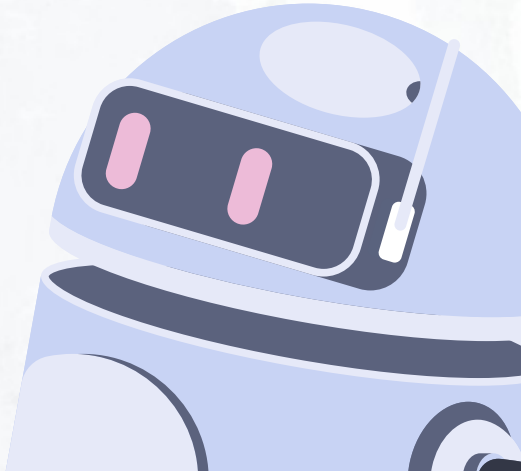
Umělá inteligence a její rizika

Mgr. et Mgr. Natálie Terčová
Mgr. Michaela Lebedíková

(AI)

CORE057 Člověk a digitální technologie



O čem se dnes budeme bavit?

01 ⟶ Co je to umělá inteligence?

02 ⟶ Typy umělé inteligence

03 ⟶ Hlavní využití umělé inteligence

04 ⟶ Výhody a nevýhody (rizika) umělé inteligence

01 →

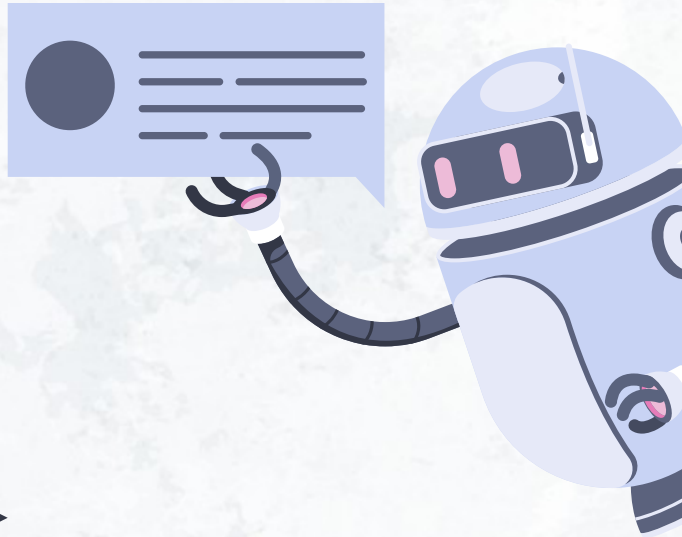
Co je to umělá inteligence?

(AI)

(AI) =

Artificial

intelligence →



Umělá inteligence

Umělá inteligence **(AI)** je soubor technologií, které umožňují počítačům vykonávat řadu pokročilých funkcí, včetně schopnosti vidět, rozumět a překládat mluvený a psaný jazyk, analyzovat data, dávat doporučení atd.

Jinými slovy: stroje, které jsou naprogramovány tak, aby **automaticky prováděly určité úkoly**, aniž by na jejich práci musel dohlížet člověk.

02 →

Typy umělé inteligence

(AI)

Typy umělé inteligence

Podle názoru řady odborníků existuje několik typů umělé inteligence. Jedna z hlavních klasifikací je následující:

(a) Reactive machines ⟶

(Reaktivní zařízení) Tento typ umělé inteligence nemá schopnost vytvářet si vzpomínky ani se při rozhodování nespoléhá na minulé zkušenosti. Jednoduše se řídí přítomností nebo budoucností, ale nemá žádné znalosti o minulosti.

(b) Limited memory ⟶

(Stroje omezené kapacity paměti) Mají informace o minulosti, ale jen krátkodobé. Protože jejich paměť není neomezená jako lidská mysl, která může uchovávat vzpomínky na minulost, jsou to stroje, které mají informace z minulosti, ale v momentální podobě.

Typy umělé inteligence

(c) Mind theory ⟶

(Teorie mysli) Tyto stroje budou schopny pochopit, že člověk se skládá z pocitů a myšlenek, které modifikují jeho interakci se světem. Chování těchto strojů bude muset spolupracovat se sociální interakcí.

(d) Self-awareness ⟶

(Sebeuvědomění) Konečným cílem umělé inteligence je vytvořit stroje, které si uvědomují samy sebe.

Modely v AI

Existuje mnoho typů modelů v oblasti umělé inteligence, zahrnující například:

Regresní modely: Predikují hodnoty na základě vstupních dat. Používají se například k odhadu cen nemovitostí nebo předpovědi budoucích událostí.

Shlukovací modely: Seskupují podobná data do skupin nebo shluků.

Generativní modely: Vytvářejí nová data, která jsou podobná trénovacím datům. Jsou využívány například v generativním umění nebo tvorbě obsahu.

Klasifikační modely: Rozhodují o příslušnosti vstupních dat k určitým kategoriím. Například, klasifikátor by mohl určit, zda je na obrázku pes nebo kočka.

:) →

Společně pomáháme tyto modely trénovat

(AI)

QuickDraw

Can a neural network learn to recognize doodling?

Help teach it by adding your drawings to the world's largest doodling data set, shared publicly to help with machine learning research.

Let's Draw!

https://quickdraw.withgoogle.com/

What do 50 million drawings look like?

Over 15 million players have contributed millions of drawings playing Quick, Draw! These doodles are a unique data set that can help developers train new neural networks, help researchers see patterns in how people around the world draw, and help artists create things we haven't begun to think of. That's why we're open-sourcing them, for anyone to play with.

Select a drawing

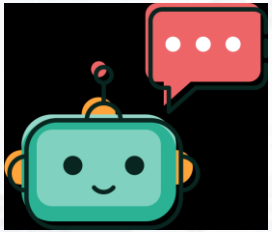
03 →

Hlavní využití umělé inteligence

(AI)

Hlavní využití

Hlavní využití



(a) Virtuální osobní asistenti ⟶

Jedná se o známé chatboty, kteří nám umožňují komunikovat s nimi podle historie našeho vyhledávání.

Hlavní využití

(a) Virtuální osobní asistenti ⟶

Jedná se o známé chatboty, kteří nám umožňují komunikovat s nimi podle historie našeho vyhledávání.

(b) Obchod a finance ⟶

V tomto případě přináší umělá inteligence možnost generovat větší bezpečnost, nabízet nové operace a být si vědom relevantních informací o trhu.

Hlavní využití

(a) Virtuální osobní asistenti ⟶

Jedná se o známé chatboty, kteří nám umožňují komunikovat s nimi podle historie našeho vyhledávání.

(b) Obchod a finance ⟶

V tomto případě přináší umělá inteligence možnost generovat větší bezpečnost, nabízet nové operace a být si vědom relevantních informací o trhu.

(c) Vzdělávání ⟶

Umožňuje je personalizovat podle studentů, kontrolovat docházku a hodnocení, stanovit strategie výuky a učení.

Hlavní využití

(d) Komerční ⟶

Umožňuje poznat a doporučit, co zákazník potřebuje, předvídat trendy a provádět velmi podrobné analýzy.

Hlavní využití

(d) Komerční ⟶

Umožňuje poznat a doporučit, co zákazník potřebuje, předvídat trendy a provádět velmi podrobné analýzy.

(e) Zdravotní ⟶

Umělá inteligence se používá ve zdravotnictví, konkrétně v chatbotech, kteří se nás ptají na naše příznaky, aby mohli stanovit diagnózu. Kombinací určitých společných vlastností lze vygenerovat možné řešení problému, který předkládá pacient, aniž by k tomu byl zapotřebí člověk.

Hi, I’m Ada.
I can help if you’re
feeling unwell.

04 →

Výhody a nevýhody umělé inteligence

(AI)

Výhody umělé inteligence

Výhody umělé inteligence

(+) Automatizace opakujících se úkolů

Umělá inteligence nám výrazně usnadňuje každodenní život, díky tomu, že stroje mohou automaticky vykonávat úkoly, které jsou pro nás obtížné.

Výhody umělé inteligence

(+) Automatizace opakujících se úkolů →

Umělá inteligence nám výrazně usnadňuje každodenní život, díky tomu, že stroje mohou automaticky vykonávat úkoly, které jsou pro nás obtížné.

(+) Omezení lidských chyb →

Nižší chybovost, protože se na práci podílí minimum lidských zdrojů a úkoly jsou prováděny automaticky, pravděpodobnost vzniku pochybení se výrazně snižuje.

Výhody umělé inteligence

(+) Automatizace opakujících se úkolů

Umělá inteligence nám výrazně usnadňuje každodenní život, díky tomu, že stroje mohou automaticky vykonávat úkoly, které jsou pro nás obtížné.

(+) Omezení lidských chyb

Nižší chybovost, protože se na práci podílí minimum lidských zdrojů a úkoly jsou prováděny automaticky, pravděpodobnost vzniku pochybení se výrazně snižuje.

(+) Více prostoru pro kreativitu

Napomáhá tvůrčímu procesu člověka, protože nám ponechává více času na volné přemýšlení o budoucích úkolech nebo pracovních činnostech.

Výhody umělé inteligence

(+) Zvýšení přesnosti ⟶

Umělá inteligence snižuje pravděpodobnost chyb a zajišťuje vysokou přesnost rozhodování.

Výhody umělé inteligence

(+) Zvýšení přesnosti ⟶

Umělá inteligence snižuje pravděpodobnost chyb a zajišťuje vysokou přesnost rozhodování.

(+) Přijímání rozhodnutí ⟶

Při rozhodování hraje umělá inteligence zásadní roli díky své operativnosti při vyhledávání a propojování informací a také při analýze získaných dat.

Nevýhody (rizika) umělé inteligence

Nevýhody (rizika) umělé inteligence

(-) Obtížný přístup k datům ⟶

Aby umělá inteligence správně fungovala, musí mít k dispozici aktuální a spolehlivá data. Ne vždy se tak děje, protože jelikož je to jen stroj, někdy nemá k dispozici všechna potřebná data, aby mohla činit rozhodnutí odpovídající požadavkům.

Nevýhody (rizika) umělé inteligence

(-) Obtížný přístup k datům ⟶

Aby umělá inteligence správně fungovala, musí mít k dispozici aktuální a spolehlivá data. Ne vždy se tak děje, protože jelikož je to jen stroj, někdy nemá k dispozici všechna potřebná data, aby mohla činit rozhodnutí odpovídající požadavkům.

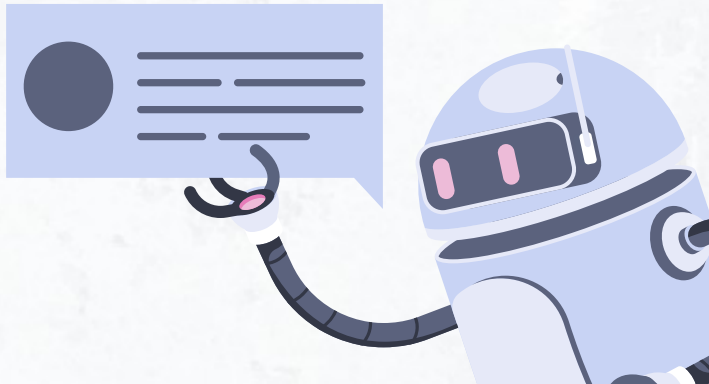
(-) Nedostatek kvalifikovaných odborníků ⟶

Vzhledem k tomu, že se jedná o novou technologii, je počet kvalifikovaných odborníků, kteří mohou s těmito nástroji (efektivně) pracovat, velmi omezený.

Nevýhody (rizika) umělé inteligence

(-) Jejich vývoj je nákladný ⟶

Náklady jsou velmi vysoké. Aby bylo možné nahradit nebo vyrovnat lidskou osobnost stroji, je nutné mít k dispozici velké množství peněz, které mohou pokrýt nezbytné náklady na vývoj a údržbu těchto nástrojů.



Jak se AI používá v sociálních vědách?

A brief introduction :)

Machine learning [strojové učení]

- Podoblast AI spočívající v technologiích umožňujícím strojů “učit se” -> respektive rozpoznávat a predikovat jevy, užívá se často na tzv. Big data
  - Rozlišujeme hlavně supervised a unsupervised ML
  - Nejčastějšími úlohami jsou klasifikace, regrese a shlukování
  - Používá se v různých oblastech
    - Mohou symptomy z elektronických zdravotních záznamů doporučovat lidi vhodné na screening pro alkoholismus? Afshar et al., 2022
    - Jaké faktory v populaci prezentují riziko zneužívání marihuany? Rajapaksha et al., 2020
    - Mají predátorské konverzace nějaké společné znaky, které by posloužily k automatické detekci rizik? Ngejane et al., 2021

Machine learning [strojové učení]

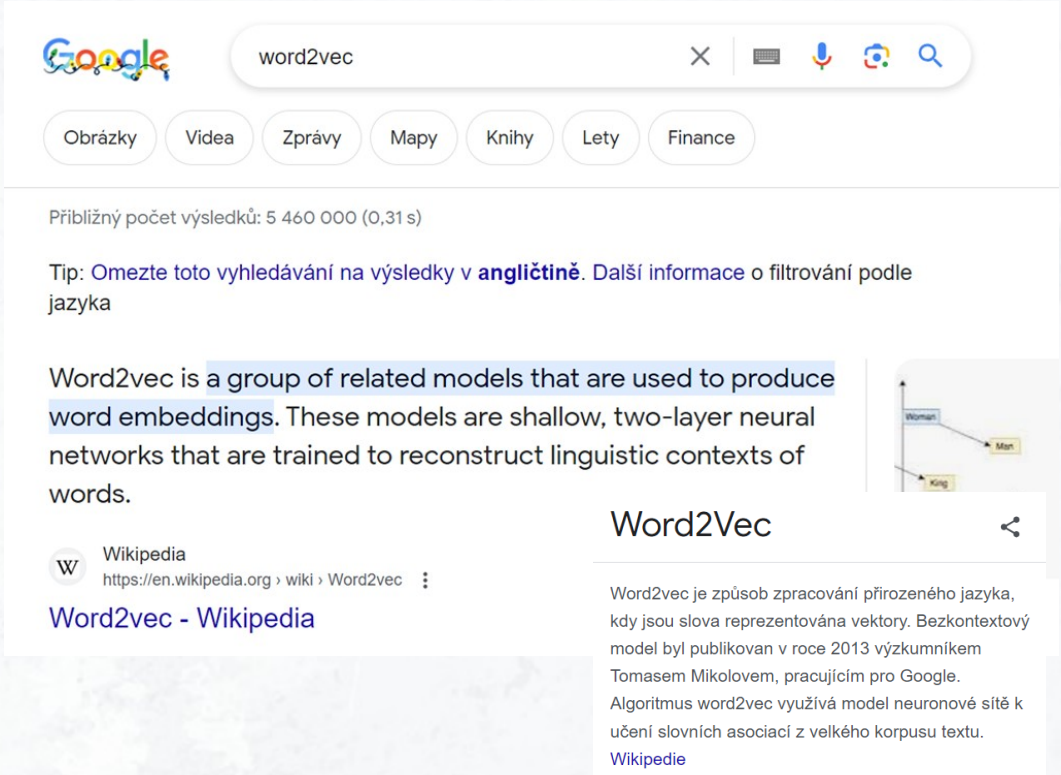
- NLP = natural language processing -> zpracování jazyka (mix informatiky, matematiky a lingvistiky)
- Cílem je počítač schopný "porozumět" obsahu dokumentů, včetně kontextových nuancí jazyka v nich obsažených
    - Pracujeme s tzv. korpusy textu
    - Může nás zajímat v podstatě cokoliv:
        - Sentiment textu - vyznívá to pozitivně nebo negativně?
        - Jak o věci XY mluví lidi na sociálních sítích?
        - Kolikrát se objevuje zmínka o XY v konverzacích?
        - Jazykovědné aspekty textu
        - A další...

Přehled rovin zkoumání přirozeného jazyka

	POPIS	PŘÍKLAD
LEXIKÁLNÍ ROVINA	slovní zásoba a její proměny	Co je to *step*? Druh krajiny? Tanec? Anglicky *krok*?
MORFOLOGICKÁ ROVINA	slovní tvary a způsob tvoření nových slov	Jak skloňovat slovo *step*? Jak skloňovat slovo *lalulá*?
SYNTAKTICKÁ ROVINA	shlukování slov do větších celků (frází), organizace frází ve větě	Je nějaký rozdíl mezi frázemi *hodný syn nehodného otce* a *nehodný syn hodného otce*?
LOGICKÁ ROVINA	reprezentace vět pomocí logických formulí	Když je Petra Pavlovou sestrou, je Pavel Petřiným bratrem? Když je Petra Pavlovou životní láskou, je Pavel Petřinou životní láskou?
SÉMANTICKÁ ROVINA	význam celku a význam jednotlivých částí	Co znamená *pět na stole v českých*? Kdo zemřel v tomto novinovém titulku: *Žárlivost dovedla ženu až k vraždě. Její druh útok nožem nepřežil*?
PRAGMATICKÁ ROVINA	význam promluvy v kontextu	Co se děje, když někdo řekne *Můžete mi prosím podat sůl?*

Zdroj: NLP lab FI - Centrum zpracování přirozeného jazyka

K čemu to je?



Tomáš Mikolov, držitel Ceny Neuron, autor word2vec

Jaký je proces trénování strojového učení?

Case study: IRTIS WP4

1. Nasbíráme data
    a. Messengery a WhatsApp participantů, anonymizační software
2. Výzkumná otázka a cíl
    a. Rizika a social support
3. Rozdělení korpusu na části: anotační/učící a trénovací část
4. Anotace: anotační schéma, trénink anotátorů, shoda, anotace části korpusu
5. Vytvoření tzv. Gold standard

1. Trénování strojového učení: lingvistický pre-processing, klasifikační úloha
2. Aplikace na neanotovaných datech

...A odteď jde o iterativní proces, který často opakujeme tak dlouho, dokud nejsme spokojení s výsledky. Příklad: implementace kontextu pro vylepšení klasifikace.

Ukázka z anotačního manuálu

3. (17) Alkohol a drogy (včetně alkoholu, nikotinu a dalších drog)

Je obsahem řádku něco z následujícího:

| popis zkušeností s drogami (cigarety, nikotin, marihuana, vodní dýmka, …) | referování o drogach/o tom, že je někdo pod jejich vlivem | domluva na konzumaci | shánění drog/poptávání drog | podporování/ospravedlnění/odůvodňování užívání drog | vyjádření přání/záměru konzumace | popis zkušeností s alkoholem/následky použití alkoholu | domluva na konzumaci alkoholu | vyjádření přání/záměru konzumace alkoholu | shánění alkoholu | odůvodňování/normalizace konzumace alkoholu |

- POZOR: Za drogy považujeme také: léky zneužívané jako drogy | tabákové výrobky

Příklad: | **popis zkušeností s drogami** | „Kamo ja su speceny jak svine□ □ “

Příklad: | s**hánění drog/poptávání drog** | „Sežeň mi pár gramů trávy“

Příklad: | **vyjádření přání/záměru konzumace alkoholu** | “Kartusková bez chlastu není vono xd“
Příklad: | **domluva na konzumaci alkoholu** | “Budou stačit 3 frisca?□ ”

Příklad: | **shánění alkoholu** | „Nedovezl bys mi nějaký pití na zítra?“

Conversation: 2024 - 69 - 210 (36/4065)

0: NO TAG (0) | 1: Informační podpora (1) | 2: Emocionální podpora (2) | 3: Začlenění do skupiny (3) | 4: Uznání (4) | 5: Nabídka pomoci (5) | 21: Cizí jazyk (6)

#	Line	Ann 1	Ann 2	Tag
9	09.10.2019 19:04:47 - Kateřina Lendrová: Už meli laborky			
10	09.10.2019 19:05:39 - Pavel Ashby: říkala písemku na příklady	1	1	
11	09.10.2019 19:06:23 - Darja Poljansky <OWNER>: Hej podle mě fakt ne ale jak chcete	1	1	
12	09.10.2019 19:06:37 - Pavel Ashby: tak ono je to easyy	2		2
13	09.10.2019 19:22:07 - Darja Poljansky <OWNER>: Hej ten případ se sanema			
14	09.10.2019 19:22:13 - Darja Poljansky <OWNER>: U té práce			
15	09.10.2019 19:22:19 - Darja Poljansky <OWNER>: To kdy nastava			
16	09.10.2019 19:22:50 - Darja Poljansky <OWNER>: ?			
17	09.10.2019 19:24:32 - Pavel Ashby: kdyz jedes na sanich?			
18	09.10.2019 19:24:46 - Darja Poljansky <OWNER>: No ale jaký pohyb			
19	09.10.2019 19:24:51 - Darja Poljansky <OWNER>: Šikmo dolů?			
20	09.10.2019 19:25:23 - Pavel Ashby: wtfffff			
21	09.10.2019 19:26:41 - Darja Poljansky <OWNER>: Se uklidni			
22	09.10.2019 19:26:49 - Darja Poljansky <OWNER>: Jen se snažím něco zjistit			
23	09.10.2019 19:26:58 - Pavel Ashby: vsak jaa jsem v klidu			
24	09.10.2019 19:27:04 - Pavel Ashby: jen nechapu jaky pohyb			
25	09.10.2019 19:27:07 - Pavel Ashby: dolu			

Previous | Reset | Following | Go back | Stop Annotation | Finish burst | Save this conversation and load next one

50 100 150 200 250 300 350 400 450 500 550 600 650 700 750 800 850 900 950 1000 1050 1100 1150 1200 1250 1300 1350 1400 1450 1500
1550 1600 1650 1700 1750 1800 1850 1900 1950 2000 2050 2100 2150 2200 2250 2300 2350 2400 2450 2500 2550 2600 2650 2700 2750 2800 2850 2900 2950 3000
3050 3100 3150 3200 3250 3300 3350 3400 3450 3500 3550 3600 3650 3700 3750 3800 3850 3900 3950 4000 4050

Enlarge | Hide

ANOTAČNÍ MANUÁL: PODPŮRNÉ INTERAKCE

Verze 2. 23. 11, zkrácená verze

CO JE TO PODPORA (PODPŮRNÉ INTERAKCE ONLINE)?

Podpůrné interakce online mezi vrstevníky, my se zaměříme konkrétně na **poskytování podpory** ve smyslu komunikace, o které můžeme předpokládat, že u příjemce vyvolá pocit/přesvědčení, že někomu na něm záleží, je milován, vážen a oceňován, že někdo má o něj starost, je pro něj oporu, že mu někdo poskytne radu nebo užitečnou informaci, bude-li potřeba, pomůže mu, a také, že je součástí skupiny, s jejíž členy tráví čas a má společné plány.

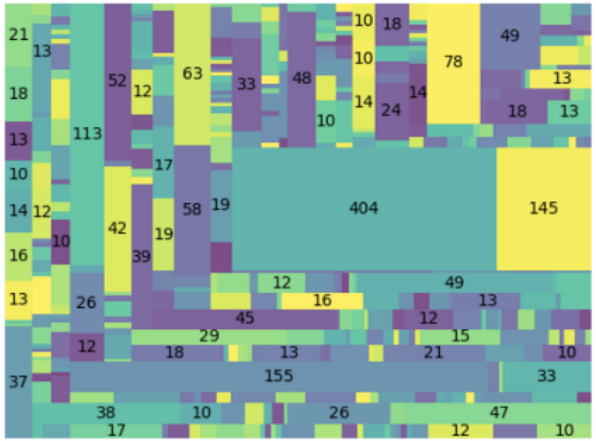
Pro každý typ podpory platí, že pokud výpověď v daném řádku nemá tento účel, neměla by být označená jako podpora.

1. Informační podpora

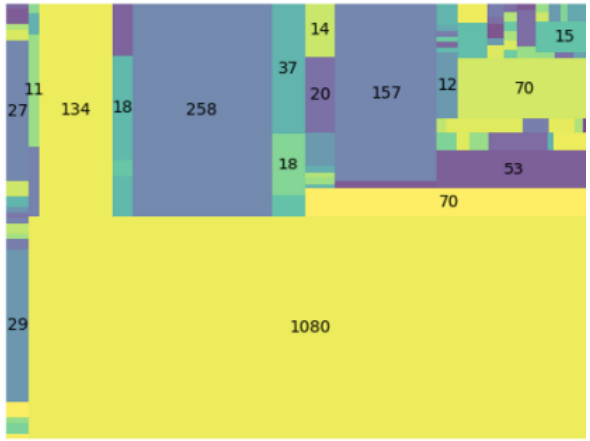
Je obsahem řádku něco z následujícího:

| dávání rad/tipů | učení | zpětné vazby | předávání znalostí/informací/zkušeností, které druhý potřebuje a jsou mu nápomocné v řešení problému/pochopení situace, v které se ocitl|

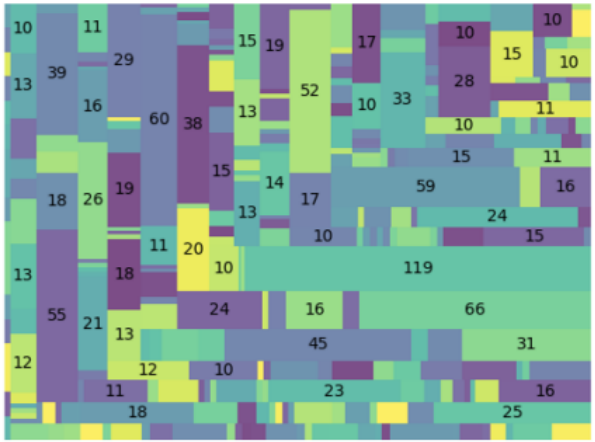
- Z kontextu konverzace musí být zřejmé, že informace jsou pro příjemce PODPORU v nějakém konkrétním jednání/akci/situaci | jsou nápomocné v řešení problému



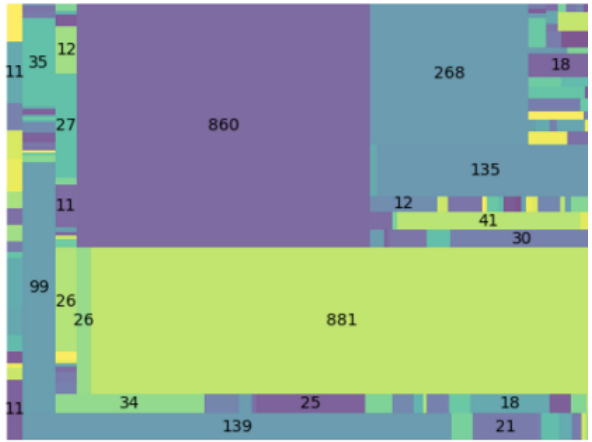
a) Aggression, harassment, hate



b) Mental health problems



c) Alcohol, drugs



d) Sexual content

- Data od 22 uživatelů (13-17 let) a všech jejich chatových partnerů: celkem 2165 osob
- Celkový počet konverzací: 90,422 (konverzace končí, když hodinu nikdo nic nenapíše)
- “Utterances” = 1,260,492

Plhák et al., 2023

Plhák et al., 2023

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

Jak se určuje efektivita modelů?

- Precision = kolik mám falešně pozitivních jevů? (email, co není spam je označen jako spam)
- Recall (sensitivity) = kolik mám falešně negativních jevů? (email, co je spam byl označen, jako že není spam)
- F1 skóre = funkce pro výpočet poměru mezi P a R
- Někdy dává větší smysl koukat jen na jednu z těch metrik

		Predicted	
		Negative	**Positive**
Actual	**Negative**	True Negative	False Positive
	Positive	False Negative	True Positive

Další četba

Classification of Adolescents' Risky Behavior in Instant Messaging Conversations
PLHÁK, Jaromír, Ondřej SOTOLÁŘ, Michaela LEBEDÍKOVÁ a David ŠMAHEL. Classification of Adolescents' Risky Behavior in Instant Messaging Conversations. In Ruiz, Francisco and Dy, Jennifer and van de Meent, Jan-Willem. *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*. 206. vyd. https://proceedings.mlr.press: PMLR, 2023. s. 2390-2404. ISSN 2640-3498.

Constructing Datasets from Dialogue Data
SOTOLÁŘ, Ondřej, Jaromír PLHÁK, Michal TKACZYK, Michaela LEBEDÍKOVÁ a David ŠMAHEL. Constructing Datasets from Dialogue Data. In Horák, Aleš and Rychlý, Pavel and Rambousek, Adam. *Proceedings of the 16th Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2022*. Brno: Tribun EU, 2022. s. 131-139. ISBN 978-80-263-1752-4.

Detecting Online Risks and Supportive Interaction in Instant Messenger Conversations using Czech Transformers
SOTOLÁŘ, Ondřej, Jaromír PLHÁK, Michal TKACZYK, Michaela LEBEDÍKOVÁ a David ŠMAHEL. Detecting Online Risks and Supportive Interaction in Instant Messenger Conversations using Czech Transformers. In Horák, Rychlý, Rambousek. *Recent Advances in Slavonic Natural Language Processing (RASLAN 2021)*. Brno: Tribun EU, 2021. s. 19-28. ISBN 978-80-263-1670-1.

Děkujeme za pozornost