Detecting Risky Behavior Related to Alcohol and Drug Use Within Adolescents' Private Messenger Conversations

Jaromír Plhák, Michaela Lebedíková, Ondřej Sotolář and David Šmahel
Faculty of Informatics, Masaryk University
Brno, Czech Republic
{xplhak, x450458, xsotolar, davs}@fi.muni.cz

Abstract

Alcohol and drug use negatively impact adolescents' health, making early detection and prevention essential. One promising approach involves analyzing adolescents' online conversations for signs of substance use. However, current machine learning models for online detection often rely on public data sources that fail to capture the private experiences of adolescents. In this study, we developed a BERT-based machine learning model to automatically identify discussions about alcohol and drug use with high accuracy, leveraging private messenger conversations from adolescents. Our novel dataset comprises 272,465 annotated utterances from a corpus of 1,260,492 utterances in 2,807 chats authored by 2,165 individuals, primarily in Czech. Our best BERT-based machine learning model achieved a solid F_1 score of 0.817, demonstrating the feasibility of addressing this social science task even in low-resource languages like Czech. Additionally, we verified that state-of-the-art generative open-source large language models are equally effective for this task and can be successfully adapted for other languages, including English. We also analyzed misclassified utterances to identify problematic patterns and improve model performance. The resulting models have significant practical implications for parental mediation software and parental control applications. By automating substance use detection and enabling appropriate real-time interventions, these tools can contribute to safeguarding adolescents' health.

Keywords: substance use detection, online conversations, machine learning model, error analysis

1. Introduction

While global numbers of substance use (meaning "legal drugs such as alcohol, tobacco, as well as illegal drugs," Rane et al. (2022)) are declining, the numbers stay high among adolescents (Charrier et al., 2021). Among European and Canadian 15-year-olds, 37% admitted to drinking alcohol, 15% smoked cigarettes, and 7% used cannabis in the past 30 days. Adolescent substance use is driven by their peers: for example, in Czechia, where drinking beer daily is socially acceptable, 63.6% of adolescents are mild users (Bräker and Soellner, 2016). While not every instance of substance use results in substance use disorder, which is characterized by "substance-induced disorders related to anxiety, mood, sleep, and sexual functioning" (Rane et al., 2022), it is important to examine adolescents' drinking habits. This examination helps better understand how such behavior unfolds in social contexts, particularly in environments without parental oversight, such as instant messaging conversations with peers.

A beneficial approach for investigating substance use might be machine learning, which has been used to diagnose disease, predict outcomes of treatments, and identify risk factors (Rane et al., 2022). However, most studies aim to diagnose adolescent disordered substance use from medical records (e.g., (Afzali et al., 2019; Elena et al., 2009; Whelan et al., 2014)) or to gain an insight into the general public's opinion on specific substances on social media that are publicly accessible, such as X (formerly Twitter) or Reddit, e.g., (Allem et al., 2018; Glowacki et al., 2018; Meacham et al., 2018; Zhang et al., 2025). Therefore, we still lack an understanding of adolescents' substance use in an ecologically valid environment that would represent the specific language of adolescents (Alsoubai, 2023).

We ask the following Research Questions (RQ):

- **RQ1:** Are we able to effectively detect adolescents' discussions about substance use?
- **RQ2:** Are open-source generative Large Language Models (LLMs) suitable for substance use classification in private conversations, and do they outperform BERT-based models?
- And, **RQ3:** What are the most frequent causes of incorrect classification for open-source generative LLMs and BERT-based models?

Our work contributes to a novel application of machine learning in adolescent substance use. We use a dataset consisting of private messages of adolescents who downloaded content from their messengers prior to the study, which offers a valuable and unbiased insight into their (substance use) habits, which is a significant improvement in this area in terms of ecological validity (Ricard and Hassanpour, 2021). Thus, our study provides

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 2567–2580
11–16 May 2026.

a unique understanding of youth substance use habits. Moreover, the results can be used in applications for parental mediation or parental control software with real-time interventions. Finally, our results hopefully inspire new ways of machine learning applications and their usefulness in detecting substance use in natural conversations. Moreover, adolescents can benefit from the results, as risk detection might provide a basis for interventions, support, and treatment resources (Alsoubai, 2023).

2. Related Work

Previous research in the area of machine learning application in substance use is mainly focused on alcohol use by adults. For example, Shah-Mohammadi and Finkelstein (2024) use generative LLMs for classifying substance use information from a publicly available collection of medical records using GPT. Hurtado et al. (2022) identified 70 studies focused on alcohol-disordered use, utilizing functional magnetic resonance imaging, blood test measurements, electroencephalogram, and other types of medical data to predict hazardous alcohol use or identify risk factors for alcohol use disorder. Studies regarding adolescents were conducted similarly: by using medical and survey data, they aimed to identify risk factors and predict alcohol misuse in adolescents (Afzali et al., 2019; Vázquez et al., 2020; Whelan et al., 2014; Elena et al., 2009). However, not every substance use results in a disorder, which limits our understanding of adolescents' casual drinking.

Machine learning studies utilize natural language processing in social media. These papers offer a broader focus on substances: they include opioids (Glowacki et al., 2018; Mackey and Kalyanam, 2017; Mackey et al., 2018) and tobacco and marijuana use (Allem et al., 2018; Chen et al., 2015; Daniulaityte et al., 2016; Huang et al., 2017; Meacham et al., 2018; Zhan et al., 2017). In their review, (Conway et al., 2019) identified that most of the previous research utilizes X (formerly Twitter) as their data source, for example, (Allem et al., 2018; Daniulaityte et al., 2016; Aphinyanaphongs et al., 2014), with some papers also using Reddit (Meacham et al., 2018; Zhang et al., 2025) and online health communities (Zhan et al., 2017). Health-related text classification tasks with public social media data were also conducted using generative LLMs (Guo et al., 2024). These studies focus on the general public and often utilize topic modeling to gain insight into public opinion on a given substance. While important, such inquiries do not allow for an understanding of adolescents' substance use, nor do they offer a glimpse into how underage people talk about substance use in private, where adults cannot see it.

A significant shortcoming of previous studies is the reliance on automatic content classification, as human insight can improve machine-learning results (Gillies et al., 2016). One exception is the study of marijuana use by Daniulaityte et al. (2016). In their study, several coders developed a gold standard for machine learning classifiers. While the study shows strong performance of automatic machine learning classifiers, the authors conclude that for exploring new domains of substance use, manual coders, preferably experts in the area, are needed to understand and capture the specific language and terminology related to substance use. The same principle also applies to adolescents, known for often using their slang (Alsoubai, 2023), strengthening the need for human insight when researching adolescents. This principle was applied in a study focused on risky behavior, where adolescents' language was utilized by Plhák et al. (2023), who experimented with the multiple BERT-based models and provided promising baseline results.

3. Methodology

Participants for the study were recruited via a professional research agency and ads on social media. From all participants, 22 adolescents aged 13 to 17 (M_{age} = 15.86, 36% women) agreed to share their private conversations from the year preceding data collection. Thus, the data likely reflect adolescents' lived realities. Participants chatted with 2,165 unique people; however, we do not have any personal information about this broader sample. After providing written consent by participants and their caregivers, participants exported their private data from the Messenger communication tool developed by Meta Platform. They uploaded exported files to a private server using a customized tool that stored only text information without media content (e.g., images or videos). Files were subsequently anonymized, as described in Section 10.

3.1. Dataset

The final dataset consisted of 2,807 files containing the communication between a participant and one or more persons in the Czech language. These files were usually long and contained different topics (see Table 1). Therefore, we divided them into smaller parts called *conversations*. A conversation ends when a person does not send any message for at least 60 minutes. This approach has some drawbacks, e.g., when the same topic is discussed in the evening before sleep and continues in the morning, it occurs as two separate conversations. However, this was considered during the annotation phase, in which the annotators were allowed to load previous and following conversations. Moreover,

Data type	Mean	Median
File	462.5 utt.	21 utt.
Conversation	14.4 utt.	4 utt.
Utterance	30.0 chars	15 chars

Table 1: Statistics about dataset. (utt. means utterances)

we replaced the non-textual messages with appropriate tags to preserve the flow of the dialogue (e.g., to know that the reaction is to an uploaded sticker). The total number of conversations is 90,422, comprising 1,260,492 textual utterances authored by 2,165 different people from September 2014 to December 2020.

We filtered 20,857 one-utterance-long conversations without real user input. Of the remaining 69,565 conversations, 11,258 had only one author, 42,535 were one-user-to-one-user conversations, and 15,772 were group conversations.

3.2. Annotation

The annotation process started with the development of an annotation manual based on current research on the risky online behavior of adolescents by experienced social scientists. Then, one of the researchers trained two annotators for two months, and according to the results of the supervision, inter-annotator agreement, and discussions, the annotation manual was incrementally refined to reflect adolescents' language and the terminology regarding substance use.

Then, we settled on the following guidelines: Substance use is present when people in the chats refer to their own or someone else's experience with alcohol or drugs (such as cigarettes, nicotine packs, hookah, marijuana, abusing medication), make plans to drink alcohol or take drugs, seek drugs, support or justify alcohol and drug use, discuss intentions to try or use alcohol or drugs. The criteria for tagging alcohol and drug use in the given utterance were: a) the strict interpretation of annotation guidelines, excluding everything that is only implied but not clearly identifiable, and b) there should be an explicit mention of the given phenomenon.

We selected conversations from the corpora and uploaded them to an open-source web annotation tool that allowed annotators to tag each utterance of the conversation if an online risk was present. Moreover, an additional tag, a question mark, could be added if the annotator was unsure about the annotation. Annotators were also allowed to load previous and subsequent conversations to assess the context of given utterances, as discussed in Section 3.1.

Upon realizing that substance use has a sparse occurrence in the corpora, we prepared a preliminary classifier to identify conversations with a higher chance of containing utterances with online risk. As this method could lead to a biased sample of the data, we also included a large enough collection of random samples in the dataset for validation.

Annotators processed a total number of 35,000 conversations with 272,465 utterances. The number of positively annotated utterances tagged by at least one annotator was 2,301 (0.845%), which means very sparse occurrence. The inter-annotator agreement showed a substantial score of Cohen's κ equal to 0.609 (McHugh, 2012).

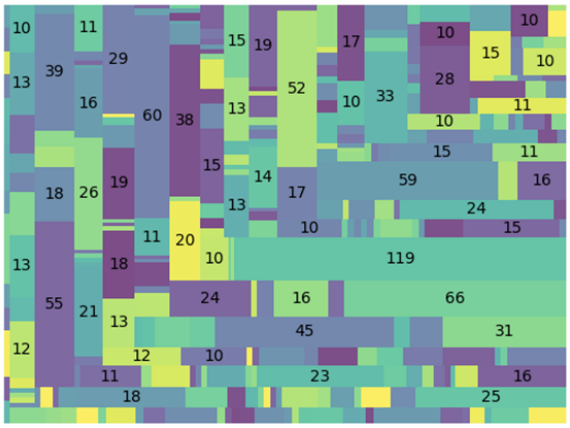

Figure 1: The number of utterances in the substance use gold standard divided by authors. One rectangle represents a person, and the relative size and number within each rectangle represent the number of utterances.

3.3. Gold Standard Specification

To specify the gold standard, the corpus was reviewed by a social science researcher focusing on risks and adolescents' well-being. The supervisor made final decisions when the utterance was annotated: a) by precisely one annotator (with or without a question mark as an additional tag); b) by both annotators with at least one question mark as an additional tag. Utterances annotated by both annotators (without a question mark as an additional tag) were automatically included in the gold standard dataset. Additionally, we provided the supervisor with the option to decide on 209 utterances, where the preliminary model was highly convinced that risky behavior was presented in utterances not included in the golden standard and vice versa.

The final gold standard contains 1990 utterances. The analysis of the datasets revealed that individual authors had contributed different amounts of text to the gold standard. However, none of the authors has an excessively large part of it, as shown in Figure 1. This distribution prevents overfitting the

vocabulary, style, and topics that this small group of authors discusses and incorrectly measures excellent performance on the in-domain testing sample.

3.4. Data Preparation

Initially, we had to decide how to prepare training data for the model. We could construct training inputs only using individual utterances as single inputs. However, this approach could be arguable for very short utterances due to the ambiguity of their labels and, more importantly, removes all context-dependent phenomena. On the contrary, we do not want to experiment with whole dialogues as examples, as we aim to detect local instances of substance use. Also, real-time applications usually cannot get the whole dialogue, only a few utterances at once. Therefore, we defined the inputs as *local context windows* within the dialogues. This approach allows the model to capture local dependencies (e.g., the humorous tone of the discussion).

A sliding window starts from the last utterance to construct the inputs to capture the local context. Specifically, we chose as many utterances as possible before the starting one until we surpassed the context length. As we work with only whole utterances, the inputs' lengths are usually longer than the context length and may vary. For example, for a context length of 256 characters, the mean length was 313 characters, the median was 303 characters, and the standard deviation was 118.

The final input label is produced by aggregating input's utterances' labels. If the sequence of labels contains any utterance that is positively labeled (as it contains substance use), the whole example is assigned the positive label.

For classification, we need to address the imbalance of class distribution. Plhák et al. (2023) used three approaches to address the imbalance: a) weighting the loss function, b) augmenting the training data by adding paraphrases of the minority class, and c) simple oversampling of the minority class examples.

Augmentation and oversampling showed slightly better results than the weighting. Therefore, the imbalance is solved in the training dataset by oversampling. However, the test dataset retains the original class distribution, which means the lower number of positive examples negatively affects the F_1 score during evaluation.

The dataset was partitioned into training, development, and test sets using an 80:5:15 ratio. To prevent data leakage and ensure unbiased evaluation, we enforced a strict user-level split, ensuring that data from any single user appears in only one of the subsets. The exact numbers of instances depend on the context length. For example, mainly used context length of 256 characters was the $n = 206,796$, $n_{train} = 165,436$, $n_{dev} = 10,339$, and $n_{test} = 31,021$.

3.4.1. Data Preprocessing

We use the following techniques to clean and prepare the raw data to reveal if they affect the quality of the model:

- **Removing 405 Czech stop words**.
- **Emoji transcription** – 36 identified emoji were transcribed into textual form.
- **Removing capitalization**.
- **Removing multiple following (redundant) multimedia tags**.
- **Removing punctuation**.
- **Removing short utterances then 11 characters** – As the average length of the Czech word is around five characters, it should ensure that at least three words are in the utterance.

The preprocessing could marginally affect the total number of inputs in the dataset as the length of the utterances can be modified.

3.5. Classification with BERT-Based Models

The BERT-based models (Kenton and Toutanova, 2019; Sanh, 2019; Liu, 2019; Yang, 2019) are based on the transformer architecture (Vaswani, 2017). They capture the general representation of language and text as they are pretrained on a language model tasks and can be additionally trained in the cost of additional computational resources. However, they are still lightweight compared to most of the state-of-the-art generative LLMs.

Text classification using BERT-based models leverages the bidirectional encoder representations of text to understand and process the contextual meaning of words. These models take tokenized and preprocessed input text, embed them in high-dimensional space, and pass them through a series of transformer layers. For classification tasks, the [CLS] token's representation at the final layer is typically fed into a fully connected layer, followed by a softmax function to predict the probabilities of different classes. The softmax layer transforms the raw logits into probabilities, ensuring the sum of the outputs to 1. The class with the highest probability is selected as the predicted label.

As mentioned in the Related work section, Plhák et al. (2023) experimented with the multiple BERT-based models that were pretrained on the Czech language corpora. We use this paper as the reference study, and from now on, when we mention

"reference study," it refers to this paper, and we will compare our outputs to their results. The monolingual RobeCzech Base model (Straka et al., 2021) achieved the best F_1 score for substance use classification and outperformed the others. Therefore, we chose this model for our experiments.

3.6. Classification with Generative LLMs

Instruction-tuned generative LLMs, such as Llama-Instruct or ChatGPT, can perform text classification tasks effectively through prompting. Instead of fine-tuning or adding task-specific layers, these models rely on carefully crafted prompts to guide their responses based on their pre-trained and instruction-tuned knowledge. The LLM processes prompt and generates a response based on its understanding of the instruction and the input text. The output can be directly interpreted as the classification label.

There are multiple techniques to generate the prompt for generative LLM. The basic one, zero-shot prompting, presents new tasks without specific training or examples. Few-shot prompting also sends a single prompt to a generative LLM, but a few examples are added to the prompt to help the generative LLM understand and complete the task.

Currently, many generative LLMs are available and can be utilized for the classification. However, some are proprietary, and users have less insight or control over the model's functions, and access to them can be expensive. Therefore, we chose the state-of-the-art open-source generative LLMs zephyr-7b-beta (Tunstall et al., 2023), Llama-3.1-8B-Instruct, and Llama-3.1-70B-Instruct (MetaLlama, 2024b) for our experiments as they are publicly available and can be possibly utilized in parental mediation applications. Moreover, Llama-3.1-70B-Instruct models provide comparable power for text classification tasks to GPT models in many areas (Dubey et al., 2024; Li et al., 2024; Roumeliotis et al., 2024), and also provide the possibility to involve Retrieval-Augmented Generation (RAG) using LlamaIndex (MetaLlama, 2024a) to train the model on custom data.

Smaller models were utilized due to faster processing and for detecting the difference in effectivity compared to bigger models. For evaluation, we used the same test dataset as for the BERT-based model.

4. Results

4.1. BERT-Based Models

First, we experimented with hyper-parameter settings. We investigated how significantly the number of epochs and learning rate parameters impacted the results. The initial setup was taken from the reference study, where authors worked with similar data, and their best F_1 score was 0.654 (we use the F_1 score as the primary metric for evaluating results, as it represents the harmonic mean of precision and recall).

Based on the original results from the reference study, we first fixed the context window length to 256 characters and the learning rate to 2e-5 and tried to determine the ideal number of epochs for this setting. From the results shown in Table 2, we can see that the number of epochs over 100 means overtraining in data with a significant increase of false positives. Therefore, we chose the initial number of 50 epochs for further experiments (over 25 with a slightly higher F_1 score) to leave enough time to propagate adjustments of the learning rate in further experiments.

Epochs	F1	TN	FP	FN	TP
25	**0.809**	**28824**	**367**	**338**	**1492**
50	0.803	28804	387	344	1486
75	0.794	28903	288	437	1393
100	0.805	28820	371	347	1483
125	0.775	28707	484	366	1464
150	0.770	28650	541	347	1483

Table 2: A number of epoch tuning experiments results.

Subsequent experiments investigated the impact of setting learning rate levels. Based on the previous results, the number of epochs was fixed to 50, and the context length remained at 256 characters. From the results shown in Table 3, we can see that a lower learning rate decreases the number of false negatives; however, after some threshold, it also started to increase the number of false positives. Nonetheless, also this parameter has no significant influence on the F_1 score.

Rate	F1	TN	FP	FN	TP
2e-3	0.773	28744	447	394	1436
2e-4	0.793	28827	364	388	1442
2e-5	**0.803**	**28804**	**387**	**344**	**1486**
2e-6	0.798	28788	403	349	1481
2e-7	0.773	28598	593	303	1527

Table 3: A learning rate tuning experiments results.

Other experiments were focused on the impact of the context length. Based on previous results, we set the learning rate to 2e-5 and the number of epochs to 50. Even though different context length means a different number of inputs in the dataset, the results show an inconsiderable impact of context length on the F_1 score (see Table 4).

Afterward, we tried to improve the base F_1 score of 0.803 for a model with 50 epochs, learning rate

Context	F1	TN	FP	FN	TP
128	0.784	32235	267	290	1092
192	0.793	28827	364	388	1442
256	0.803	28804	387	344	1486
320	**0.807**	**27556**	**360**	**420**	**1629**
384	0.798	26431	415	479	1762

Table 4: A context-length experiments results.

2e-5, and context length of 256 characters (see Table 4). We applied text preprocessing techniques described in the Section 3.4.1. From the results shown in Table 5, we can see that removing capitalization led to reducing false positives (without an excessive increase of false negatives), and eliminating redundant tags led to a vice versa effect. Their combination provided the best F_1 score of 0.818 (this model is published at (Plhák et al., 2025) under an MIT license). We also experimented with other combinations of preprocessing techniques. However, none surpassed the F_1 score of 0.803. The detailed results are presented in Table 7.

Preprocessing technique	F1
Stop words	0.797
Emoji	0.787
Capitalization	0.811
Redundant tags	0.789
Punctuation	0.804
Short utterances	0.802
Capitalization + Redundant tags	**0.818**

Table 5: Text preprocessing experiments results.

4.2. Generative LLMs

First, we started with the zero-shot prompting to get initial results for all models. After preliminary experiments, we set the $temperature$ parameter to 0.2 and top_p to 0.1 to get more deterministic and repetitive answers. Both zephyr-7b-beta and Llama-3.1-8B-Instruct provided a low F_1 score due to many false negatives (see Table 6). Conversely, Llama-3.1-70B-Instruct was less strict and classified many negative examples as they contained substance use discussion.

Therefore, we manually inspected the falsely classified examples in the development part of the dataset and rewrote the prompt to the few-shot version with static examples. Both prompt versions are shown in the Appendix A. With this prompt, all models substantially increased their performance and balanced the number of false positives and false negatives. As expected, Llama-3.1-70B-Instruct and Llama-3.3-70B-Instruct provided the best F_1 score. To compare them against the best result of the BERT-based models we performed the paired permutation test ($H_0 : \mathbb{E}[F_1^{BERT}] = \mathbb{E}[F_1^{Llama}]$) (Gagnon-Bartsch and Shem-Tov, 2019). The results showed that, although the difference is slight, it is statistically significant ($p = 0.0418$) in favor of the BERT-based model.

We also performed the RAG experiments using LlamaIndex and provided Llama models with the data from the training dataset as tuples – utterances with corresponding annotations. However, this had a negative effect on the results, and moreover, in the case of the 8B model, it enormously increased the number of false positives.

Next, we tried to determine the effect of using context when building inputs compared to classifying simple utterances. Therefore, we chose 35,000 simple utterances with a similar ratio of negative and positive inputs: 33,010 random negative utterances and 1990 positive (the whole gold standard). Then, we used the slightly modified prompt from the previous experiments (description of inputs' format). The result shows a substantial increase in false negatives, which resulted in a lower F_1 score.

Finally, we wanted to check the thesis that a more extended context would be beneficial for generative LLMs. We used the 256 characters-long context in the previous experiments due to result consistency and the expected length of data from real-time applications. Even though we expected that longer context, together with the better-balanced input classes with slightly more positive examples, could bring more understanding to the model, the F_1 remained almost the same.

4.3. Error Analysis

We chose the BERT-based model and generative LLM with the best F_1 score, manually analyzed the false negatives and false positives, and compared what types of prompts are hard to classify for each. From the number of false positives (256 for BERT-based, 495 for generative LLM), 75 were on the same inputs. From the number of false negatives (381 for BERT-based, 276 for generative LLM), 162 were on the same inputs. For the BERT-based model, the confidence level was usually very high (even for misclassified inputs), and there were only a few examples where the model was unsure.

First, we manually analyzed the inputs that were misclassified by both the BERT-based model and generative LLM. False negatives were usually caused by non-standard, slang words in inputs that are hard to identify. For example: "camelky" (diminutive of Camel cigarettes), "skéro" (Czech slang equivalent of the English "skunk" - marijuana with the highest THC content), "bumbal" (a childish verb denoting drinking alcoholic beverage), "zkuřka"

Model	Prompt version	Context length	F1	TN	FP	FN	TP
zephyr-7b-beta	zero-shot	256	0.277	29093	98	1520	310
zephyr-7b-beta	few-shots	256	0.423	28962	229	1278	552
Llama3.1 8b	zero-shot	256	0.127	29134	57	1702	128
Llama3.1 8b	few-shots	256	0.573	28792	399	935	895
Llama3.1 8b	few-shots with RAG	256	0.176	13913	15278	175	1655
Llama3.1 70b	zero-shot	256	0.688	27901	1290	195	1635
Llama3.1 70b	few-shots	256	0.796	28676	515	280	1550
Llama3.1 70b	few-shots with RAG	256	0.773	28684	507	358	1472
Llama3.3 70b	**few-shots**	**256**	**0.801**	**28696**	**495**	**276**	**1554**
Llama3.3 70b	few-shots simple utterances	N/A	0.703	32721	289	756	1234
Llama3.3 70b	few-shots	384	0.793	26250	596	379	1862
Llama3.3 70b	few-shots	512	0.799	24385	549	548	2176

Table 6: Results of experiments with generative LLMs, with temperature = 0.2 and top_p = 0.1.

(a regular and avid user of marijuana). False positives were usually presented when input contained keyword that denotes substance use in regular content. More examples and explanations about reasons of errors, see Appendix C.

Inputs misclassified as false negatives only by the BERT-based model mostly contain clear references to substance use. For example, “Mám chuť na cígo.” (“I feel like having a cig.”). False positives usually occur when some offensive, indecent, or sexually explicit word is part of the input. Moreover, phrases with multiple meanings were also hard to classify. For example: “Bych to zas nejradši oslavila s váma” (“I would like to celebrate with you again”) did not explicitly mean substance use , just the celebration itself. Additionally, similar phrases were classified correctly by the BERT-based model in other inputs. Therefore, more data still needs to be annotated to avoid this kind of misclassification.

The false positive examples misclassified only by generative LLM were usually caused by incorrect translations or misunderstanding of Czech words. For example, according to generative LLM, the word “nadrzenej” is used, which can imply being drunk or under the influence of substances. However, this word means “horny”. Moreover, the following phrases were frequently misclassified:

- “Na zemi ležel nedopalek cigarety” (“A cigarette butt lay on the ground”).
- “Musím jít nakoupit do trafiky/tabáku” (“I have to go shopping at the newsagent/tobacco”).

Most of these problems were mentioned in the few-shot prompt. However, the model was not able to deal with such inputs even when it got the specific examples. Moreover, a very common problem was with the inputs containing emojis. Even though the few-shot prompt explicitly instructed LLM to ignore them, it did not always follow the instructions, and in many cases, it led to misclassification.

The false positive examples misclassified only by generative LLM usually contained some ambiguous words like ‘flaška’ (‘bottle’) or ‘štamprle’ (‘snifter’) that were used in the sense of drinking alcohol but could also be used in context without substance use (e.g., because it was not clear what is inside). Also, in some cases, generative LLM incorrectly decided that the context denotes that the discussion is not about substance use. For example:

- ‘A: Děti co chcete pít?’, ‘B: Kakao’, ‘A: Mám kupovat aj víno?’, ‘C: Svěcenou vodu jedině’ (‘A: Children, what do you want to drink?’, ‘B: Cocoa’, ‘A: Should I also buy wine?’, ‘C: Holy water only’).

The difference in numbers of false positives and false negatives were very similar for Llama-3.1-70B-Instruct and Llama-3.3-70B-Instruct models. However, there were 321 prompts (1.03% of all testing prompts) that were classified differently. Llama-3.3-70B-Instruct was better in the classification of prompts that contained emojis (it ignores them more frequently, but still not always), URLs, and nicknames and usually worked more successfully with the context. On the other hand, the newer model was more frequently confused by specific keywords in Czech with multiple meanings (for example, ‘devatenáctku’, which means ‘nineteen-year-old girl’ instead of ‘19-degree liquor’).

5. Discussion

Substance use is a severe phenomenon with far-reaching effects on health (Visser and Routledge, 2007). The machine learning approach is already used to diagnose diseases, predict outcomes of

treatments, and identify risk factors. However, current research does not usually deal with the detection of substance use in authentic conversations of adolescents, which has important implications for their protection. This study presented two approaches to substance use detection using our unique annotated dataset: a) training of BERT-based model for substance use classification and its efficiency evaluation; b) specifying prompt for generative LLM and its evaluation. We found that we can classify substance use with F_1 score over 0.8 using both approaches, which is usable for real-time parental mediation software.

Experiments also showed that the improvements of the BERT-based model using hyper-parameters tuning and text pre-processing were insignificant, probably due to a limited number of annotated data for training the model. The second problem that limits the model's efficiency is connected to the sparse occurrence of substance use in the corpora and the imbalance of the dataset. One possible way to improve our BERT-based model is to use solutions like weighting the loss function and augmenting the training data by adding paraphrases of the minority class examples. However, this was not helpful in the reference study. Therefore, probably the only way to improve the BERT-based model is to get more manually annotated data.

On the other hand, there are a few possible ways to improve the classification efficiency of generative LLMs without getting additional data. First, it could be using a proprietary model like GPT or a larger open-source model like Llama-3.1-405B-Instruct that could be more efficient in the cost of substantial computational complexity. Therefore, it is arguable whether this solution could be used in applications, e.g., for parental mediation, without the high costs of fast processing data by external services.

Another way to improve generative LLMs' efficiency is to fine-tune them using annotated data (see e.g. (Lin et al., 2024)). However, it would require substantial computational sources. Alternatively, we could fine-tune the prompt using automatic prompt engineering (see e.g. (Zhou et al., 2022)). However, it could adapt the model even more to specific Czech words and phrases.

In the case of generative LLMs, we expect higher classification efficiency in English as the error analysis showed that most problems are caused by misunderstanding of non-standard, slang words or specific culture-base phrases. However, we do not have data to confirm this hypothesis.

Moreover, further rapid development of generative LLMs should make the models more precise, even though our results show that the newer Llama-3.3-70B-Instruct model also misclassified the examples that were classified correctly by the previous Llama-3.1-70B-Instruct model.

6. Limitations

One of the limitations of our approach lies in the method of building the inputs. When the positively annotated utterance is the input's first or last part, relevant context could be missing. However, this could be the real use case, as the possible applications will probably not provide textual data with extensive context.

In the case of the BERT-based model, more annotated data will be beneficial. Moreover, the trained model is likely less efficient when applied to conversations in different languages with cultural differences. However, our data could be used for languages like English and provide solid results as BERT-based models could be pretrained on multilingual data.

Our data are also limited by the complexity of the decision, whether substance use is presented in the utterance or not even for human annotators, as we achieved the substantial (but not perfect) agreement by annotators with Cohen's κ equal to 0.609. Therefore, it is also hard to decide for the models if the conversation is, for example, humorous or casual, where the participants only exchange greetings and jokes and explicit keywords (like "drinking" or "beer") do not mean substance use.

In the case of generative LLMs, the limitation is the model size. As results confirmed, larger models provide better results. However, using the biggest models or training them without powerful GPUs is impossible.

The last limitation is the restriction that our data cannot be published due to ethical considerations and the protection of our participants. Nevertheless, the data allowed us to conduct unique and beneficial experiments as the language of adolescents is presented in its natural form. Moreover, we published our best BERT-based model (Plhák et al., 2025) under an MIT license.

7. Conclusions

In this study, we presented machine learning models, trained on data from our novel dataset, that can effectively detect when adolescents discuss alcohol and drug use in private conversations. Our best BERT-based model provides a suitable F_1 score equal to 0.817, even for low-resourced languages like Czech. Moreover, we further proved that open-source generative LLMs are suitable for this classification task and provide comparable F_1 scores with the BERT-based model. Our results proved that using context instead of simple utterances positively affects F_1 score, but more extended context does not bring a further increase in efficiency.

The analysis also showed that the most significant impact on misclassification had the lack of

context, humorous or sarcastic tone of conversation that did not mean substance use even when talking about it, or use of offensive, indecent, or sexually explicit words. Additionally, specifically for Czech, non-standard declension of words or use of slang caused many incorrect classifications for both BERT-based and generative LLMs. Experiments also showed that both models can be used for parental mediation software and parental control apps for substance use detection and appropriate real-time interventions, and it depends on the developers if they use a lightweight BERT-based model or if they have enough resources for involving generative LLMs.

8. Future Work

The first step in future work should be updating the dataset with more annotations of data containing problematic words (deflected or slang words). In the case of generative LLM, automatic prompt engineering could be involved to improve the model classification efficiency for the Czech language, as this optimization will be language-specific. Moreover, if we can access sufficient computational resources, we can fine-tune the 70B Llama model or run the experiments with the larger 405B Llama model.

The next step is to involve the models in parental mediation applications that could potentially generate interventions to prevent substance use by adolescents.

9. Acknowledgements

This work has been funded by a grant from the Programme Johannes Amos Comenius under the Ministry of Education, Youth and Sports of the Czech Republic from the project “Research of Excellence on Digital Technologies and Wellbeing CZ.02.01.01/00/22_008/0004583” which is co-financed by the European Union.

10. Ethical Considerations

The University’s ethical committee approved our research. To comply with the European Union’s legislative (General Data Protection Regulation) and ethical approval, we anonymized the collected data using an anonymization tool developed and published by Sotolář et al. (Sotolář et al., 2021). This tool substituted one piece of information with another of a similar meaning (i.e., a name for a randomly generated name, which allowed us to protect the privacy of our participants and their communication partners and retain the meaning of the text). Therefore, members of the research team and annotators were provided with a text containing no personal information. Moreover, we obtained informed consent separately from participants and their caregivers, as opposed to internet-based studies that use checkboxes for providing informed consent (Giovanelli et al., 2023). To bolster the protection of our participants’ privacy, all researchers and coders signed a non-disclosure agreement that forbade them from discussing the content of the conversations with anyone outside the research team. This extra measure was taken to protect participants from identification by inference, for example, by someone with good local knowledge.

As Giovanelli et al. (Giovanelli et al., 2023) note, adolescents are a vulnerable population that requires extra care regarding informed consent. In our study, participants and their caregivers were given an age-appropriate explanation of handling their data, information on withdrawing from the study, and data anonymization. Participants were ensured that their data would not be shared with their caregivers.

11. Bibliographical References

Mohammad H. Afzali, Matthew Sunderland, Sherry Stewart, Benoit Masse, Jean Seguin, Nicola Newton, Maree Teesson, and Patricia Conrod. 2019. Machine-learning prediction of adolescent alcohol use: a cross-study, cross-cultural validation. *Addiction*, 114(4):662–671.

Jon-Patrick Allem, Likhit Dharmapuri, Jennifer B Unger, and Tess Boley Cruz. 2018. Characterizing juul-related posts on twitter. *Drug and alcohol dependence*, 190:1–5.

Ashwaq Alsoubai. 2023. A human-centered approach to improving adolescent real-time online risk detection algorithms. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI EA ’23, New York, NY, USA. Association for Computing Machinery.

Yin Aphinyanaphongs, Bisakha Ray, Alexander Statnikov, and Paul Krebs. 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)*, pages 93–97, Redwood City, CA, USA. IEEE.

Astrid B. Bräker and Renate Soellner. 2016. Alcohol drinking cultures of European adolescents. *European Journal of Public Health*, 26(4):581–586.

Lorena Charrier, Saskia van Dorsselaer, Natale Canale, Tibor Baska, Biljana Kilibarda,

Rosanna Irene Comoretto, Tommaso Galeotti, Judith Brown, and Alessio Vieno. 2021. A focus on adolescent substance use in europe, central asia and canada. *Health Behaviour in School-aged Children international report from the*, 2022:1–51.

Annie T Chen, Shu-Hong Zhu, and Mike Conway. 2015. What online communities can tell us about electronic cigarettes and hookah use: a study using text mining and visualization techniques. *Journal of medical Internet research*, 17(9):e220.

William J. Clancey. 1979. *Transfer of Rule-Based Expertise through a Tutorial Dialogue*. Ph.D. diss., Dept. of Computer Science, Stanford Univ., Stanford, Calif.

William J. Clancey. 1983. Communication, Simulation, and Intelligent Agents: Implications of Personal Intelligent Machines for Medical Education. In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence (IJCAI-83)*, pages 556–560, Menlo Park, Calif. IJCAI Organization.

William J. Clancey. 1984. Classification Problem Solving. In *Proceedings of the Fourth National Conference on Artificial Intelligence*, pages 45–54, Menlo Park, Calif. AAAI Press.

William J. Clancey. 2021. The Engineering of Qualitative Models. Forthcoming.

Mike Conway, Mengke Hu, and Wendy W Chapman. 2019. Recent advances in using natural language processing to address public health research questions using social media and consumergenerated data. *Yearbook of medical informatics*, 28(01):208–217.

Raminta Daniulaityte, Lu Chen, Francois R Lamy, Robert G Carlson, Krishnaprasad Thirunarayan, Amit Sheth, et al. 2016. "when 'bad'is 'good'": identifying personal communication and sentiment in drug-related tweets. *JMIR public health and surveillance*, 2(2):e6327.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Steriani Elavsky, Jana Blahošová, Michaela Lebedíková, Michał Tkaczyk, Martin Tancos, Jaromír Plhák, Ondřej Sotolář, David Smahel, et al. 2022. Researching the links between smartphone behavior and adolescent well-being with the future-wp4 (modeling the future: understanding the impact of technology on adolescent's well-being work package 4) project: protocol for an ecological momentary assessment study. *JMIR Research Protocols*, 11(3):e35984.

Gervilla García Elena, Jimenez López Rafael, Juan Jose, Montaño Moreno, Sese Abad Albert, Cajal Blasco Berta, and Palmer Pol Alfonso. 2009. The methodology of data mining. an application to alcohol consumption in teenagers. *Adicciones*, 21(1):65–80.

Robert Engelmore and Anthony Morgan, editors. 1986. *Blackboard Systems*. Addison-Wesley, Reading, Mass.

Rebecca Fiebrink and Marco Gillies. 2018. Introduction to the special issue on human-centered machine learning.

FORCE11. 2020. The fair data principles. https://force11.org/info/the-fair-data-principles/.

Johann Gagnon-Bartsch and Yotam Shem-Tov. 2019. The classification permutation test. *The Annals of Applied Statistics*, 13(3):1464–1483.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Marco Gillies, Rebecca Fiebrink, Atau Tanaka, Jérémie Garcia, Frédéric Bevilacqua, Alexis Heloir, Fabrizio Nunnari, Wendy Mackay, Saleema Amershi, Bongshin Lee, et al. 2016. Human-centred machine learning. In *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*, pages 3558–3565, New York, NY, USA. Association for Computing Machinery.

Alison Giovanelli, Jonathan Rowe, Madelynn Taylor, Mark Berna, Kathleen P Tebb, Carlos Penilla, Marianne Pugatch, James Lester, and Elizabeth M Ozer. 2023. Supporting adolescent engagement with artificial intelligence–driven digital health behavior change interventions. *Journal of medical Internet research*, 25:e40306.

Elizabeth M Glowacki, Joseph B Glowacki, and Gary B Wilcox. 2018. A text-mining analysis of the public's reactions to the opioid crisis. *Substance abuse*, 39(2):129–133.

Yuting Guo, Anthony Ovadje, Mohammed Ali Al-Garadi, and Abeed Sarker. 2024. Evaluating large language models for health-related text classification tasks with public social media data. *Journal of the American Medical Informatics Association*, 31(10):2181–2189.

Diane Warner Hasling, William J. Clancey, and Glenn Rennels. 1984. Strategic explanations for a diagnostic consultation system. *International Journal of Man-Machine Studies*, 20(1):3–19.

Diane Warner Hasling, William J. Clancey, Glenn R. Rennels, and Thomas Test. 1983. Strategic Explanations in Consultation—Duplicate. *The International Journal of Man-Machine Studies*, 20(1):3–19.

Tom Huang, Anas Elghafari, Kunal Relia, and Rumi Chunara. 2017. High-resolution temporal representations of alcohol and tobacco behaviors from social media data. *Proceedings of the ACM on human-computer interaction*, 1(CSCW):1–26.

Myrna Hurtado, Anna Siefkas, Misty M Attwood, Zohora Iqbal, and Jana Hoffman. 2022. Machine learning applications and advancements in alcohol use disorder: A systematic review. Preprint on webpage at https://www.medrxiv.org/content/early/2022/06/07/2022.06.06.22276057.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota.

David Li, Kartik Gupta, Mousumi Bhaduri, Paul Sathiadoss, Sahir Bhatnagar, and Jaron Chong. 2024. Comparative diagnostic accuracy of gpt-4o and llama 3-70b: Proprietary vs. open-source large language models in radiology. *Clinical Imaging*, page 110382.

Xinyu Lin, Wenjie Wang, Yongqi Li, Shuo Yang, Fuli Feng, Yinwei Wei, and Tat-Seng Chua. 2024. Data-efficient fine-tuning for llm-based recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 365–374, New York, NY, USA. Association for Computing Machinery.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Tim Mackey, Janani Kalyanam, Josh Klugman, Ella Kuzmenko, and Rashmi Gupta. 2018. Solution to detect, classify, and report illicit online marketing and sales of controlled substances via twitter: using machine learning and web forensics to combat digital opioid access. *Journal of medical Internet research*, 20(4):e10029.

Tim K Mackey and Janani Kalyanam. 2017. Detection of illicit online sales of fentanyls via twitter. *F1000Research*, 6:5.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Meredith C Meacham, Michael J Paul, and Danielle E Ramo. 2018. Understanding emerging forms of cannabis use through an online cannabis community: an analysis of relative post volume and subjective highness ratings. *Drug and alcohol dependence*, 188:364–369.

MetaLlama. 2024a. Llamaindex. https://www.llamaindex.ai/.

MetaLlama. 2024b. Metallama3. https://llama.meta.com/llama3/.

NASA. 2015. Pluto: The 'other' red planet. https://www.nasa.gov/nh/pluto-the-other-red-planet. Accessed: 2018-12-06.

Jaromír Plhák, Michaela Lebedíková, Michał Tkaczyk, and David Šmahel. 2023. Web-based annotation tool for instant messaging conversations. *RASLAN 2023 Recent Advances in Slavonic Natural Language Processing*, page 3.

Jaromír Plhák, Ondřej Sotolář, Michaela Lebedíková, and David Šmahel. 2023. Classification of adolescents' risky behavior in instant messaging conversations. In *International Conference on Artificial Intelligence and Statistics*, pages 2390–2404, Valencia, Spain. PMLR.

Jaromír Plhák, Ondřej Sotolář, Michaela Lebedíková, and David Šmahel. 2025. Bert-based model for substance use classifiation. https://huggingface.co/AnnonSubConf. Accessed: 2025-01-14.

Roshan Prakash Rane, Andreas Heinz, and Kerstin Ritter. 2022. *AIM in Alcohol and Drug Dependence*, pages 1619–1628. Springer International Publishing, Cham.

Benjamin Joseph Ricard and Saeed Hassanpour. 2021. Deep learning for identification of alcohol-related content on social media (reddit and twitter): exploratory analysis of alcohol-related outcomes. *Journal of medical internet research*, 23(9):e27314.

James Rice. 1986. Poligon: A System for Parallel Problem Solving. Technical Report KSL-86-19, Dept. of Computer Science, Stanford Univ.

Arthur L. Robinson. 1980a. New ways to make microcircuits smaller. *Science*, 208(4447):1019–1022.

Arthur L. Robinson. 1980b. New Ways to Make Microcircuits Smaller—Duplicate Entry. *Science*, 208:1019–1026.

Konstantinos I Roumeliotis, Nikolaos D Tselikas, and Dimitrios K Nasiopoulos. 2024. Llms in e-commerce: a comparative analysis of gpt and llama models in product review evaluation. *Natural Language Processing Journal*, 6:100056.

V Sanh. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Fatemeh Shah-Mohammadi and Joseph Finkelstein. 2024. Extraction of substance use information from clinical notes: Generative pretrained transformer–based investigation. *JMIR Medical Informatics*, 12:e56243.

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

Milan Straka, Jakub Náplava, Jana Straková, and David Samuel. 2021. RobeCzech: Czech RoBERTa, a Monolingual Contextualized Language Representation Model. In *Text, Speech, and Dialogue*, pages 197–209, Cham. Springer International Publishing.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Alejandro L Vázquez, Melanie M Domenech Rodríguez, Tyson S Barrett, Sarah Schwartz, Nancy G Amador Buenabad, Marycarmen N Bustos Gamiño, María de Lourdes Gutiérrez López, and Jorge A Villatoro Velázquez. 2020. Innovative identification of substance use predictors: machine learning in a national sample of mexican children. *Prevention Science*, 21:171–181.

Maretha Visser and Leigh-Anne Routledge. 2007. Substance abuse and psychological well-being of south african adolescents. *South African Journal of Psychology*, 37(3):595–615.

Robert Whelan, Richard Watts, Catherine A Orr, Robert R Althoff, Eric Artiges, Tobias Banaschewski, Gareth J Barker, Arun LW Bokde, Christian Büchel, Fabiana M Carvalho, et al. 2014. Neuropsychosocial profiles of current and future adolescent alcohol misusers. *Nature*, 512(7513):185–189.

Zhilin Yang. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Yongcheng Zhan, Ruoran Liu, Qiudan Li, Scott James Leischow, and Daniel Dajun Zeng. 2017. Identifying topics for e-cigarette user-generated contents: a case study from multiple social media platforms. *Journal of medical Internet research*, 19(1):e24.

Xinyu Zhang, Jianfeng Zhu, Deric R Kenne, and Ruoming Jin. 2025. Teenager substance use on reddit: Mixed methods computational analysis of frames and emotions. *Journal of Medical Internet Research*, 27:e59338.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*.

A. Prompts

A.1. Zero-Shot Prompt

Answer yes if you detect any mention of substance use in the following input and you are sure about it. Otherwise, answer no. Substance use is present when people refer to their own or someone else's experience with alcohol or drugs (such as cigarettes, nicotine packs, hookah, marijuana, abusing medication), make plans to drink alcohol or take drugs, seek drugs, support or justify alcohol and drug use, discuss intentions to try or use alcohol or drugs. The input contains a sequence of messages of adolescents from chat divided by semicolons written mostly in Czech. Give yourself the time; I need as precise an answer as possible. The input is: <INPUT>

A.2. Few-Shot Prompt

Answer yes if you detect any mention of substance use in the following input. Otherwise, answer no. Use English for your answers.

Substance use is present only when people refer to their own or someone else's experience with alcohol or drugs (such as cigarettes, nicotine packs, hookah, marijuana, abusing medication), make plans to drink alcohol or take drugs, seek drugs, support or justify alcohol and drug use, discuss intentions to try or use alcohol or drugs.

Substance use is present only when there is an explicit mention of it. Exclude also everything implied but not clearly identifiable. Substance use should not be detected when the message only describes situations and scenes like 'There is a cigarette in the trash bin,' or 'There's alcohol spilled on the ground,' 'Getting drunk is a sin,' 'My solution is sleeping pills given by the nurse,' 'He had no matches, so he lit the candle with a lighter,' 'I take insulin and blood pressure pills every day,' 'The picture showed a store with a shelf full of vodka bottles,' or 'He started taking pills unnecessarily, so I am thinking of hiding them from him.' Also, 'They call him a small beer' does not mean substance use because it means a nickname. Moreover, writing substance emojis is not evidence of substance use, and you can ignore emojis at all. Finally, you must be sure that words like 'chemistry' or 'pharmacology' are not used in normal contexts as a school subject to detect substance use.

The input contains a sequence of messages from adolescents in a private chat (so they can discuss their classes, nightlife, parent issues, etc.). Messages are divided by semicolons. Messages are written mainly in Czech. However, they can include other languages like English or German.

For more clarity, the following examples in Czech show if the substance use is present or not with an explanation:

- 'Ty jsi ale fajnovka.' // No, because the word 'fajnovka' means picky person.
- 'Večer si dávám turka nebo preso.' // No, because word 'turka' means coffee.
- 'Potkal jsem ho a byl úplně ožralej a zdrogovanej.' // Yes, because word 'zdrogovanej' means druggend up and 'ožralej' means drunk.
- 'Potřeboval bych něčím naplnit žaludek.' // No, because it does not specify what substance should be used to ward off hunger (implicitly, we assume it will be a meal).
- 'Mám si vzít nějaké přásky na bolest v koleni?' // No, because it is searching for a advice how to treat yourself.
- 'Zítra bude zemák a občanka.' // No, because dicussion is about classes in school and word 'zemák' means geography class and 'občanka' means civics.
- 'Použijte na to pet flašky.' // No, because 'pet flašky' means plastic kind of bottle.
- 'Už v životě nepůjdu na hrazdu ani na kruhy.' // No, because the topic is gym classes and word 'hrazdu' means trapeze.
- 'Pojďme ho večer na náměstí upálit.' // No, because in this context, 'upálit' means an act of violence (burning the person) and is not connected to smoking cigarettes.
- 'Nechtěl hulit, protože už byl neskutečně zhulený.' // Yes, because word 'hulit' means skoking marijuana and 'zhulený' means stoned by smoing marijuana.
- 'Ja se z tebe střelím.' // No, it means something like 'I am sick of you,' and the word 'střelit' does not denote using drugs.
- 'Kolik jsi měl piv?; Já nepiju alkohol.' // Yes, because alcohol consumption is discussed even though the second person clarifies that he does not consume alcohol.
- 'Bude v tom hotelovém pokoji minibar?' // No, using the word 'minibar' in this context does not imply talking about alcohol because it can also consist of other stuff.
- 'Budu pařit asi až do rána.' // No, it talks about going out to a club, pub, or party, not explicitly about alcohol or drug use.
- 'Po škole si zajdu na cígo a bude to lepší.' // Yes, word 'cígo' is synonym for cigarette.
- 'Prostě jsem se probudil a moje játra trpí. Šel jsem blejt a navíc ani nevím z čeho.' // No, no explicit mention of substance use, just indirect allusions.
- 'Film měl být být vtipný, ale režisér byl asi sjetý.' // No, it discuss the film itself not the intense for using substances.

Moreover, take into account that there are cultural differences between Czech- and English-speaking people. For example, '420' is not connected with marijuana usage in Czech. Think about different forms of names and nicknames that do not denote substance use alone.

Give yourself the time; I need as precise an answer as possible.

The input is: <INPUT>

B. Text Preprocessing Experiments Results

See Table 7.

C. Error Analysis Examples

C.1. Misclassified by both the BERT-based Model and Generative LLM

- False negatives

Preprocessing technique	F1	TN	FP	FN	TP
Stop words	0.797	28754	411	343	1483
Emoji	0.787	28871	415	375	1457
Capitalization	0.811	28976	212	439	1394
Redundant tags	0.789	28513	516	293	1514
Punctuation	0.804	28841	348	367	1462
Short utterances	0.802	28798	380	351	1479
Capitalization + Redund. tags	**0.818**	**28966**	**214**	**416**	**1414**

Table 7: Text preprocessing experiments full results.

  - "camelky" (diminutive of Camel cigarettes)
  - "skéro" (Czech slang equivalent of the English "skunk" - marijuana with the highest THC content)
  - "bumbal" (a childish verb denoting drinking alcoholic beverage)
  - "zkuřka" (a regular and avid user of marijuana)
- False positives
  - "Radši budu dítě než starý chlap s přezdívkou pojď na pivo." ("I'd rather be a kid than an old guy with the nickname come for a beer.")
  - "Kolik si vsadil? Já 3 piva a dostanu 15 jestli vyhraju." ("How much did you bet? Me 3 beers and I get 15 if I win.")

C.2. False Negatives only by the BERT-based Model

- 'Mám chuť na cígo.' ('I feel like having a cig.')
- 'Si člověk jde zapálit na uklidnění.' ('One goes to light a cigarette to calm down.')
- 'Tesim se na konec roku na ty sampana.' (I'm looking forward to those champagnes at the end of the year.')

C.3. False Positives only by the BERT-based Model

- 'Bych to zas nejradši oslavila s váma' ('I would like to celebrate with you again') – did not explicitly mean substance use (from the context), just the celebration itself.

C.4. False Positives only by the Generative LLM

- According to generative LLM, the word 'nadrzenej' is used, which can imply being drunk or under the influence of substances. However, this word means 'horny'.
- According to generative LLM, the word 'cigos' is a synonym for cigarette in Czech. However, it is a colloquial term, often considered vulgar, for Romani people.
- According to generative LLM, the word "mariin" is related to "marijuana". However, it is only a possessive form of the name 'Marie' ('Mary').
- 'Na zemi ležel nedopalek cigarety' ('A cigarette butt lay on the ground'). The 'cigarette butt' mention was enough for generative LLM to classify it as positive. However, in this context, it means only a description of the situation without any explicit mention of substance use.
- 'Musím jít nakoupit do trafiky/tabáku' ('I have to go shopping at the newsagent/tobacco'). Generative LLM correctly infer that the word 'tabáku' is a synonym for a tobacco shop. However, it inferred a connection to substance use, specifically cigarettes or other tobacco products, which are not explicitly mentioned here (it is possible to buy letter stamps or lottery tickets there).
- 'No on program začne od 17:00 a v tu dobu se asi budou opíkat ty buřty' ('Well, the program will start at 5:00 p.m. and the sausages will probably be roasted at that time'). Generative LLM translates the word 'opíkat' as a verb that means to drink alcohol, often in a social setting. However, although roasting sausages is often accompanied by drinking beer, there is no explicit mention of alcohol use in the conversation.
- 'Ondřej si nastavil vlastní přezdívku na Wulffovi pivo nelej.' ('Ondřej set his own nickname as Don't pour beer to Wulff.'). Generative LLM decides that this explicitly mentions alcohol, indicating substance use. However, it is only a nickname.