(1)
print(result)
```

C HYPERPARAMETER SETTINGS

```python
# general hyperparameters
ARGS = {
    "model_name": "ufal/robeczech-base",
    "epochs": 10,
    "target_names": ["0", "1"],
    "per_device_batch": 128,
    "learning_rate": 1e-5
}

# linear warmup to 1/3 of epoch
w_steps = int((ARGS["epochs"] * len(train_texts)) / (3 * ARGS["per_device_batch"] * ARGS["visible_devices"]))

# use F1 for choosing best model
training_args = TrainingArguments(
    ...
    load_best_model_at_end=True,
    metric_for_best_model="f1",
    ...
)

# use CE loss weighted by the training example ratio (positive/negative)
class CustomTrainer(Trainer):
    def compute_loss(self, model, inputs, return_outputs=False):
        labels = inputs.get("labels")
        outputs = model(**inputs)
        logits = outputs.get("logits")
        loss_fct = nn.CrossEntropyLoss(weight=torch.tensor(ARGS["class_ratio"]).to("cuda"))
        loss = loss_fct(logits.view(-1, self.model.config.num_labels), labels.view(-1))
        if return_outputs:
            return (loss, outputs)
        return loss
```