

Využití umělé inteligence a velkých jazykových modelů pro transkripci Terms of Use

Matyáš Flekač

Bakalářská práce
2025



Univerzita Tomáše Bati ve Zlíně
Fakulta aplikované informatiky

Univerzita Tomáše Bati ve Zlíně
Fakulta aplikované informatiky
Ústav informatiky a umělé inteligence

Akademický rok: 2024/2025

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

(projektu, uměleckého díla, uměleckého výkonu)

Jméno a příjmení: Matyáš Flekač
Osobní číslo: A22472
Studijní program: B0613A140020 Softwarové inženýrství
Forma studia: Prezenční
Téma práce: Využití umělé inteligence a velkých jazykových modelů pro transkripci Terms of Use
Téma práce anglicky: Using Artificial Intelligence and Large Language Models for Transcription of Terms of Use

Zásady pro vypracování

- Vypracujte literární rešerši k dané problematice.
- Navrhněte vhodný formát dat pro zpracování.
- Zpracujte data z webové stránky tosdtr.org.
- Navrhněte vlastní řešení zpracování dat pomocí ChatGPT API.
- Zpracujte možnosti využití pro další webové stránky a podmínky užití.
- Vhodně prezentujte výsledky práce.

Forma zpracování bakalářské práce: **tištěná/elektronická**

Seznam doporučené literatury:

1. PECINOVSKÝ, Rudolf. Python: kompletní příručka jazyka pro verzi 3.11. Grada. Knihovna programátora (Grada). Praha: Grada Publishing, 2023. ISBN 978-80-271-3891-3.
2. Jednoduše. Praha : Euromedia Group. Universum (Euromedia Group). Praha: Euromedia Group, 2023. ISBN 978-80-242-9293-9.
3. Manning, Christopher D., Raghavan, Prabhakar, Schütze, Hinrich. Introduction to Information Retrieval. Cambridge: Cambridge University Press, 2008. 482 p. ISBN 978-0521865715.
4. Jurafsky, Daniel, Martin, James H. Speech and Language Processing. 3rd ed. London: Pearson, 2023. 1024 p. ISBN 978-0131873216.

Vedoucí bakalářské práce: **Ing. Petr Žáček, Ph.D.**
Ústav informatiky a umělé inteligence

Datum zadání bakalářské práce: **27. října 2024**

Termín odevzdání bakalářské práce: **2. června 2025**



L.S.

doc. Ing. Jiří Vojtěšek, Ph.D. v.r.
děkan

prof. Mgr. Roman Jašek, Ph.D., DBA v.r.
ředitel ústavu

Ve Zlíně dne 17. listopadu 2024

PROHLÁŠENÍ AUTORA BAKALÁŘSKÉ PRÁCE

Beru na vědomí, že

- odevzdáním bakalářské práce souhlasím se zveřejněním své práce podle zákona č. 111/1998 Sb., v platném znění bez ohledu na výsledek obhajoby;
- bakalářská práce bude uložena v elektronické podobě v univerzitním informačním systému a bude dostupná k nahlédnutí;
- jedno vyhotovení bakalářské práce v listinné podobě bude ponecháno Univerzitě Tomáše Bati ve Zlíně k uložení;
- na moji bakalářskou práci se plně vztahuje zákon č. 121/2000 Sb. o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon) ve znění pozdějších právních předpisů, zejm. § 35 odst. 3;
- podle § 60 odst. 1 autorského zákona má Univerzita Tomáše Bati ve Zlíně právo na uzavření licenční smlouvy o užití školního díla v rozsahu § 12 odst. 4 autorského zákona;
- podle § 60 odst. 2 a 3 mohu užít své dílo – bakalářskou práci – nebo poskytnout licenci k jejímu využití jen s předchozím písemným souhlasem Univerzity Tomáše Bati ve Zlíně, která je oprávněna v takovém případě ode mne požadovat přiměřený příspěvek na úhradu nákladů, které byly Univerzitou Tomáše Bati ve Zlíně na vytvoření díla vynaloženy (až do jejich skutečné výše);
- pokud bylo k vypracování bakalářské práce využito softwaru poskytnutého Univerzitou Tomáše Bati ve Zlíně nebo jinými subjekty pouze ke studijním a výzkumným účelům (tj. k nekomerčnímu využití), nelze výsledky bakalářské práce využít ke komerčním účelům;
- pokud je výstupem bakalářské práce jakýkoliv softwarový produkt, považují se za součást práce rovněž i zdrojové kódy, popř. soubory, ze kterých se projekt skládá; neodevzdání této součásti může být důvodem k neobhájení práce.

Prohlašuji, že

- jsem na bakalářské práci pracoval(a) samostatně a použitou literaturu jsem řádně citoval(a); v případě publikace výsledků budu uveden(a) jako spoluautor;
- odevzdaná verze bakalářské práce a verze elektronická nahraná do IS/STAG jsou obsahově totožné.

Ve Zlíně, dne

.....

podpis autora

ABSTRAKT

Tato bakalářská práce se zabývá využitím velkých jazykových modelů, konkrétně modelu ChatGPT, pro automatizovanou analýzu podmínek užívání online služeb. Cílem bylo navrhnout vhodný datový formát, který umožní efektivní zpracování smluvních dokumentů, a vytvořit aplikaci, která tato data analyzuje. Řešení kombinuje veřejně dostupná data z projektu ToS;DR a vlastní vstupy od uživatele. V rámci testování byly porovnány různé typy promptů a výstupy modelu byly ověřeny vůči oficiálním hodnocením. Výsledky ukazují, že model dokáže generovat přehledné shrnutí, identifikovat riziková ustanovení a nabídnout doporučení k úpravě textu. Aplikace je navržena tak, aby byla srozumitelná pro běžné uživatele a zároveň využitelná i v odborném prostředí.

Klíčová slova: podmínky užití, velké jazykové modely, ChatGPT, ToS;DR, analýza textu

ABSTRACT

This bachelor's thesis focuses on the use of large language models, specifically ChatGPT, for the automated analysis of terms of service of online platforms. The main goal was to design an appropriate data format for processing legal documents and to implement an application capable of analyzing such data. The solution combines public data from the ToS;DR project with custom user input. During testing, different prompt types were compared, and model outputs were validated against official ToS;DR summaries. The results show that the model can produce clear summaries, identify problematic clauses, and offer suggestions for improvement. The application is designed to be accessible to everyday users while also being suitable for legal or academic use.

Keywords: terms of use, large language models, ChatGPT, ToS;DR, text analysis

Rád bych poděkoval vedoucímu této práce Ing. Petru Žáčkovi, Ph.D. za jeho ochotu, vstřícnost a cenné rady, které mi během psaní této práce velmi pomohly a také za jeho čas, který mi věnoval během našich schůzek.

V rámci této bakalářské práce byla využita umělá inteligence jako nástroj pro analýzu a zpracování textů. Konkrétně byla nasazena v projektu, který se zaměřil na porozumění a interpretaci smluvních podmínek služeb s využitím jazykového modelu GPT.

OBSAH

ÚVOD.....	13
I TEORETICKÁ ČÁST	14
1 UMĚLÁ INTELIGENCE.....	15
1.1 HISTORIE A VÝVOJ AI.....	15
1.1.1 Turingův test	16
1.2 STROJOVÉ UČENÍ	17
1.2.1 Neuronové sítě	17
1.2.2 Hluboké učení	19
1.3 SUPERVIZOVANÉ A NESUPERVIZOVANÉ UČENÍ	19
1.4 ZPĚTNOVAZEBNÍ UČENÍ.....	20
2 ZPRACOVÁNÍ PŘIROZENÉHO JAZYKA (NLP).....	21
2.1 PŘEDZPRACOVÁNÍ TEXTU V NLP	21
3 GENERATIVNÍ UMĚLÁ INTELIGENCE	22
3.1 ZÁKLADY UČENÍ GENERATIVNÍ UMĚLÉ INTELIGENCE	22
3.2 REŽIMY UMĚLÉ INTELIGENCE	22
3.2.1 Režim učení (learning).....	22
3.2.2 Režim provozu (inference).....	23
4 VELKÉ JAZYKOVÉ MODELÝ	24
4.1 VÝVOJ A UČENÍ VELKÝCH JAZYKOVÝCH MODELŮ	24
4.2 PŘÍPADY UŽITÍ VELKÝCH JAZYKOVÝCH MODELŮ	26
4.2.1 Generování textu	26
4.2.2 Virtuální asistenti	26
4.2.3 Generování kódu	26
4.2.4 Analýza sentimentu	26
5 CHATGPT	27
5.1 VÝHODY CHATGPT	28
5.2 NEVÝHODY CHATGPT	28
6 VYUŽITÍ UMĚLÉ INTELIGENCE PRO TRANSKRIPCI.....	30
6.1 PRINCIPY FUNGOVÁNÍ TRANSKRIPCE	30
6.2 VYUŽITÍ AUTOMATICKÉ TRANSKRIPCE	30
6.3 VÝHODY AUTOMATICKÉ TRANSKRIPCE	31
7 PROBLEMATIKA PODMÍNEK UŽITÍ (TERMS OF USE).....	32
7.1 TYPY TERMS OF USE	32
7.1.1 Browsewrap.....	32
7.1.2 Clickwrap	32

8	STRÁNKY S PROBLEMATIKOU TERMS OF USE	34
8.1	PRIVACYSPY	34
8.2	ToS;DR (TERMS OF SERVICE; DIDN'T READ).....	35
8.3	TOSDR API.....	36
8.3.1	Service API.....	36
8.3.2	Case API.....	36
8.3.3	Document API.....	37
8.3.4	Search API.....	37
8.3.5	Privacy Shields API.....	37
8.3.6	Error Codes.....	37
8.4	OCHRANA OSOBNÍCH ÚDAJŮ A GDPR	37
9	METODY SBĚRU DAT	39
9.1	WEBCRAPING.....	39
9.1.1	BeautifulSoup.....	39
9.2	API.....	40
9.2.1	API vs web scraping.....	40
10	PYTHON	41
II	PRAKTICKÁ ČÁST	42
11	PRŮZKUM WEBOVÉ STRÁNKY TOSDR.ORG	43
11.1	ÚČEL TOSDR.ORG	43
11.2	JAK TO FUNGUJE	43
11.3	OBSAH STRÁNEK	43
11.4	VYUŽITÍ PROJEKTU TOSDR.ORG	44
12	NÁVRH FORMÁTU DAT PRO ZPRACOVÁNÍ.....	45
12.1	POŽADAVKY NA VSTUPNÍ DATA.....	45
12.2	PŘÍSTUP K NÁVRHU FORMÁTU	45
12.3	VYUŽITELNOST NAVRŽENÉHO FORMÁTU	46
13	EXTRAKCE A PŘEDZPRACOVÁNÍ DAT Z TOSDR.ORG.....	47
13.1	PŘÍSTUP K API.....	47
13.2	ZÍSKÁVÁNÍ A VÝBĚR ÚDAJŮ.....	47
13.3	VYTVÁŘENÍ VSTUPU PRO JAZYKOVÝ MODEL	48
13.4	AUTOMATIZACE, KONTROLA CHYB A STABILITA	49
14	POROVNÁNÍ A ODLIŠNOSTI NAVRHOVANÉHO ŘEŠENÍ	51
15	IMPLEMENTACE ŘEŠENÍ POMOCÍ CHATGPT API	53
15.1	UŽIVATELSKÉ ROZHRANÍ A OVLÁDÁNÍ APLIKACE.....	55
16	ROZŠÍŘENÍ NA DALŠÍ WEBOVÉ STRÁNKY A PODMÍNKY UŽITÍ	57
16.1	MOŽNOST RUČNÍHO VKLÁDÁNÍ PODMÍNEK	57

16.2	MOŽNOST AUTOMATIZOVANÉHO NAČÍTÁNÍ Z JINÝCH ZDROJŮ	58
16.3	PODMÍNKY MIMO OBLAST TECHNOLOGIÍ	58
16.4	VYUŽITÍ V PRÁVNÍ A VZDĚLÁVACÍ PRAXI	58
16.5	BUDOUCÍ SMĚRY ROZŠÍŘENÍ	59
17	PREZENTACE VÝSLEDKŮ A VYHODNOCENÍ ŘEŠENÍ.....	60
17.1	POUŽITÍ VÝCHOZÍHO PROMPTU	60
17.2	VÝSTUPY ZALOŽENÉ NA VÝCHOZÍM PROMPTU	60
17.2.1	Github.....	60
17.2.2	Facebook	61
17.2.3	Youtube	62
17.3	ALTERNATIVNÍ PROMPT A JEHO VLIV NA VÝSTUP	63
17.4	POZOROVANÉ ROZDÍLY MEZI PROMPTY	64
17.4.1	Github.....	64
17.4.2	Facebook	65
17.4.3	Youtube	66
17.5	ANALÝZA VLASTNÍHO DOKUMENTU A PŘÍZPŮSOBNÍ VÝSTUPU	67
17.6	ANALÝZA VLASTNÍCH ToS POMOCÍ KREATIVNÍHO PROMPTU	69
17.7	POROVNÁNÍ VÝSTUPŮ S OFICIÁLNÍM SHRNU TÍM Z ToS;DR	70
17.8	POSTUP POROVNÁNÍ	70
17.8.1	Github.....	71
17.8.2	Facebook	72
17.8.3	Youtube	72
17.9	ČASOVÁ PROMĚNLIVOST VÝSTUPŮ	73
17.9.1	Výsledky analýzy	73
17.10	SHRNU TÍ VÝSLEDKŮ	75
18	SPUŠTĚNÍ APLIKACE	78
	ZÁVĚR	79
	SEZNAM POUŽITÉ LITERATURY.....	81
	SEZNAM OBRÁZKŮ	84
	SEZNAM TABULEK.....	85
	SEZNAM POUŽITÝCH SYMBOLŮ A ZKRATEK	86
	SEZNAM PŘÍLOH.....	87

ÚVOD

V digitální době, kdy většina uživatelů souhlasí s podmínkami užívání bez jejich čtení, představují Terms of Service (ToS) významný právní dokument s potenciálně zásadním dopadem na soukromí a práva jednotlivce. Projekt Terms of Service; Didn't Read ukazuje, že smluvní dokumenty často obsahují problematická ustanovení, která mohou být pro běžné uživatele obtížně srozumitelná. Současně se objevují nové možnosti využití umělé inteligence, především velkých jazykových modelů (LLM), pro zjednodušení a interpretaci právních textů.

Tato bakalářská práce si kladla za cíl prozkoumat možnosti využití jazykového modelu ChatGPT k automatizované analýze podmínek užívání. Cílem bylo nejen navrhnout datový formát vhodný pro vstup do modelu, ale také implementovat plně funkční aplikaci, která kombinuje data z veřejných zdrojů (např. ToS;DR) i vlastních dokumentů uživatele. Součástí řešení je také porovnání výstupů modelu s oficiálními hodnoceními a zhodnocení jeho použitelnosti pro běžné i odborné účely.

Teoretická část práce se věnuje především principům umělé inteligence, fungování velkých jazykových modelů a jejich využití v oblasti zpracování přirozeného jazyka. Součástí je také přehled existujících nástrojů pro analýzu podmínek použití, jako jsou projekty ToS;DR nebo PrivacySpy, a diskuse jejich výhod a omezení.

Praktická část navazuje návrhem a realizací vlastního nástroje pro sumarizaci Terms of Use pomocí GPT modelu. Čtenář se v ní seznámí s návrhem datového formátu, s implementací uživatelského rozhraní i s postupem vyhodnocení výstupů včetně porovnání s externími službami a reálnými příklady. Výsledkem je nástroj, který umožňuje přístupnější porozumění uživatelským podmínkám běžným uživatelům.

I TEORETICKÁ ČÁST

1 UMĚLÁ INTELIGENCE

Umělá inteligence je inteligence, která zajišťuje, že počítače a stroje dokážou simulovat lidské učení, řešení problémů, kreativitu a rozhodování. Je nedílnou součástí počítačových věd a v současné době je nejvíce rozvíjena. Přístroje se zabudovanou umělou inteligencí dokážou vidět a identifikovat objekty. Jsou schopny porozumět a odpovědět na lidský jazyk. Zvládají se učit z nových informací a zkušeností a také se sami rozhodovat bez potřeby lidské inteligence. Od roku 2024 se lidé nejvíce zaměřují na generativní umělou inteligenci. Abychom mohli pochopit generativní AI (Artificial Intelligence), musíme prvně pochopit technologie, na kterých generativní AI stojí. V kultuře je představována umělá inteligence jako konkurent lidské inteligence, někdy je zobrazována i jako existenční hrozba. Aplikace založené na umělé inteligenci vládnu ve skutečnosti jen omezenými schopnostmi. Člověk umělou inteligenci využívá každý den, aniž by si to přímo uvědomoval. [1; 2; 3]

1.1 Historie a vývoj AI

Myšlenka umělé inteligence se datuje tisíce let do minulosti až k antickým filozofům. Již v mytologii starověkého Řecka a Číny se objevil pojem „živý stroj“, avšak tuto myšlenku lidé rozvinuli až v 18. století, kdy sestrojili stroje poháněné vlastní silou neboli „automaty“. Umělá inteligence jako vědní obor je velmi mladá a nemůžeme přesně určit datum jejího vzniku, předpokládá se ale, že to bylo v 60. letech 20. století, kdy vzniklo mnoho médií na myšlenku umělého člověka. Vědci si poté začali klást otázky, zda je možné vytvořit náhradu lidského mozku, tzv. umělý mozek. [2; 3; 4]

Zájem o umělou inteligenci se nejvíce zintenzivnil v letech 1950-1956. V roce 1950 publikoval Alan Turing svou práci „Computer Machinery and Intelligence“, z níž vzniknul Turingův test, který se používá k měření počítačové inteligence. V roce 1952 byl vyvinut program na hraní dámy, který se jako první v historii naučil hru samostatně. Roku 1955 byl uspořádán seminář o „umělé inteligenci“, který vedl John McCarthy, což bylo považováno za první použití tohoto sousloví. [3; 4]

V letech 1957–1979 došlo k dozrávání umělé inteligence. V 70. letech byl v Japonsku sestrojen první robot, nebo také první autonomní vozidlo. [4]

Léta 1980–1987 označujeme jako rozmach umělé inteligence, ten byl způsoben průlomovými objevy ve výzkumu, ale také díky finančním prostředkům vlády. Stále

populárnější se stávala technika hlubokého učení a používání expertního systému, díky kterému se počítače mohli učit z vlastních chyb a dělat samostatná rozhodnutí. [4]

V dnešní době byl zaznamenán velký nárůst AI nástrojů pro běžné použití jako například virtuální asistenti, vyhledávače atd. V roce 2012 byla neuronová síť vycvičena k rozpoznávání koček na obrázku, aniž by měla nějaké informace o pozadí. V roce 2015 byla podepsána dohoda o zákazu používání autonomních zbraní pro válečné účely. V roce 2016 byl vytvořen první robotický občan, který měl realistický lidský vzhled a schopnost replikovat emoce a komunikovat. V roce 2017 vytvořil Facebook dva chatboty s umělou inteligencí, aby spolu komunikovali a vyjednávali, tyto chatboti se nakonec vzdali angličtiny a vytvořili si svůj vlastní jazyk pro komunikaci. V roce 2020 bylo zahájeno beta testování modelu GPT-3, který využívá učení k vytváření kódu, poezie a jiných jazykových a písemných úloh. Je to první model, který vytváří obsah a je téměř k nerozeznání od obsahu vytvořeného člověkem. [4]

1.1.1 Turingův test

Jedná se o metodu, která určuje, zda stroj dokáže vykazovat lidskou inteligenci. Pokud se stroj dokáže zapojit do konverzace s člověkem bez toho, aniž by byl detekován jako stroj, je považováno, že prokázal lidskou inteligenci. Alan Turing vynalezl tento test za účelem určení, zda dokáže být počítač tak chytrý, aby byl zaměněn za člověka. [5]

Zkoušky se účastní tři účastníci, člověk jako tazatel, člověk jako odpovídající a stroj. Tazatel ani jednoho z nich nevidí a komunikuje s nimi pouze pomocí textových zpráv. Klade otázky jak člověku, tak stroji. Stroj se snaží odpovídat jako člověk a skutečný člověk odpovídá přirozeně. Pokud tazatel nedokáže rozlišit, kdo je člověk a kdo stroj, můžeme říct, že stroj Turingovým testem prošel, pokud je stroj snadno rozpoznatelný, tak testem neprošel. [5]

V roce 1966 vznikli první počítače, které dokázaly oklamat lidi. V roce 1966 vznikl počítač ELIZA, který vytvářel věty ze specifických slov. Byl to jeden z prvních strojů, který byl schopný oklamat lidi. [5]

Později byl vynalezen chatbot PARRY, který imitoval paranoidního schizofrenika. Skupina psychiatrů měla za úkol detekovat a analyzovat konverzaci s reálným člověkem a tímto chatbotem. Ve 48 % dokázali identifikovat, že se jedná o stroj. Ani jeden z těchto počítačů Turingovým testem neprošel. [5]

V roce 2013 vznikl chatbot, který vykazoval inteligenci 13letého kluka. Ve 33 % porotci dokázali identifikovat, že se jedná o stroj. Pro úspěšné splnění testu musí být úspěšnost identifikace maximálně 30 %. [5]

V roce 2024 prošel testem poprvé v historii ChatGPT, odpovědi umělé inteligence byly náhodně porovnávány s odpověďmi studentů. Chování umělé inteligence a lidí bylo velmi podobné. Jelikož je tento model trénován na obsahu vytvořeném lidmi, tak užívá myšlenky a data od lidí, tudíž odpovědi vypadají lidsky. Kvůli tomu jsou vyvolávány pochybnosti, zda testem skutečně prošel. [5]

1.2 Strojové učení

Jednou ze základních vlastností umělé inteligence je schopnost učit se. „*Strojové učení je druh umělé inteligence, který umožňuje počítačovým systémům učit se, jak provádět úkoly, aniž by k tomu musely být explicitně naprogramovány.*“ [2, s. 59] Toto učení využíváme u složitějších úkolů, kde programátoři nejsou schopni napsat potřebné algoritmy. Do strojového učení zařazujeme vytváření modelů trénováním algoritmu tak, aby byl schopný predikovat a dělat rozhodnutí na základě získaných dat. Ke strojovému učení se využívá velké množství dat, které mohou být označena štítky nebo neoznačena. Zahrnuje velkou škálu možných technik, které umožňují počítačům se učit a dělat závěry z dat, aniž by byly naprogramovány pouze pro konkrétní úkony. Existuje několik druhů technik strojového učení a algoritmů jako například lineární regrese, logistická regrese, SVMs, KNN. Každý z těchto přístupů je navržen na různé druhy problémů a dat. [1; 2; 6]

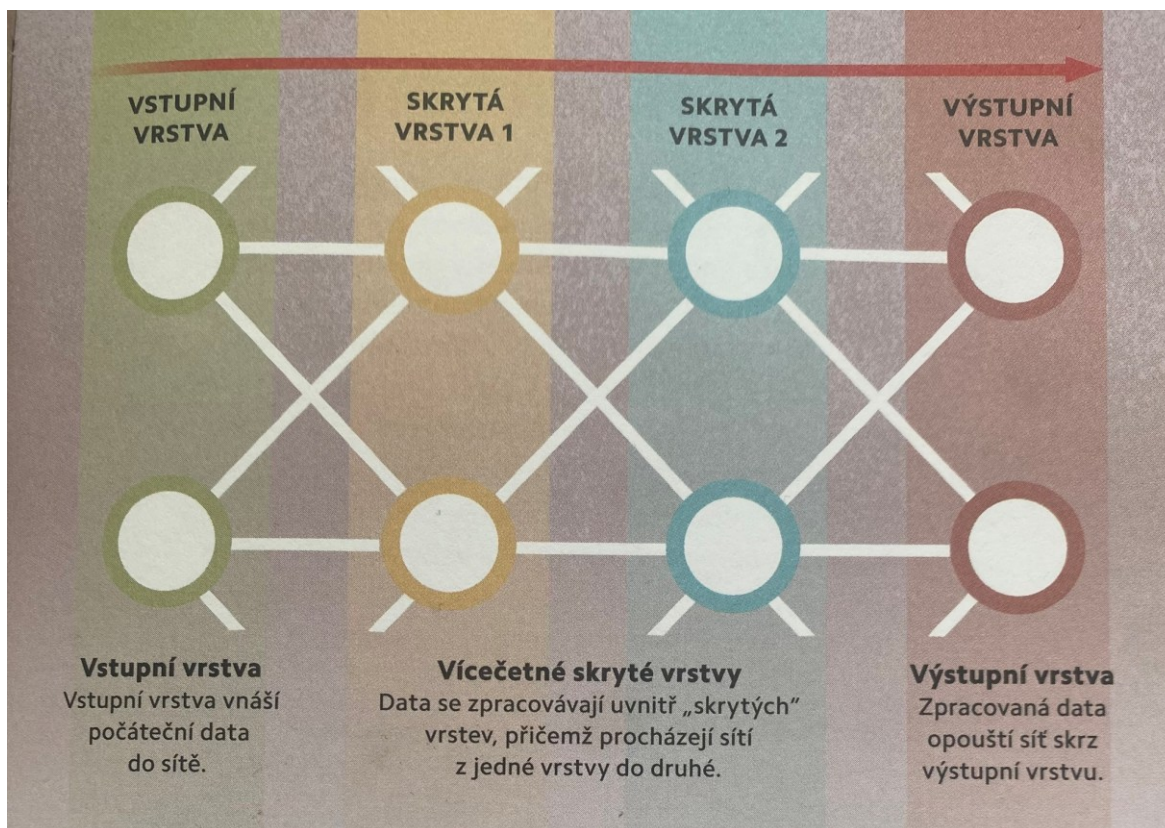
Jednou z nejlehčích forem strojového učení je učení pod dohledem, které zahrnuje použití datových sad k trénování algoritmu ke klasifikaci dat nebo přesnému předpovídání výsledků. V tomto druhu učení spárují lidé každý trénovací příklad s výstupem. Cílem je, aby se model naučil mapování mezi vstupy a výstupy na trénovacích datech, a aby toto mohl predikovat na nových nikdy neviděných datech. [1]

1.2.1 Neuronové síť

Jedním z nejpoužívanějších typů strojového učení je algoritmus zvaný neuronová síť. Neuronové síť jsou vymodelovány podle struktury a funkce lidského mozku. Tyto síť se skládají z propojených vrstev uzlů tzv. umělých neuronů, které jsou poskládány do vícečetných vrstev. Uzly ve vrstvách zpracovávají, přijímají a odesílají data do dalších vrstev v síti, dokud neexistuje nějaký výstup. Uzly pracují samostatně a každý může být

naprogramován tak, aby zacházel s daty, jakým způsobem chceme. K učení neuronové sítě se využívají cvičná data, tím se síť učí, jak generovat očekávané výsledky. [1; 2]

Data putují z uzlů v jedné vrstvě do uzlů ve vrstvě další. První vrstva vždy obsahuje přicházející data neboli vstup. Po této vrstvě následují vrstvy skryté, ve kterých probíhá zpracování. Data v těchto vrstvách nejsou viditelná pro uživatele tak jako na vstupu a výstupu, proto je nazýváme skryté. Zpracovaná výsledná data putují do výstupní vrstvy. [2]



Obrázek 1 Stavba neuronové sítě [2, s. 77]

Všechny umělé neuronové sítě mají stejnou základní strukturu, ale existují i složitější. Například hluboké neuronové sítě mohou mít několik skrytých vrstev, využívají se v hlubokém učení, rekurentní neuronové sítě zajišťují generování spojů mezi uzly v předchozích nebo následujících vrstvách. [2]

Generativní antagonistické sítě – využívají dvě konkurenční umělé neuronové sítě, které rozdělujeme na generátor a diskriminátor. Generátor využívá neštitkovaná data a vytváří z nich data falešná pro zmatení druhé neuronové sítě. Diskriminátor obdrží tyto data a snaží se poznat ty falešné. Pokud se mu to povede, generátor vygeneruje nová falešná data, která jsou složitější na rozlišení. Pokud neuspěje, pokouší se data rozpoznat účinnějším způsobem. [2]

Dopředné neuronové sítě – skládají se z jednoduché neuronové sítě, ve které informace putují pouze dopředu. Neobsahuje zpětnovazební smyčky a výstupy nejsou vráceny zpět v podobě vstupů. Obsahuje jediný umělý neuron, který se při užití gradientního sestupu dokáže strojově učit. [2]

Rekurentní neuronové sítě – data mohou putovat zpět za použití zpětnovazebních smyček. Využívají se ke zpracování dat, které mají specifické pořadí, například jazyk. Data jsou nedotčena, protože zachovávají strukturu sekvenčních dat a jejich závislosti. Využívají se k odhadnutí následujícího výstupu z datové sekvence a ke zpracování přirozeného jazyka. [2]

1.2.2 Hluboké učení

Hluboké učení představuje efektivní metodu strojového učení. Využívá vícevrstvé neuronové sítě nazývané hluboké neuronové sítě, které dokážou přesněji simulovat komplexní rozhodování lidského mozku. Hluboké neuronové sítě obsahují vstupní vrstvu, dále alespoň tři, ale většinou stovky skrytých vrstev a výstupní vrstvu. Na rozdíl od klasických strojových učících modelů, které obvykle obsahují pouze jednu nebo dvě skryté vrstvy. Podobně jako u všech neuronových sítí, data tečou z vrstvy vstupní do vrstev skrytých, zde jsou data zpracována uzly a předána do uzlů v další vrstvě. Díky velkému množství vrstev dokážou hluboké neuronové sítě zpracovávat data rychle a přesně. [1; 2]

Díky těmto vícenásobným vrstvám je možné učení bez dohledu. Může být automatizována extrakce funkcí z velkých neoznačených a nestrukturovaných dat a vytvářena jejich vlastní predikce, co tyto data představují. [1]

Jelikož hluboké učení nepotřebuje lidské zásahy, umožňuje strojové učení ve velkém měřítku. Hodí se pro zpracování přirozeného lidského jazyka, pro počítačové vidění a dalších úkolů, které zahrnují rychlou a přesnou identifikaci komplexních vzorců a vztahů ve velkém množství dat. [1]

1.3 Supervizované a nesupervizované učení

Supervizovaným učením označujeme typ strojového učení, kdy se umělá inteligence učí z dat označených datových setů. Člověk označí jak data vstupní, tak data výstupní štítky, díky čemuž se umělá inteligence naučí, jaký mají mezi sebou vztah. Po přidání nových vstupních dat se k předpovědi výstupů využívá funkce, která představuje souhrnný název pro vstupy, výstupy a pravidlo, které určuje jejich vztah. [2]

Nesupervizované učení se využívá v nešítkováných datových setech k odhalení skrytých struktur. [2]

Existuje i tzv. semi-supervizované učení, které stojí mezi učením supervizovaným a nesupervizovaným. Zde se využívají z části označené datové sety, což zapříčiňuje větší spolehlivost než učení nesupervizované. [2]

1.4 Zpětnovazební učení

V tomto typu učení provádí umělá inteligence úkoly metodou pokus-omyl. Umělá inteligence se naprogramuje tak, aby byla schopna rozpoznat odměny a tresty. Podle toho, zda úkol vyřeší správně nebo špatně, dostane zpětnou vazbu buď pozitivní nebo negativní. *„Umělá inteligence se naučí, že úspěch je dobrý a neúspěch špatný, a pokouší se úkol provádět opakovaně, dokud nedosáhne odměny“.* [2, s. 74]

2 ZPRACOVÁNÍ PŘIROZENÉHO JAZYKA (NLP)

Slouží ke komunikaci mezi člověkem a strojem. Jedná se o soubor výpočetních technik pro automatickou analýzu a reprezentaci lidských jazyků. Tyto metody jsou založeny na teoretických poznatcích. Počítačové programy využívající NLP zpracovávají texty v přirozeném jazyce a následně je interpretují do jazyka zpracovatelného počítačem. Počítačové programy využívají metodu zpracování přirozeného jazyka pro interpretaci lidského jazyka. Moderní modely zpracování přirozeného jazyka jsou postaveny na strojovém učení, přesněji na hlubokém učení, čerpají z lingvistiky a dokážou zpracovávat neštitkovaná data. „*Procesory zpracování přirozeného jazyka tvoří základ systémů určených pro překlad z jednoho jazyka do druhého.*“ [3, s. 54] Výzkum zpracování přirozeného jazyka pomohl k vývoji a nastartování éry generativní umělé inteligence a velkých jazykových modelů a je součástí našeho každodenního života. [3; 7; 8; 9; 10]

2.1 Předzpracování textu v NLP

Zpracování přirozeného jazyka obsahuje několik různých technik, které umožňují analýzu textu. Jako první se texty převedou na malá písmena z důvodu, že hodně počítačových programů rozlišuje velikosti písmen. [8; 9; 10]

Dalším krokem je tokenizace neboli rozdělení textu na jednotlivá slova nebo části slov. V některých textech jako například v angličtině, kde jsou texty rozděleny mezerami, je tokenizace jednodušší než v jiných jazycích, kde slova rozdělena mezerami nejsou. [8; 9; 10]

Jako další se provede stemming, což znamená, že se texty zkrátí na jejich základní tvar odstraněním koncovek. Hned poté proběhne lemmatizace, což je převod slov na jejich slovníkový tvar. [7; 8; 9; 10]

Nakonec se provede odstranění slov jako „je“, „pro“ a další, které dávají lidským větám větší smysl, ale pro analýzu významu nejsou zásadní. [8; 9; 10]

3 GENERATIVNÍ UMĚLÁ INTELIGENCE

Generativní umělá inteligence nám umožňuje vytvářet obsah jako je přirozený jazyk, obrázky, zvuky, videa nebo kód. Využívá pokročilé modely strojového učení. „*Generování obrázků založené na umělé inteligenci existuje již od šedesátých let minulého století a dokáže využívat celou škálu klasických i statických technik.*“ [2, s. 117] Výstup je založený na poskytnutých vstupech od uživatele. Nejběžnější způsob pro komunikaci s generativním AI je pomocí chatovacích aplikací, kde se jako vstup využívá přirozený jazyk. Aplikace, které používají pro vstup přirozený jazyk, využívají velké jazykové modely ke zpracování přirozeného jazyka. [2; 11]

3.1 Základy učení generativní umělé inteligence

V moderní AI se vytváří model na velmi výkonném hardware, tento model se cvičí a učí na velkém množství dat. Modelu je předkládán vstupní vzorec a následně se měří odchylka jeho výstupu od výstupu očekávaného. Tento proces je iterativní a opakuje se, dokud model nedosáhne požadované úrovně přesnosti. [11]

3.2 Režimy umělé inteligence

Obvykle se v AI modelu oddělují dva režimy, jedním z nich je režim učení (learning) a druhý režim provozu (inference). Režimy mají odlišné algoritmy fungování a nároky na výpočetní čas. Často se také provozují na rozdílném hardware. [11]

3.2.1 Režim učení (learning)

V tomto procesu se model učí z dostupných dat, aby získal schopnosti provádět požadované úkony. Tento režim probíhá v několika krocích. [12]

Prvním krokem je sběr dat. Tento krok je základem tréninku umělé inteligence. Kvalita dat ovlivňuje, jak bude náš model efektivní a přesný. Musíme být schopni získat a shromáždit relevantní data podle cílů projektu. Můžeme použít různé metody sběru dat, například webscaping z internetových stránek, pomocí API a další. Pro úspěch modelu musíme mít k dispozici kvalitní data. [12]

Druhým krokem je předzpracování a čištění dat. V tomto kroku čistíme sesbíraná data a odstraňujeme duplicity, nepotřebné informace a řešíme problém s chybějícími a poškozenými daty. Správně předzpracovaná data vedou k výkonnějšímu modelu, lepší přesnosti a lepším výsledkům. [12]

V dalších krocích musíme data označit, aby jim model porozuměl a mohl se z nich učit. Tento proces označování označuje nezpracovaná data štítky, díky kterým jsou strojově čitelná. Existují dva typy označování dat a těmi jsou ruční označování, kdy data označuje člověk. Tato metoda je spolehlivá, ale časově náročná. Druhým typem je označení automatizované, kdy si modely umělé inteligence data označují sami, pomocí naučených vzorů. V případě, že si model není jistý s označením dat, může data předat k ručnímu označení. [12]

V následujícím kroku jsou data rozšířena prováděním malých změn, čímž vytváříme data nová. Díky tomuto můžeme zvýšit kvalitu modelu, jelikož obsahuje více dat. [12]

Data rozdělíme na testovací a tréninkovou sadu. Pomocí tréninkové sady se model učí a pomocí testovací sady se hodnotí jeho odpovědi. Různé rozdělení dat může model ochránit před nadměrným přizpůsobením a pomoci mu zpracovávat data nová, se kterými nikdy nepracoval. [12]

Dále se musíme vyhnout zkreslení a musíme provést vyvážení dat, aby náš model neupřednostňoval určité odpovědi před jinými. Díky tomuto dosáhneme spolehlivějšího modelu umělé inteligence, který poskytuje objektivní odpovědi. Zmírnění kreslení dosahujeme několika možnými způsoby jako jsou podvzorkování, převzorkování a syntéza dat. [12]

V dalším kroku musíme provést ladění několika různých parametrů, mezi které řadíme například počet vrstev v neuronové síti a rychlost učení. Tento krok nazýváme ladění hyperparametrů. Snažíme se najít nejlepší možnou kombinaci těchto parametrů, aby byl model co nejvýkonnější. [12]

Po kompletním učení modelu je nutné zkontrolovat funkčnost. Hodnocení se provádí na základě toho, jak přesné odpovědi dokáže model generovat na základě nových dat. Po tomto kroku se model uvede do režimu provozu. [12]

3.2.2 Režim provozu (inference)

Po úspěšném trénování je model nasazen do produkčního prostředí, kde zpracovává nové, dosud neviděné vstupy a poskytuje predikce nebo rozhodnutí v reálném čase. V tomto režimu model aplikuje naučené vzory a vztahy z trénovacích dat na aktuální data, aby poskytl užitečné výstupy. [12; 13]

4 VELKÉ JAZYKOVÉ MODELY

Velké jazykové modely spadají mezi základní modely strojového učení a jsou trénovány na velkém množství dat. Mají schopnosti porozumět přirozenému jazyku a generovat jej. [14]

Tyto modely jsou klíčovým prvkem generativní umělé inteligence, která si v poslední době získala velkou popularitu. Může se zdát, že se velké jazykové modely objevily spolu s vývojem generativní AI, ale není tomu tak. Hodně společností s nimi pracuje už několik let a využívají je k rozvoji technologií zpracování přirozeného jazyka. Došlo k tomu společně s pokroky ve strojovém učení, neuronových sítí a modelech transformátorů, které poskytují strukturu pro tyto systémy. [14]

Velké jazykové modely jsou univerzální a lze je používat na široké spektrum aplikací. Představují zásadní pokrok v oblasti zpracování přirozeného jazyka a umělé inteligence a jsou dostupné veřejnosti prostřednictvím rozhraní jako například ChatGPT od firmy OpenAI, Llama od společnosti Meta. [14]

Tyto jazykové modely umožňují porozumění kontextu a generování přirozeného textu podobným způsobem, jako by ho napsal člověk. Jsou schopny překládat texty mezi jazyky, shrnovat dlouhé texty, odpovídat na otázky, asistovat při psaní anebo také generovat programovací kód. Díky velkému množství parametrů dokážou zachytit složité vzorce v jazyce a provádět hodně úloh souvisejících s jazykem. Jejich využití můžeme vidět například v chatbotech, virtuálních asistentech, pomoci při výzkumu a další. [14]

4.1 Vývoj a učení velkých jazykových modelů

Velké jazykové modely jako například GPT-4 fungují na odlišném principu než tradiční AI modely. Tyto modely využívají hlubokého učení, aby dokázaly analyzovat a generovat jazyk. Obvykle jsou postaveny na architektuře transformátorů, například na generativním předtrénovaném transformátoru (GPT), který pracuje efektivně se sekvenčními daty. [14]

Tyto modely obsahují velké množství vrstev neuronových sítí, přičemž každá z nich obsahuje mnoho parametrů, které jsou dovedeny k dokonalosti během trénování. [14]

Velké jazykové modely se učí bezkontextově neboli bez přímé interakce s konkrétními případy, díky tomu se učí gramatice, významu slov a vztahů mezi nimi. Modelu se předloží velké množství textů (v řádu miliard stránek), model tyto texty sekvenčně prochází a učí se z předchozích slov predikovat ty následující. Na první pohled to může vypadat jako jednoduchý systém, ale při velkém počtu interních parametrů modelu se ukazuje, že má

model schopnosti vyšších forem učení jako je rozpoznávání kontextu, řešení matematických úloh a porozumění cizím textům. [14]

Při trénování se trénovací text rozdělí na tokeny, každý token představuje unikátní textovou jednotku. Tokenem může být slovo, jeho část nebo také kombinace interpunkce a slov. Každý token má specifické ID, díky kterému lze text vyjádřit jako sekvenci těchto ID. [15]

Po rozdělení textu na tokeny, se ke každému tokenu přiřadí kontextový vektor, známý jako vkládání. Tento vektor obsahuje více číselných hodnot, které reprezentují sémantické vlastnosti daného tokenu. Hodnoty vektoru se určují na základě četnosti společného výskytu tokenů nebo podobnosti jejich kontextu. [15]

Úkolem modelu je na základě předchozích tokenů předpovědět, jaký token bude následovat. Každému tokenu v sekvenci se přiřazují váhy, které určují jeho vliv na predikci dalšího tokenu. Následně se provede výpočet, který s využitím těchto vah a vkládání vypočítá pravděpodobné hodnoty vektoru následujícího tokenu. Model poté vybere nejpravděpodobnější token, který pokračuje v sekvenci. [15]

Tento proces probíhá opakovaně pro každý token v řadě, přičemž každá nově vygenerovaná sekvence slouží jako vstup pro další iteraci. Výstup se tedy skládá postupně, token po tokenu, podobně jako u automatického doplňování textu, kde se výsledky přizpůsobují na základě dosud zadaných vstupů. [15]

Během trénování model zná celou sekvenci tokenů, ale ignoruje všechny tokeny následující za aktuálně zpracovávaným tokenem. Predikovaný vektor dalšího tokenu se porovná se skutečnou hodnotou a na základě výpočtu chyby se model postupně upravuje tak, aby minimalizoval ztrátu a zlepšil svou přesnost. [15]

Výkon modelu lze také navýšit pomocí technik jako prompt engineering, doladění modelu, nebo zpětnovazební učení s lidským dohledem. Tyto metody nám pomáhají odstranit nežádoucí zkreslení, nenávislné projevy a tzv. halucinace, neboli nepřesné informace, které vznikají často jako nežádoucí vedlejší produkt kvůli školení na velkém množství nestrukturovaných trénovacích dat. [14; 15]

Správné vyladění modelu je potřeba k tomu, aby byly připraveny k nasazení, kde se klade důraz na spolehlivost a také, aby nevystavovaly firmy k nechtěné odpovědnosti. [14]

4.2 Případy užití velkých jazykových modelů

Velké jazykové modely přinášejí změny do velkého množství průmyslů, včetně financí, zdravotnictví, financí a další. Vylepšují konverzační umělou inteligenci ve virtuálních asistentech a chatbotech, poskytují lepší interakce a uživatelskou zkušenost. Mají uplatnění v různých oblastech a mohou zefektivňovat mnoho procesů. [14]

4.2.1 Generování textu

Je možné generovat rozsáhlé texty na základě zadání, tyto modely jsou schopny generovat texty emailů, blogových článků. [14]

4.2.2 Virtuální asistenti

Jedná se o programy, které díky zpracování přirozeného jazyka dokážou napodobovat lidskou řeč. Mohou poskytovat odpovědi na dotazy zákazníků, provádět běžné úkoly a zajišťovat automatickou 24hodinovou online podporu. [14]

4.2.3 Generování kódu

Pomáhají programátorům při psaní kódu, při sestavování aplikací, při debuggování aplikací, nebo také s překládáním kódu mezi různými programovacími jazyky. [14]

4.2.4 Analýza sentimentu

Modely dokážou rozpoznat tón a emoce zákazníka v textu, často se využívají k analýze zpětné vazby od zákazníků. [14]

5 CHATGPT

Jedná se o chatovacího bota s umělou inteligencí, který využívá zpracování lidského jazyka k tvorbě konverzací, které vypadají jako lidské. Jeho služby využívají miliardy lidí po celém světě. Dokáže odpovídat na dotazy různých témat a také je schopný generovat texty na webové stránky, seminární práce, kód anebo také odpovědi na emaily. [16; 17]

Jedná se o formu generativní umělé inteligence, která dokáže generovat obrázky, videa a texty podle příkazu zadaného uživatelem. Služba je podobná různým online chatovacím asistentům na webových stránkách, kterým lidé pokládají dotazy a modely jim odpovídají. [16]

GPT popisuje způsob, jakým model zpracovává požadavky a vytváří odpovědi. ChatGPT je trénován pomocí posilovaného učení, které využívá zpětnou vazbu od lidí a modely odměn, jež hodnotí nejlepší odpovědi. Tato zpětná vazba pomáhá zlepšovat ChatGPT díky strojovému učení, což vede k lepším odpovědím v budoucnosti. [16]

Tento model byl vytvořen firmou OpenAI, která se zabývá výzkumem umělé inteligence a byl spuštěn v listopadu 2022. Tato firma byla založena v roce 2015 a je podporována několika významnými firmami jako Microsoft. Mezi zakladatele patří Elon Musk a Sam Altman. [16]

ChatGPT využívá hluboké učení a pomocí transformátorových neuronových sítí vytváří texty podobné lidským. Na základě sekvence tréninkových dat se snaží transformátor předpovědět další text. Na rozdíl od jiných chatbotů si dokáže ChatGPT zapamatovat otázky a tím dodává konverzaci větší plynulost. [16]

Trénování probíhá na začátku na obecných datech a následně se přesune k datům, které jsou více konkrétní. Tento model byl trénován na online textu použitého z databází, přesněji se jednalo asi o 300 bilionů slov. Jednalo se o data z knih, webových stránek, ze kterých se naučil lidský jazyk, následně začal používat skripta, aby se naučil základní konverzace. Uživatelé fungují jako lidští trenéři, kteří mají možnost hodnotit odpovědi, u každé zprávy mají možnost dát palec nahoru nebo dolů podle toho, jak jim odpověď pomohla, případně napsat i zpětnou vazbu. [16; 17]

Lidé ho nejčastěji využívají k programování a hledání chyb v kódu, k psaní emailů, řešení matematických problémů, k vytváření textů článků, k hledání klíčových slov pro web, ale také ke skládání hudby. [16]

5.1 Výhody ChatGPT

Tento chatovací bot má několik výhod, které se ale neustále rozrůstají tím, jak se program vyvíjí. Mezi hlavní výhody patří: [18]

Dostupnost – jelikož je poháněn umělou inteligencí a je online, můžeme jej žádat o pomoc kdykoliv a odkudkoliv. [18; 19]

Ušetření času – je možné jej využít pro hledání odpovědí bez nutnosti ručního hledání. Algoritmy tohoto modelu hledají dostupné informace a během několika sekund nám dají odpověď. [19]

Stále se učí – ChatGPT se neustále učí z interakce s lidmi. Shromažďuje informace o lidech, jak se chovají, jaké mají požadavky a pomocí těchto informací se neustále zdokonaluje. [19]

Cena – Nejnovější model je i pro uživatele, kteří neplatí žádné předplatné, ale pouze na omezený počet příkazů. Starší model je dostupný všem, bez omezení. Lidé, co si platí tuto službu, mohou využívat novějšího modelu společně s nahráváním obrázků a dokumentů a dalšími výhodami. [17; 18]

Aktualizace – probíhají neustále aktualizace a vylepšení, aby zůstal model co nejpřesnější a mohl poskytovat co nejlepší odpovědi. Firma OpenAI pravidelně školí ChatGPT s novými daty, což pomáhá získávat nové znalosti. [18]

Sebevzdělávání – tento model může fungovat jako učitel, který nám může pomoci s řešením úkolů. Dokáže nám dát přesný postup, jak problém vyřešit a tím zdokonalit naše vědomosti. [18]

5.2 Nevýhody ChatGPT

I když má ChatGPT mnoho výhod, existuje i velké množství nevýhod a omezení, na které si uživatelé musí dát pozor. Mezi ně řadíme následující: [16]

Zaujatost a předsudky – Tréninková data, na kterých ChatGPT pracuje, mohou obsahovat předsudky, tyto předsudky mohou ovlivňovat výsledné odpovědi. I když se vyvíjejí metody pro minimalizaci těchto předsudků, stále mohou přetrvávat. [18]

Nedostatečné pochopení lidské komunikace – ChatGPT ne vždy plně chápe složitost jazyka. To může vést k povrchním odpovědím, které postrádají hlubší porozumění. Někdy mohou jeho reakce znít uměle a nepřirozeně. [16]

Etické problémy – Generování obsahu umělou inteligencí může vyvolat otázky ohledně autenticity a pravdivosti. ChatGPT například nedokáže poskytovat analýzu nebo odkazy na zdroje, což může vést k neúplným informacím a potenciálně šířit neověřené nebo chybné údaje. [16; 18]

Problémy s plagiátorstvím – Vzhledem k tomu, že ChatGPT může vytvářet obsah podobný chráněným materiálům, může vzniknout otázka porušování autorských práv, což je stále nejednoznačné, zvláště v akademickém a kreativním kontextu. [18]

Absence emocionální inteligence – ChatGPT nemá schopnost chápat lidské emoce v komunikaci. To může způsobit, že jeho odpovědi budou nevhodné nebo neadekvátní pro danou situaci. [16; 18]

Omezená kontrola nad kontextem – ChatGPT může mít problémy s porozuměním kontextu, což může vést k odpovědím, které jsou správné, ale nevhodné nebo nepřesné v konkrétním kontextu [18]

Politická nestrannost a zaujatost – Ačkoli ChatGPT se vyhýbá zapojení do politických diskusí, někdy může vykazovat skrytý favoritismus k určitým politickým názorům, což může být vnímáno jako zaujatost. [16]

6 VYUŽITÍ UMĚLÉ INTELIGENCE PRO TRANSKRIPCI

Technologie automatického rozpoznávání řeči (ASR) představuje významný milník v oblasti zpracování přirozeného jazyka a umělé inteligence. Její vliv je patrný napříč různými oblastmi lidské činnosti, od vzdělávání, přes zdravotnictví, až po média a korporátní sféru. ASR umožňuje převod mluveného slova do psané podoby, čímž usnadňuje přístup k informacím, podporuje inkluzi osob se specifickými potřebami a výrazně zefektivňuje pracovní procesy. [20]

6.1 Principy fungování transkripce

ASR systémy využívají pokročilé algoritmy strojového učení, které analyzují zvukový signál a převádějí jej na text. Tento proces obvykle zahrnuje následující kroky: [20]

- Detekce a nahrání mluveného projevu. [20]
- Úprava zvukového záznamu (např. odstranění šumu, normalizace hlasitosti). [20]
- Segmentace signálu a jeho analýza pomocí akustických a jazykových modelů. [20]
- Rekonstrukce celých vět na základě pravděpodobnostních výpočtů. [20]

Moderní systémy dále využívají komponenty z oblasti zpracování přirozeného jazyka (NLP), které umožňují porozumění kontextu a zvyšují přesnost přepisu. V některých případech je výsledný text doplněn o lidskou revizi, která slouží ke korekci chyb a dosažení vyšší přesnosti, zejména v případě legislativních požadavků na přístupnost. [20]

6.2 Využití automatické transkripce

Uplatnění technologie automatického rozpoznávání řeči nalezneme v různých sektorech, mezi které patří: [20]

- Vzdělávání – podpora studentů se sluchovým postižením, jazykovou bariérou nebo poruchami pozornosti prostřednictvím titulků a přepisů. [20]
- Právo – řešení nedostatku soudních zapisovatelů, zaznamenávání výpovědí a soudních jednání. [20]
- Zdravotnictví – automatizovaný přepis lékařských záznamů, poznámek z vyšetření a operací. [20]

- Média a firemní sektor – živé titulkování, transkripce jednání, zpřístupnění obsahu pro zaměstnance a zákazníky. [20]

6.3 Výhody automatické transkripce

Kombinace umělé a lidské inteligence umožňuje dosáhnout vyšší kvality přepisu, která je zásadní zejména v kontextech, kde jsou stanoveny legislativní nároky na přesnost.

K hlavním výhodám patří: [20]

- Rychlost zpracování velkého množství dat. [20]
- Možnost rozlišování jednotlivých mluvčích. [20]
- Přizpůsobení specifické terminologii nebo odbornému jazyku. [20]
- Možnost trénování systému na konkrétní datové sady. [20]
- Škálovatelnost a ekonomická efektivita. [20]

7 PROBLEMATIKA PODMÍNEK UŽITÍ (TERMS OF USE)

Tento dokument zaručuje uživatelům důvěru ve webovou stránku, kterou používají. Terms of Use neboli podmínky užití je právní dohoda mezi majitelem webu a uživatelem webu. Měly by se objevovat na každé webové stránce. Jedná se o pravidla pro využívání stránek nebo služby, které musí být dodržovány uživatelem. Také sděluje, jaké práva majitel stránky má v případě, že jsou tyto pravidla porušena. Existují 3 důležité důvody, proč mít na stránce Terms of Use. [21; 22]

Sdělení uživatelům o podmínkách služeb jednodušeji – tím, že si uživatel přečte vaše podmínky užití, může zjistit více o službách, které nabízíte, a jaké jsou jejich podmínky. Zároveň to uživatelům dodává důvěru ve webovou stránku, protože se o ni majitel neustále stará. [22]

Právní ochrana pro majitele webu – pokud v podmínkách webu uvedeme, že nechceme, aby byl náš duševní majetek sdílen, bude tento dokument sloužit jako naše právní ochrana v případě, že by to někdo porušil. [22]

Kontrola akce uživatelů – díky podmínkám používání můžete kontrolovat, co uživatelé na vašich stránkách dělají, a tím se vyhnout nežádoucím aktivitám. [22]

7.1 Typy Terms of Use

Rozlišujeme dva typy, oba typy jsou prakticky to stejné, liší se jen ve formě, ve které jsou představeny uživatelům webových stránek. [22]

7.1.1 Browsewrap

Jedná se o častější verzi podmínek užití. Na webu, nejčastěji v patičce, se nachází odkaz na stránku s podmínkami. Uživatel musí tento odkaz najít a navštívit sám, aby si tyto podmínky přečetl a byl s nimi obeznámen. [22; 23]

7.1.2 Clickwrap

Uživateli webu se zobrazí vyskakovací okno, ve kterém musí potvrdit přečtení podmínek užití kliknutím na tlačítko souhlasit, popřípadě zaškrtnutím souhlasu. Dokud uživatel nepotvrdí přečtení podmínek, nemůže si prohlížet některé části stránky nebo někdy dokonce i celou stránku. Tím docílíme, že uživatel nemůže podniknout žádné právní kroky, protože souhlasil a přijmul podmínky užití. V případě prodeje zboží jsou webové stránky

v některých zemích povinné zobrazovat podmínky užití tímto způsobem. Ve většině případů se potom objevuje zaškrtačací pole při odeslání objednávky. [22; 23]

8 STRÁNKY S PROBLEMATIKOU TERMS OF USE

S rostoucím důrazem na digitální soukromí a transparentnost online služeb vznikají nástroje, které se snaží zjednodušit orientaci v často komplikovaných a rozsáhlých dokumentech, jako jsou podmínky užívání nebo zásady ochrany osobních údajů. Dva z nejvýznamnějších nástrojů v této oblasti představují projekty PrivacySpy a ToS;DR.

8.1 PrivacySpy

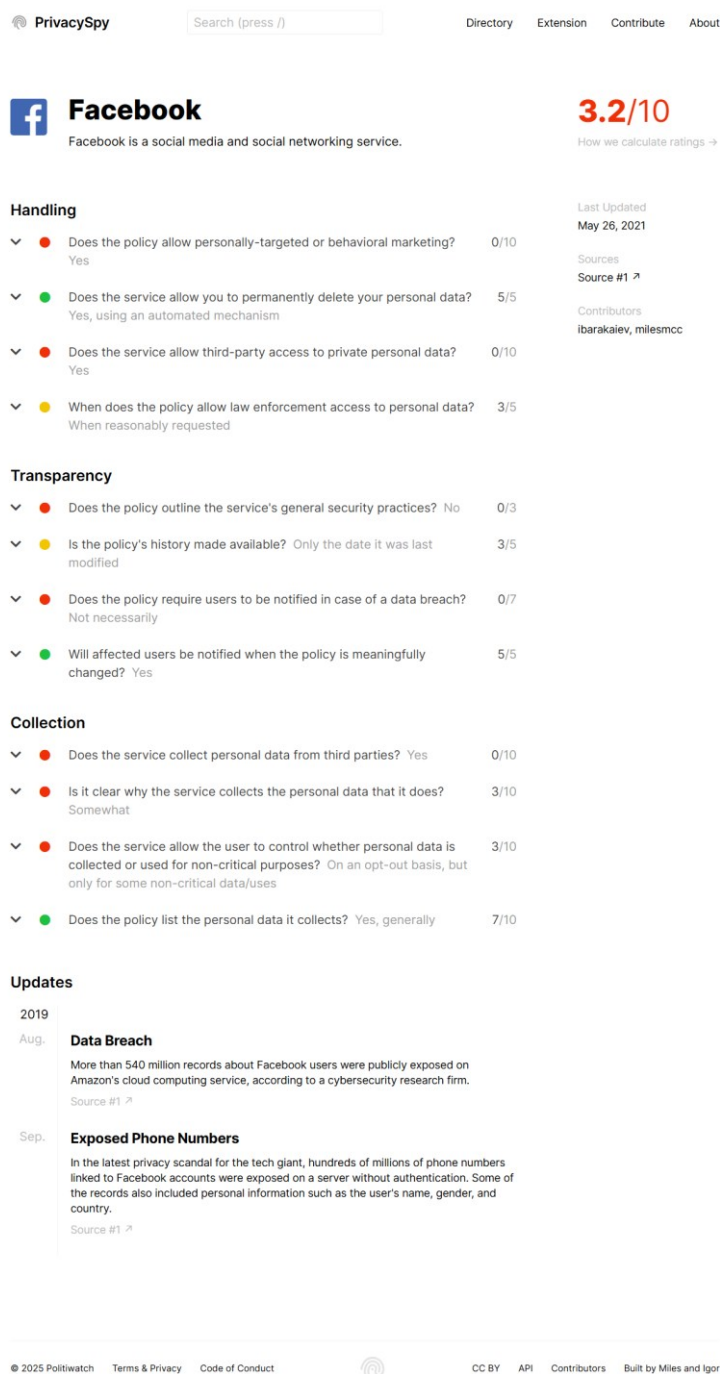
Projekt PrivacySpy vznikl jako reakce na nízkou čitelnost a srozumitelnost zásad ochrany osobních údajů, které uživatelé často přehlíží. Jeho cílem je zpřístupnit a zjednodušit tyto dokumenty široké veřejnosti, aniž by bylo nutné je celé detailně pročítat. [24]

PrivacySpy využívá konzistentní hodnoticí rubriku, pomocí níž analyzuje jednotlivé zásady a přiřazuje jim skóre na desetibodové škále. Toto skóre reflektuje míru transparentnosti a důvěryhodnosti přístupu dané společnosti k nakládání s osobními údaji. Hodnocení je doplněno o poznámky a citace, které zajišťují transparentnost celého procesu. [24]

Projekt si klade za cíl nejen informovat uživatele, ale také motivovat společnosti k vyšší odpovědnosti, a to nejen za samotnou existenci zásad, ale i za jejich kvalitu a srozumitelnost. Kromě samotného hodnocení dokumentů PrivacySpy také sleduje jejich změny a aktualizace. V případě odhalení nevhodného nakládání s osobními daty je tento stav zaznamenán a zohledněn v hodnocení dané služby. [24]

Součástí projektu je i slovník používaných pojmů a transparentní metodika výpočtu skóre. Například rozdíl mezi kritickým a nekritickým použitím dat je důležitým hodnoticím kritériem. Zatímco využití e-mailu pro resetování hesla je považováno za kritické, jeho využití pro behaviorální marketing už nikoli. [24]

PrivacySpy tak usiluje o vytvoření prostředí, kde budou zásady ochrany osobních údajů nejen právní povinností, ale také smysluplným a srozumitelným nástrojem pro informované rozhodování uživatelů. [24]



PrivacySpy Search (press /) Directory Extension Contribute About

Facebook 3.2/10
Facebook is a social media and social networking service. [How we calculate ratings →](#)

Handling Last Updated May 26, 2021
[Sources](#)
[Source #1 ↗](#)
[Contributors](#)
ibarakalev, milesmcc

- Does the policy allow personally-targeted or behavioral marketing? 0/10
Yes
- Does the service allow you to permanently delete your personal data? 5/5
Yes, using an automated mechanism
- Does the service allow third-party access to private personal data? 0/10
Yes
- When does the policy allow law enforcement access to personal data? 3/5
When reasonably requested

Transparency

- Does the policy outline the service's general security practices? No 0/3
- Is the policy's history made available? Only the date it was last modified 3/5
- Does the policy require users to be notified in case of a data breach? 0/7
Not necessarily
- Will affected users be notified when the policy is meaningfully changed? Yes 5/5

Collection

- Does the service collect personal data from third parties? Yes 0/10
- Is it clear why the service collects the personal data that it does? Somewhat 3/10
- Does the service allow the user to control whether personal data is collected or used for non-critical purposes? On an opt-out basis, but only for some non-critical data/uses 3/10
- Does the policy list the personal data it collects? Yes, generally 7/10

Updates

2019

Aug. **Data Breach**
More than 540 million records about Facebook users were publicly exposed on Amazon's cloud computing service, according to a cybersecurity research firm.
[Source #1 ↗](#)

Sep. **Exposed Phone Numbers**
In the latest privacy scandal for the tech giant, hundreds of millions of phone numbers linked to Facebook accounts were exposed on a server without authentication. Some of the records also included personal information such as the user's name, gender, and country.
[Source #1 ↗](#)

© 2025 Politiwatch Terms & Privacy Code of Conduct CC BY API Contributors Built by Miles and Igor

Obrázek 2 Ukázka hodnocení na stránce Privacyspy [25] (vlastní zpracování – screenshot)

8.2 ToS;DR (Terms of Service; Didn't Read)

Projekt ToS;DR vznikl v roce 2012 jako reakce na všeobecně rozšířený jev, kdy uživatelé běžně souhlasí s podmínkami služeb, aniž by je četli. Inspirací pro název byla populární internetová zkratka TL;DR (Too Long; Didn't Read), která ironicky reflektuje délku a komplikovanost právních dokumentů na internetu. [26]

Počátky projektu sahají do roku 2011, kdy byla na Chaos Communication Campu poblíž Berlína představena základní idea: vytvořit otevřený nástroj, který uživatelům poskytne srozumitelné shrnutí smluvních podmínek různých online služeb. Iniciativu odstartoval Hugo Roy, který se zasloužil o právní analýzu dokumentů. Projekt od té doby funguje jako komunitní open-source iniciativa, do níž přispívají dobrovolníci, a která je financována neziskovými organizacemi a dary jednotlivců. [26]

Hodnocení jednotlivých služeb probíhá veřejně a transparentně, za účasti širší komunity. Výsledky jsou poskytovány jako otevřená data a software. Cílem je nejen informovat uživatele o tom, jak služby nakládají s jejich daty a právy, ale také vyvíjet tlak na společnosti, aby své podmínky zjednodušovaly a zpřehledňovaly. [26]

ToS;DR je tak důležitým nástrojem v boji proti tzv. „největší lži internetu“, tedy iluzi, že uživatelé vědomě a plně souhlasí s podmínkami, které ve skutečnosti nikdy nečetli. [26]

8.3 TOSDR API

Projekt **ToS;DR** poskytuje několik RESTful API rozhraní, která umožňují interakci s jeho databází služeb a jejich smluvních podmínek. Níže je přehled dostupných API:

8.3.1 Service API

Toto API umožňuje získávat informace o službách uložených v databázi ToS;DR. Poskytuje dva hlavní endpointy:

- **GET /service/v3/**: Aktuální verze pro získání seznamu všech služeb nebo konkrétní služby podle ID. developers.tosdr.org
- **GET /service/v2/**: Starší verze s podobnou funkcionalitou.

Oba endpointy vracejí data ve formátu JSON.

8.3.2 Case API

Toto rozhraní slouží k získávání informací o jednotlivých případech (tzv. "cases") z databáze ToS;DR. Umožňuje získat konkrétní případ podle jeho ID nebo seznam všech případů. Endpoint:

GET /case/v2/: Vrací data ve formátu JSON.

8.3.3 Document API

Toto API poskytuje přístup k dokumentům souvisejícím se službami, jako jsou smluvní podmínky nebo zásady ochrany osobních údajů. Umožňuje získat seznam dokumentů nebo konkrétní dokument podle ID.

8.3.4 Search API

Umožňuje vyhledávání služeb v databázi ToS;DR na základě zadaného dotazu. Například:

GET /search/v5/:

8.3.5 Privacy Shields API

Toto API poskytuje "privacy shields" pro služby, což jsou grafické odznaky zobrazující hodnocení služby z hlediska ochrany soukromí. Endpoint:

https://shields.tosdr.org/<locale>_<service_id>.svg: Vrací SVG obrázek s hodnocením služby.

8.3.6 Error Codes

ToS;DR API používá specifické chybové kódy k indikaci různých stavů a chyb. Tyto kódy jsou implementovány napříč různými API rozhraními a je důležité je zohlednit při implementaci klientských aplikací.

8.4 Ochrana osobních údajů a GDPR

Jedním z hlavních důvodů, proč se vůbec zabývat analýzou podmínek užívání (Terms of Use), je ochrana osobních údajů. V současnosti téměř každá online služba nějaká data o uživatelích sbírá, ať už se jedná o e-mailovou adresu, IP adresu, informace o chování na webu nebo platební údaje. To, jakým způsobem jsou tato data zpracovávána a komu jsou předávána, bývá často ukryto právě v těchto rozsáhlých dokumentech, které většina uživatelů nikdy nečte.

Díky nařízení GDPR (Obecné nařízení o ochraně osobních údajů), platnému v rámci Evropské unie, mají uživatelé jasně definovaná práva. Patří mezi ně například právo vědět, jaké údaje o nich, kdo uchovává, právo na výmaz osobních údajů, právo na jejich ochranu nebo možnost kdykoli odvolat souhlas se zpracováním. Služby by měly být v tomto směru transparentní a umožnit uživatelům mít kontrolu nad tím, co se s jejich daty děje.

Problémem však zůstává, že podmínky užívání tyto skutečnosti často popisují složitě, nejasně nebo je skrývají mezi dalšími částmi textu. Právě proto má smysl takový dokument analyzovat a extrahovat z něj klíčové informace, zejména s ohledem na to, zda služba respektuje práva uživatelů podle GDPR, nebo naopak sbírá a sdílí údaje bez dostatečného vysvětlení.

Navrhované řešení umožňuje jednoduše zjistit, jak se služba staví k ochraně soukromí, jaká jsou potenciální rizika, a jestli potencionálně umožňuje snadné odstranění účtu nebo export osobních dat. To všechno jsou věci, které by měl mít uživatel předem na očích, aby se mohl rozhodnout, jestli službu chce používat, nebo raději zvolí jinou.

Právě proto může být analýza klíčovým nástrojem pro ochranu soukromí uživatelů v digitálním světě.

9 METODY SBĚRU DAT

Existuje několik různých zdrojů, z nichž můžou pocházet testovací data pro trénování modelů umělé inteligence. Pochopení, jak sběr dat probíhá je důležité pro pochopení, jak se modely umělé inteligence učí. [27]

9.1 Web scraping

Jedná se o automatické získávání velkého množství dat z webových stránek. Jde převážně o nestrukturovaná HTML (Hypertext Markup Language) data, které je následně možné převádět do strukturovaných formátů jako například JSON nebo CSV. [28; 29]

Web scrapery můžou získávat všechna data z webové stránky nebo jenom data, které uživatel požaduje. Nejideálnější je, pokud si uživatel zadá přesná data, díky čemuž urychlí získávání těchto dat. [28]

Jako první programu sdělíme URL (Uniform Resource Locator) adresu stránky, kterou chceme procházet. Následně si načte HTML kód stránky, přičemž některé pokročilejší scrapery dokážou získávat i jiné programovací jazyky než HTML. Program z HTML kódu vyčte data a uloží je do definovaného formátu. [28]

Existuje několik typů web scraperů, jako například Self-built Web Scrapers, které si vytvoří člověk sám. Jsou k tomu potřeba programovací znalosti. Dále mohou být Pre-built Web Scrapers, které uživatel pouze stáhne a může je používat. Existují také scrapery v podobě rozšíření v prohlížeči, které ale mají omezené funkce. [28]

Nejčastěji se pro programování vlastního web scraperu využívá programovací jazyk Python, který obsahuje několik knihoven jako je Scrapy nebo BeautifulSoup, které jsou vytvořené speciálně pro účel web scraperů. [28]

9.1.1 BeautifulSoup

Jedná se o jednu z nejoblíbenějších knihoven v jazyce Python, která usnadňuje získávání dat z HTML a XML souborů, existuje již od roku 2004. [30]

Poskytuje snadnou manipulaci s HTML strukturou, intuitivní nástroje pro vyhledávání a úpravu HTML dokumentu a umožňuje najít pouze konkrétní prvky. Automaticky převádí vstup na Unicode a UTF-8 a tím zajišťuje, že se o převod nemusí starat uživatel. [30]

Tato knihovna umožňuje snadnou extrakci dat z webových stránek a jejich rychlou konverzi do prakticky využitelné podoby. [30]

9.2 API

Některé webové stránky nabízejí používání jejich dat pomocí tzv. API. Představuje alternativu webscrapingu. Díky tomu se můžeme vyhnout zpracování HTML kódu a máme přímý přístup k JSON nebo XML datům. Jedná se o soubor pravidel a protokolů, díky kterým mezi sebou mohou softwarové aplikace vzájemně komunikovat. Definuje pravidla, jak mezi sebou mohou aplikace komunikovat a jaké informace si mohou předávat. [31; 32]

Mnoho populárních služeb zveřejňuje svá API pro vývojáře, aby mohli pracovat s jejich daty. Například Github nabízí informace o repozitářích a uživateli, Reddit nabízí API s přístupem k příspěvkům a komentářům a mnoho dalšího. [32]

9.2.1 API vs webscraping

Zda využít API nebo webscraping záleží na několika parametrech, jako dostupnost dat, oprávnění k přístupu a nároky na zdroje. V tabulce níže můžeme vidět a porovnat podstatné faktory, které rozhodují o tom, kterou službu využít. [32]

Tabulka 1 API vs Webscraping

Webscraping	API
Možnost získávat data z jakékoliv veřejné stránky	Je limitováno na stránky, které nabízejí veřejnou API
Může být blokován stránkami, které mají systémy proti botům	Může obsahovat omezení na základě zásad použití
Získává data v nezpracovaném HTML, je potřeba provést čištění a převod dat	Data jsou strukturována v JSON/XML a připravena k použití
V případě velkých webových stránek pomalejší	Díky přímému přístupu k datům je mnohem rychlejší
Musíme dávat pozor na podmínky stránek, aby webscraping nebyl nelegální	Přesně stanovené podmínky užití API dle majitele

(zdroj: [32])

10 PYTHON

Python patří mezi jeden z nejpopulárnějších a nejrozšířenějších programovacích jazyků. Je ideální pro začátečníky díky jeho univerzálnosti a jednoduchosti. Slouží k navrhování jednoduchých programů, ale obsahuje prostředky i pro vytváření velkých aplikací. Jedná se o jazyk, který má velký rozsah využití jako je vývoj webových stránek, vývoj aplikací, automatizace, analýza dat a využívá se také v umělé inteligenci. Tento jazyk hraje klíčovou roli například v doporučovacím algoritmu Netflixu či v software, který řídí autonomní vozidla. [33; 34]

Popularita Pythonu stále roste i díky tomu, že obsahuje mnoho knihoven, které slouží k ulehčení programování. Programátor se tak může zaměřit na jeho úkol a nemusí programovat jiné základní podprogramy. Často se využívá pro analýzu dat a strojové učení, kde nabízí široké možnosti vizualizace a statistického zpracování. Jeho jednoduchá syntaxe a srozumitelnost umožňuje, aby ho využívali nejen profesionální vývojáři, ale i vědci, analytici nebo administrátoři pro automatizaci běžných úkolů. [33; 34]

Ve webovém vývoji slouží Python k vytváření serverových aplikací a správě databází. Frameworky jako Django a Flask usnadňují práci na back-endu a umožňují rychlý vývoj robustních webových řešení. [33]

Python také nachází uplatnění v testování softwaru, kde pomáhá s automatizací testovacích procesů a odhalováním chyb v kódu. Nástroje jako Green a Requestium umožňují efektivní testování nových funkcionalit. [33]

Další oblastí, kde Python vyniká, je automatizace opakujících se úkolů. Pomocí skriptů lze zjednodušit činnosti, jako je hromadné přejmenovávání souborů, extrakce dat z webových stránek nebo automatické odesílání e-mailů. [33]

II PRAKTICKÁ ČÁST

11 PRŮZKUM WEBOVÉ STRÁNKY TOSDR.ORG

Webová platforma Terms of Service Didn't Read se zaměřuje na analýzu a hodnocení podmínek užívání a zásad ochrany osobních údajů různých internetových služeb. Jejím hlavním cílem je usnadnit uživatelům orientaci v často složitých a obsáhlých právních dokumentech, které většina lidí běžně nečte. Nabízí proto stručný a srozumitelný přehled o tom, jak jednotlivé platformy nakládají s osobními údaji uživatelů a jaká práva a povinnosti z jejich používání vyplývají.

11.1 Účel tosdr.org

Tento projekt vznikl jako komunitní iniciativa s cílem upozorňovat na nepřehledné a často nevýhodné podmínky, které technologické společnosti nastavují ve svůj prospěch. Funguje na takovém principu, že dobrovolníci z celého světa se podílejí na analýze smluvních podmínek různých služeb a shrnují nejdůležitější informace, které mají zásadní dopad na uživatele.

11.2 Jak to funguje

Platforma poskytuje hodnocení internetových služeb na základě několika klíčových kritérií. Každá z nich získává známku od A (velmi dobře) po E (velmi špatné), která odráží míru ochrany soukromí a férovosti podmínek. Posuzované aspekty jsou rozděleny na pozitivní (například transparentnost, možnost smazání účtu, ochrana uživatelských práv) a negativní (sledování aktivit, sdílení osobních údajů s třetími stranami, omezování práv uživatelů).

Ke každému bodu je připojen podrobný komentář spolu s odkazem na konkrétní pasáž v podmínkách dané služby. To umožňuje uživatelům rychle získat přehled o tom, jak s jejich daty služba zachází, aniž by museli číst dlouhé právní texty.

11.3 Obsah stránek

Na stránce tosdr.org lze nalézt několik hlavních sekcí, které zahrnují:

- Přehled hodnocených služeb – seznam online platform s jejich aktuálním ratingem a odkazy na podrobné analýzy.
- Detailní analýzy služeb – každá analyzovaná služba má svou vlastní stránku s podrobným rozbořem jednotlivých bodů a jejich hodnocením.

- Seznam klíčových bodů – souhrn nejčastěji posuzovaných oblastí, například možnost smazání účtu, podmínky pro sdílení osobních údajů nebo transparentnost smluvních podmínek.
- Možnost přispění – stránka umožňuje uživatelům se aktivně zapojit do hodnocení služeb a přidávat vlastní analýzy.

11.4 Využití projektu tosdr.org

Díky této platformě mají lidé možnost lépe porozumět podmínkám, které odsouhlasují při používání různých online služeb, a činit informovanější rozhodnutí. Jelikož je projekt neustále aktualizován komunitou, reflektuje změny v podmínkách jednotlivých platforem a přispívá k větší transparentnosti digitálního prostoru.

Tento nástroj zároveň vytváří tlak na technologické společnosti, aby poskytovaly srozumitelnější a férovější podmínky. Pro běžné uživatele tak představuje cenný zdroj informací, který jim pomáhá chránit jejich práva na internetu.

12 NÁVRH FORMÁTU DAT PRO ZPRACOVÁNÍ

Pro úspěšné zpracování smluvních podmínek pomocí velkých jazykových modelů (LLM), jako je ChatGPT, je důležité pracovat s takovým formátem dat, který bude vhodný jak pro načítání z externích zdrojů, tak pro následné použití při sestavování vstupů (promptů) pro model. Jelikož tyto modely nejsou určeny k přímému zpracování technických struktur jako JSON, bylo potřeba navrhnout meziformát, který technická data převede do podoby přirozeného jazyka. Tento formát zároveň musí být dostatečně flexibilní, aby mohl být použit i pro různé typy dokumentů.

12.1 Požadavky na vstupní data

Základním požadavkem při návrhu bylo, aby byl výsledný formát dobře čitelný nejen pro model, ale i pro samotného uživatele. Zdroje dat, jako je veřejné API projektu ToS;DR, sice poskytují obsah ve strukturované podobě (např. JSON), ale tato forma je pro jazykový model příliš technická. Pokud má být výstup modelu kvalitní a srozumitelný, je nutné vstupní informace upravit do formy, která bude lépe odpovídat očekávání modelu, tedy přirozený, srozumitelný a kontextově bohatý text.

Zároveň bylo nutné, aby formát zachovával ty nejdůležitější informace, mezi které patří:

- obsah jednotlivých tvrzení nebo ustanovení z podmínek,
- hodnocení nebo interpretace daných bodů (pokud je dostupná),
- případné souvislosti nebo další kontext, který pomůže modelu pochopit, čeho se text týká.

12.2 Přístup k návrhu formátu

Při návrhu bylo zvoleno jednoduché, ale rozšiřitelné řešení. Namísto využití technických struktur se předpokládá převod dat do běžného textu, který bude model schopen dobře analyzovat. Struktura výsledného textu je navržena tak, aby obsahovala všechny klíčové informace z původních dat, ale zároveň nebyla zahlcena zbytečnými technickými detaily.

Jedním z důležitých aspektů je také možnost škálování. Formát musí fungovat jak při analýze jedné služby, tak i při zpracování více dokumentů najednou. Stejně tak musí být připraven na různou kvalitu a rozsah dat, od několika málo bodů až po rozsáhlé podmínky složitějších platforem.

Dalším požadavkem je možnost úprav nebo rozšíření. V budoucnu může být potřeba doplnit do struktury další informace, například právní kvalifikaci jednotlivých ustanovení, překlady, rizikovost nebo doporučení pro uživatele. Navržený formát proto musí být dostatečně otevřený a flexibilní.

12.3 Využitelnost navrženého formátu

Výhodou navrženého přístupu je, že umožňuje pracovat s daty z různých zdrojů, nejen z API ToS;DR, ale také s vlastními podmínkami vloženými uživatelem. Ať už se jedná o jednoduchý textový dokument, výstup z databáze nebo ručně zkopírované podmínky, formát umožňuje sjednotit jejich podobu do jednoho vstupu, se kterým může model dále pracovat.

Tento formát je navíc snadno pochopitelný i pro člověka, což je důležité v situacích, kdy má uživatel možnost si vstup před odesláním zkontrolovat. Přispívá to nejen k transparentnosti celého systému, ale i ke kvalitnějším výsledkům, protože promyšlená struktura vstupu vede k přesnějším a srozumitelnějším odpovědím modelu.

Zároveň platí, že i přestože je formát vytvořen hlavně pro jazykový model, může být využit i pro jiné účely, například pro archivaci analyzovaných podmínek, jejich další zpracování, nebo jako podklad pro porovnávání více služeb mezi sebou.

13 EXTRAKCE A PŘEDZPRACOVÁNÍ DAT Z TOSDR.ORG

Aby bylo možné využít potenciál jazykového modelu pro analýzu smluvních podmínek, bylo nejprve nutné zajistit odpovídající vstupní data. V tomto projektu byla jako hlavní zdroj informací zvolena platforma ToS;DR, která nabízí otevřenou databázi shrnutí a hodnocení podmínek užívání různých online služeb. Ta je spravována komunitně a zaměřuje se na to, aby uživatelé lépe chápali, s čím vlastně při registraci na daných platformách souhlasí.

ToS;DR zpřístupňuje svá data prostřednictvím veřejného REST API. Tato rozhraní vracejí odpovědi ve formátu JSON a umožňují přístup k detailním informacím o jednotlivých službách, včetně jejich názvu, známky hodnocení, tématických bodů a konkrétních shrnutí jednotlivých ustanovení smluvních podmínek.

13.1 Přístup k API

V rámci tohoto řešení je API využíváno ve dvou krocích. Nejprve je přes endpoint `/search/v5/` provedeno vyhledání služby podle jména, které uživatel zadá do aplikace. Vyhledání je navíc přizpůsobeno tak, aby pracovalo i s mírnými rozdíly v zápisu (např. velikost písmen, diakritika), což zvyšuje uživatelský komfort.

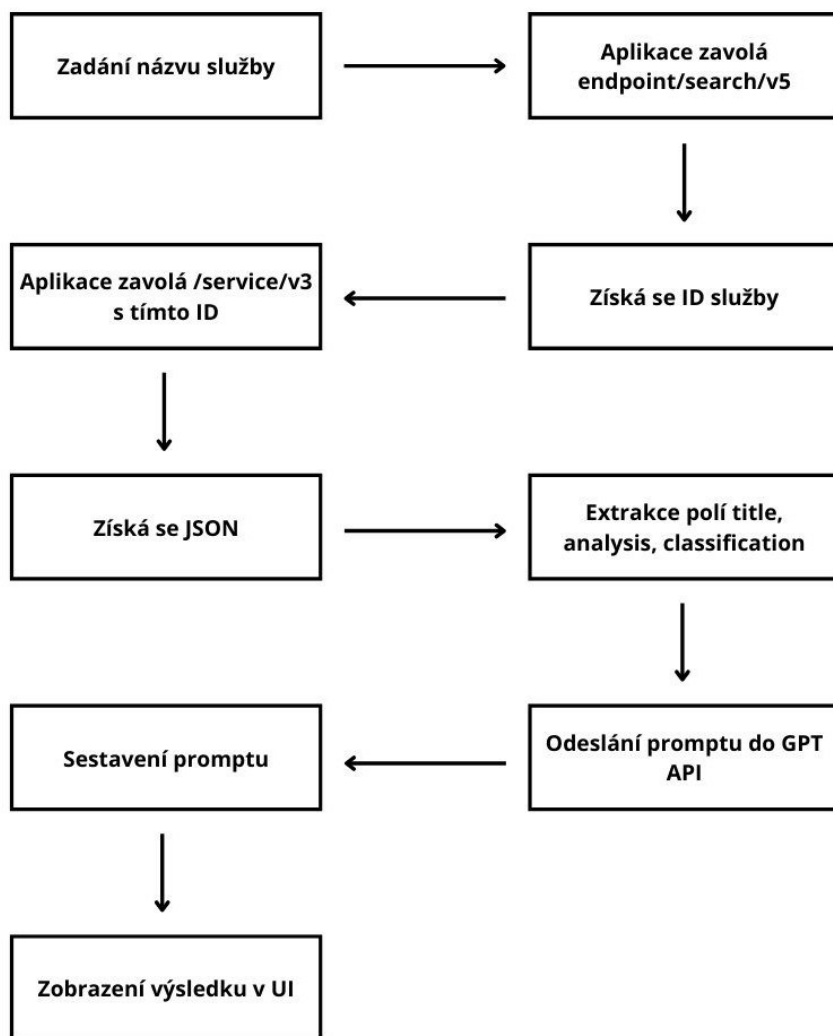
Jakmile je služba nalezena, je z jejího ID následně vyžádán detailní popis přes endpoint `/service/v3/`. Tento požadavek vrací nejen základní informace o službě, ale také seznam tzv. „points“, tedy konkrétních shrnutí jednotlivých klauzulí, která popisují např. sdílení dat, práva uživatelů, možnost smazání účtu a další důležité aspekty. Pro přístup k API je využita knihovna `requests` v jazyce Python, která umožňuje jednoduchou, ale efektivní práci s HTTP požadavky.

13.2 Získávání a výběr údajů

Po získání odpovědi z API byla provedena extrakce pouze těch údajů, které jsou relevantní pro další zpracování jazykovým modelem. Konkrétně se jedná o:

- název služby,
- celkovou známku hodnocení (rating),
- seznam jednotlivých tvrzení (points), přičemž z každého je použito pole `title` (krátké shrnutí) a `analysis` (komentář nebo vysvětlení).

Z každého bodu je kromě obsahu a vysvětlení extrahováno také pole *classification*, které určuje celkové hodnocení daného ustanovení („good“, „bad“, „neutral“). Tato informace je následně využita nejen při přípravě vstupu pro model, ale také pro barevné rozlišení bodů v grafickém rozhraní aplikace. Uživatel tak na první pohled vidí, které části jsou potenciálně rizikové, a které naopak pozitivní.



Obrázek 3 Schéma zpracování dotazu pomocí API ToS;DR (zdroj: vlastní)

13.3 Vytváření vstupu pro jazykový model

Přestože jsou data z API formálně dobře strukturovaná, nejsou vhodná pro přímé zadání do jazykového modelu. Proto je potřeba je převést do srozumitelnější a přirozenější textové

podoby. Tento krok zahrnuje vytvoření jednoduchého seznamu bodů, kde je každý záznam uveden ve formátu:

- [Název bodu]: [Vysvětlení]
- Například: Služba může sdílet vaše data s třetími stranami: ToS;DR uvádí, že služba umožňuje použití osobních údajů pro marketing.

Tento seznam pak tvoří základ promptu, který je rozšířen o uživatelské zadání, například: „Shrň hlavní výhody a nevýhody těchto podmínek. Upozorni na rizika. Na závěr dej doporučení a ohodnot' podmínky známkou.“

Tímto způsobem je zajištěno, že jazykový model dostane vstup, který je dobře srozumitelný, přehledný a obsahuje všechny důležité informace potřebné pro relevantní výstup.

13.4 Automatizace, kontrola chyb a stabilita

Celý proces je plně zautomatizován v rámci aplikace. Uživatel pouze zadá název služby a vyplní prompt. Systém poté automaticky:

- vyhledá službu v ToS;DR,
- načte její podmínky,
- sestaví text pro jazykový model,
- odešle požadavek do API ChatGPT,
- zobrazí výsledek v přehledném rozhraní,
- automaticky určí, zda se jedná o pozitivní nebo negativní bod a barevně označí.

Zároveň byly implementovány kontroly pro situace, kdy služba neexistuje, nemá žádná tvrzení, nebo dojde k chybě při načítání dat. V takovém případě uživatel obdrží srozumitelné chybové hlášení a aplikace se nespustí dál.

Díky této kombinaci automatizace, zpětné vazby a možnosti ručního zadání (například vlastního promptu) byla zajištěna, jak technická stabilita systému, tak i vysoká míra použitelnosti pro různé typy uživatelů.

Výstup je zároveň automaticky barevně označen podle hodnocení jednotlivých bodů: zeleně pro pozitivní, červeně pro negativní a oranžově pro neutrální body. Toto vizuální

zvýraznění výrazně usnadňuje orientaci v dlouhých textech a pomáhá uživatelům rychle rozpoznat klíčové informace.

14 POROVNÁNÍ A ODLIŠNOSTI NAVRHOVANÉHO ŘEŠENÍ

Navrhované řešení přináší moderní přístup k analýze smluvních podmínek (Terms of Use) pomocí umělé inteligence, konkrétně velkých jazykových modelů, jako je ChatGPT. Na rozdíl od existujících projektů, jako jsou ToS;DR nebo PrivacySpy, které fungují hlavně na základě ruční analýzy nebo předem nastavených pravidel, se tento projekt soustředí na automatické zpracování textu a jeho srozumitelné shrnutí pomocí AI.

Hlavní rozdíl spočívá v tom, že uživatel není omezený jen na služby, které jsou v databázi ToS;DR nebo jiných nástrojů. Můžeme totiž nahrát a analyzovat jakýkoliv dokument s podmínkami užívání, včetně vlastního textu nebo podmínek z méně známé služby, která doposud nebyla nikde hodnocena. Díky tomu může být nástroj použitelný i v případech, kdy jiné nástroje selhávají, protože danou službu vůbec neobsahují.

Velkou výhodou je také to, že si uživatel může sám určit, jaké informace chce z dokumentu získat. Není vázaný na předem daný seznam hodnocených bodů nebo konkrétní šablonu. Pomocí tzv. promptu si může napsat vlastní zadání, například že chce shrnutí hlavních výhod a nevýhod, upozornění na možná rizika, doporučení, nebo třeba známku jako ve škole. Díky tomu dostane výstup, který odpovídá přesně tomu, co ho zajímá.

Zajímavou funkcí je i možnost porovnat výsledky analýzy. Uživatel si může zobrazit výstup vytvořený ChatGPT a vedle toho i oficiální hodnocení ze služby ToS;DR, popřípadě vlastní služby. Porovnání je v aplikaci přehledně barevně vyznačené, což umožňuje snadno a rychle vidět rozdíly a shody mezi jednotlivými verzemi shrnutí. Navíc je v aplikaci uchovávána historie jednotlivých porovnání, takže se uživatel může kdykoliv vrátit k předchozím analýzám a přehodnotit je.

Barevné rozlišení představuje klíčový prvek celého řešení, kdy je každý bod ve výstupu označen barvou podle jeho významu: zeleně pro pozitivní aspekty (výhody), červeně pro negativní aspekty (rizika, nevýhody) a oranžově pro body s neutrálním nebo nejednoznačným charakterem. Toto zvýraznění usnadňuje rychlou orientaci ve výsledcích a šetří čas při vyhodnocování dlouhých dokumentů. Barevné značky jsou aplikovány nejen na výstup vytvořený AI modelem, ale také na data získaná z API služby ToS;DR. To umožňuje přímé vizuální srovnání oficiálního hodnocení a výstupu z umělé inteligence. Kromě toho je zajištěna plná kompatibilita i s uživatelsky nahranými dokumenty, včetně těch ve formátu PDF.

Na rozdíl od nástrojů, které pracují s pevně danými výstupy, nabízí toto řešení mnohem víc možností přizpůsobení. Uživatel má plnou kontrolu nad tím, co chce z textu zjistit, a jak má výstup vypadat. To zvyšuje nejen přesnost, ale i použitelnost celého nástroje v různých situacích.

Celkově se tedy jedná o univerzální nástroj, který kombinuje výhody umělé inteligence s jednoduchým a přístupným použitím. Není závislý na předem připravených datech, zvládne zpracovat libovolný text a výstupy jsou srozumitelné a přehledné. Díky tomu se jedná o užitečný nástroj nejen pro běžné uživatele, ale i pro odborníky z oblasti práva, ochrany soukromí nebo vývoje digitálních služeb.

Tabulka 2 Výhody navrhovaného řešení

	ToS;DR	Navrhované řešení
Způsob analýzy	Ruční analýza komunitou	Plně automatizovaná pomocí LLM (Large Language Model)
Možnost vlastní služby	Není možnost analýzy vlastních Terms of Use	Ano, je možná analýza vlastní služby a vlastních Terms of Use
Typ výstupu	Předem připravené shrnutí služeb	Vlastní výstup na základě vlastního promptu
Jazyková flexibilita	Pouze anglicky	Možnosti vícejazyčného vstupu

15 IMPLEMENTACE ŘEŠENÍ POMOCÍ CHATGPT API

Po zajištění a předzpracování vstupních dat přichází na řadu samotná analýza pomocí umělé inteligence. V rámci tohoto projektu byla k tomuto účelu využita technologie od společnosti OpenAI, konkrétně jazykový model ChatGPT, ke kterému se přistupuje přes veřejné API. Pro připojení a práci s tímto rozhraním byla využita oficiální knihovna `openai` pro Python, která umožňuje jednoduché a přehledné odesílání požadavků a zpracování odpovědí.

```
from openai import OpenAI
client = OpenAI(api_key="sk-proj-...")
```

Ukázka kódu 1 import knihovny, napojení na API key pomocí knihovny (zdroj: vlastní)

ChatGPT pracuje s principem zpráv v rámci konverzace. Každý požadavek se skládá z jedné nebo více zpráv, které mají přidělenou roli (např. `user`, `system`, `assistant`). V tomto projektu je využíván jeden hlavní vstup, který má roli uživatele (`user`) a obsahuje sestavený `prompt`. Tento `prompt` se skládá ze dvou částí: první obsahuje seznam bodů nebo shrnutí smluvních podmínek a druhou část tvoří samotné zadání od uživatele, tedy co přesně chce model s těmito daty udělat.

```
completion = client.chat.completions.create(
    model="gpt-4o-mini",
    messages=[{"role": "user", "content": final_prompt}]
)
gpt_output_text = completion.choices[0].message.content
```

Ukázka kódu 2 Odeslání promptu a získání odpovědi od modelu, nastavení rolí (zdroj: vlastní)

Výhodou tohoto přístupu je, že uživatel není nijak omezen. Může si sám napsat, jak má analýza vypadat, například „Vypiš hlavní výhody a nevýhody“, „Upozorni na rizika pro uživatele“ nebo „Zhodnot' celkové podmínky známkou jako ve škole“. Díky tomu si každý může přizpůsobit výstup podle svých potřeb a model na to dokáže reagovat velmi přesně.

V rámci projektu byl využit model `gpt-4o-mini`, který byl v době vývoje dostupný za nízkou cenu, ale přitom nabízel velmi dobré výsledky. Tento model je schopný generovat odpovědi s vysokou jazykovou kvalitou, zvládá složitější zadání a je dostatečně rychlý i pro běžné použití. Výsledná odpověď je automaticky načtena do uživatelského rozhraní, kde si ji uživatel může pohodlně přečíst, případně ji porovnat s oficiálním hodnocením služby, pokud je dostupné.

Je ale důležité zmínit, že kromě modelu gpt-4o-mini existují i jiné, často výkonnější jazykové modely. Například plná verze modelu gpt-4, která má lepší porozumění složitějším textům, nebo další alternativy od jiných firem, jako například Claude od Anthropic, Gemini od Googlu, Mistral nebo Meta LLaMA. Většina těchto modelů je ale dostupná pouze za poplatek, případně vyžaduje složitější integraci. Pro tento projekt však byla zvolena varianta, která je cenově dostupná, a přitom dostatečně výkonná. Do budoucna by bylo určitě zajímavé tyto modely vyzkoušet a porovnat, zda by výsledky analýzy nebyly ještě přesnější a kvalitnější.

Samotné zpracování požadavku probíhá tak, že uživatel v grafickém rozhraní aplikace zadá název služby nebo vloží vlastní text podmínek. Dále napíše prompt, tedy zadání pro model a klikne na tlačítko pro spuštění analýzy. Aplikace následně sestaví finální text pro model, odešle ho přes API, počká na odpověď a tu pak zobrazí uživateli.

Důležitou součástí implementace je také správa historie všech požadavků a odpovědí. Aplikace při každé analýze automaticky ukládá zadaný název služby, prompt, vstupní text i výstup modelu do souboru history.json. Uživatel si pak v levé části aplikace může kdykoliv zobrazit seznam všech předchozích analýz, vybrat konkrétní záznam a znovu si zobrazit výstup. Tato funkce výrazně zvyšuje přehlednost, umožňuje snadné porovnání více analýz a šetří čas v případech, kdy je třeba se k předchozím výsledkům vrátit nebo je zrevidovat. Každý záznam je navíc možné ze seznamu ručně smazat.

```
entry = {
    "id": str(uuid.uuid4()),
    "service_name": service_name,
    "prompt": prompt,
    "input_text": input_text,
    "gpt_result": gpt_result,
    "tosdr_data": tosdr_data
}
history.append(entry)
save_history()
```

Ukázka kódu 3 Uložení dat o provedené analýze do historie (zdroj: vlastní)

V průběhu implementace bylo zároveň potřeba ošetřit různé chybové stavy, například pokud služba nebyla nalezena v databázi ToS;DR, pokud neměla žádné hodnotitelné body, nebo pokud došlo k chybě při volání API. Všechny tyto situace jsou uživateli jasně

oznámeny pomocí hlášek v rozhraní a aplikace se s nimi umí vypořádat tak, aby nespadla nebo nezůstala viset.

Celkově se tedy jedná o poměrně jednoduché, ale funkčně silné napojení na jazykový model, které umožňuje zpracovat a zhodnotit téměř libovolný text smluvních podmínek ať už z veřejné databáze nebo z vlastní hlavy uživatele. Odpovědi, které model vrací, jsou nejen srozumitelné, ale často i překvapivě praktické a konkrétní, což z něj dělá užitečný nástroj nejen pro běžné použití, ale i pro testování, porovnávání nebo tvorbu vlastních podmínek.

15.1 Uživatelské rozhraní a ovládání aplikace

Aplikace byla navržena tak, aby byla přehledná, intuitivní a umožňovala pohodlnou práci s analýzou smluvních podmínek. Uživatelské rozhraní je rozděleno do tří hlavních částí:

Levá část (sloupec) představuje historii transkripcí, kde jsou uvedeny všechny dříve analyzované služby nebo dokumenty. Uživatel zde může vybrat konkrétní záznam a zobrazit jeho výstup.

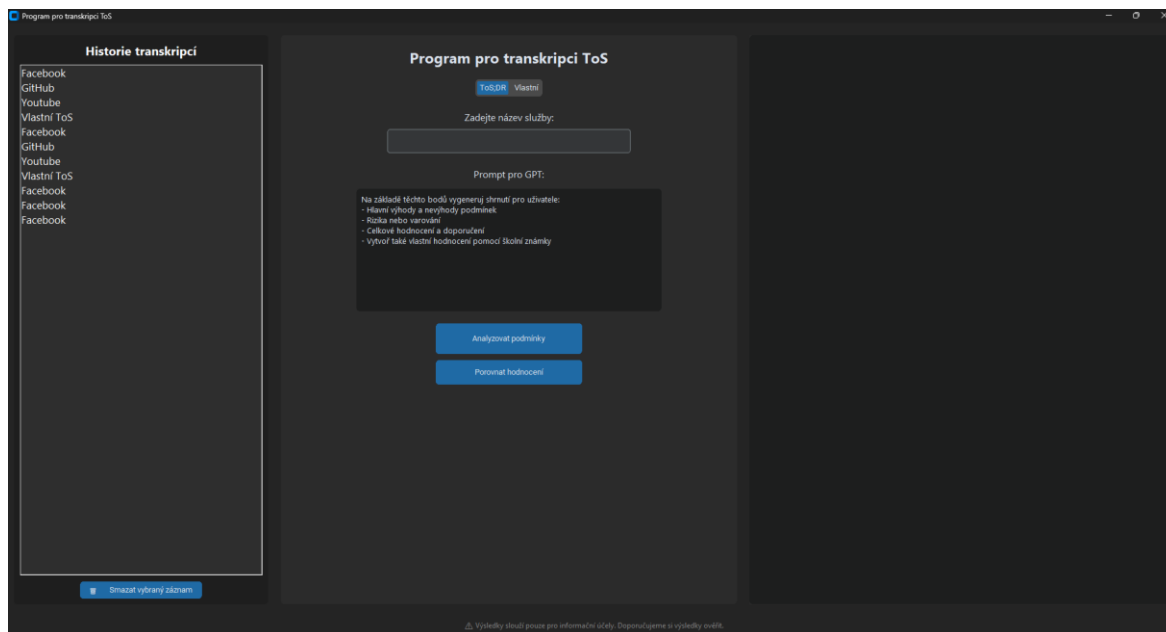
Střední část slouží k zadávání vstupních údajů a promptů. Uživatel zde volí mezi dvěma režimy: buď analyzuje veřejně známou službu (přes ToS;DR), nebo vkládá vlastní podmínky (textově či z PDF). Pod vstupem se nachází textové pole pro zadání vlastního promptu a dvě tlačítka: jedno pro spuštění analýzy, druhé pro porovnání s oficiálními body ToS;DR.

Pravá část zobrazuje výstup generovaný jazykovým modelem. Po spuštění analýzy se zde zobrazí přehledné shrnutí hlavních výhod, nevýhod, rizik, doporučení a celkového hodnocení. Výstupy jsou barevně zvýrazněny dle klasifikace (zeleně „good“, červeně „bad“, oranžově „neutral“ a „warning“), což usnadňuje rychlou orientaci v obsahu.

Součástí aplikace je také samostatné okno pro porovnání, které umožňuje zobrazit oficiální body z ToS;DR vedle odpovědi modelu a porovnat tak jejich obsah a klasifikaci.

Aplikace navíc podporuje plnou historii požadavků, možnost smazání záznamů a načtení podmínek z PDF dokumentu. To činí nástroj vhodným jak pro běžné uživatele, tak pro testování složitějších právních dokumentů.

Na spodní části aplikace se zároveň nachází zřetelné upozornění (disclaimer), že výstupy nástroje nepředstavují právní radu. Jde o automatizované shrnutí, které může obsahovat chyby, a uživatelům je doporučeno vždy si prostudovat originální znění podmínek použití.



Obrázek 4 Ukázka vzhledu aplikace (zdroj: vlastní)

16 ROZŠÍŘENÍ NA DALŠÍ WEBOVÉ STRÁNKY A PODMÍNKY UŽITÍ

Navržený systém byl původně vytvořen s cílem analyzovat podmínky užití služeb evidovaných v databázi ToS;DR, případně podmínek zadaných uživatelem ručně. Již během vývoje se však ukázalo, že navržená architektura je dostatečně flexibilní a univerzální, aby mohla být v budoucnu snadno rozšířena o další zdroje dat, typy dokumentů a možnosti použití.

Díky tomu, že celý systém je postaven na principu převodu strukturovaných nebo nestrukturovaných podmínek užití do textové formy zpracovatelné velkým jazykovým modelem, není vázán pouze na jednu konkrétní databázi. Veškeré vstupy jsou zpracovávány jako čistý text, buď získaný z API, nebo zadávaný uživatelsky, a poté odesílány do modelu ChatGPT, který generuje požadovanou výstupní analýzu. To znamená, že jakmile je zachována základní struktura vstupu, lze stejný mechanismus aplikovat i na jiné zdroje.

16.1 Možnost ručního vkládání podmínek

Jednou z klíčových funkcí, která tuto rozšiřitelnost umožňuje, je podpora vlastního zadání podmínek přímo uživatelem. Uživatel má možnost vložit libovolný text smluvních podmínek jakékoli webové stránky, i takové, která není uvedena v databázi ToS;DR. Tím je umožněna analýza nejen známých a frekventovaných služeb, ale i menších, nově vzniklých nebo méně transparentních projektů, jejichž podmínky by jinak zůstaly neanalyzované.

Tato funkce je integrovaná v uživatelském rozhraní a zajišťuje, že analýza je dostupná pro široké spektrum dokumentů, ať už se jedná o výňatek z podmínek, celý dokument nebo kombinaci několika typů informací. Tímto způsobem je systém připraven k použití i v prostředích, kde se běžně pracuje s interními nebo neveřejnými podmínkami, např. ve firmách, školách nebo vývojových týmech.

Kromě ručního vložení textu je také možné nahrát soubor ve formátu PDF, ze kterého si systém automaticky vytáhne potřebný text. Díky tomu je práce s většími dokumenty výrazně jednodušší a rychlejší. Uživatel tedy nemusí ručně kopírovat dlouhé podmínky, ale může je pohodlně načíst přímo z oficiálního dokumentu. Tento způsob vkládání rozšiřuje možnosti využití nástroje i na složitější nebo oficiální právní texty, které jsou běžně poskytovány ve formátu PDF.

16.2 Možnost automatizovaného načítání z jiných zdrojů

V případě potřeby může být systém rozšířen o schopnost automaticky stahovat podmínky užívání z veřejných webových stránek. V praxi by to mohlo být realizováno například pomocí jednoduchého crawleru, který by zadanou URL adresu navštívil, vyhledal část stránky obsahující podmínky a tuto část předal ke zpracování.

Další možností je využití specializovaných služeb, které shromažďují podmínky a zásady ochrany osobních údajů a často nabízejí export ve formátech vhodných ke zpracování. Data z těchto služeb by mohla být předána stejným způsobem jako data z ToS;DR, případně po převedení do přirozeného textu.

Tato forma rozšíření by mohla přinést výrazné zvýšení pokrytí dostupných dokumentů a zároveň umožnit hromadnou analýzu většího počtu služeb v krátkém čase, například v rámci výzkumu, auditu nebo srovnání služeb podle zaměření.

16.3 Podmínky mimo oblast technologií

Zatímco současná podoba systému je zaměřena především na online platformy a digitální služby, obdobný přístup by bylo možné uplatnit i na podmínky z jiných oblastí. Příkladem mohou být:

- všeobecné obchodní podmínky e-shopů,
- smlouvy o poskytování služeb (např. mobilní operátoři, banky, pojišťovny),
- licenční smlouvy k softwaru a aplikacím,
- zásady ochrany osobních údajů.

Uvedené typy dokumentů mají často podobnou strukturu jako Terms of Use a jazykové modely jsou schopné i tyto texty analyzovat, sumarizovat, nebo zvýraznit rizikové části. Vzhledem k tomu, že model nevyžaduje konkrétní formát vstupu a pracuje čistě s přirozeným jazykem, není nutné žádné zásadní technické přizpůsobení.

16.4 Využití v právní a vzdělávací praxi

Rozšíření tohoto nástroje se nabízí také ve vzdělávacích nebo právních kontextech. Studenti práv, pedagogové nebo školitelé v oblasti ochrany osobních údajů by mohli tento nástroj využít k rychlé orientaci v právních textech, jejich porovnávání nebo jako pomůcku pro výuku. Umožněna by byla i tvorba výukových verzí podmínek, např. „zjednodušené

podmínky pro běžného uživatele“, které by byly vygenerovány na základě detailních právních dokumentů.

Podobné použití se nabízí i v právní praxi, kde může nástroj sloužit jako rychlý pomocník pro orientační rozbor dokumentu, identifikaci slabých míst nebo generování podkladů pro další manuální analýzu. V kombinaci s dalšími nástroji pro zpracování přirozeného jazyka by mohl sloužit jako součást právní AI asistence.

16.5 Budoucí směry rozšíření

Do budoucna se nabízí řada možných rozšíření, která by mohla zvýšit hodnotu a použitelnost nástroje. Patří sem například:

- podpora více jazyků (např. automatická detekce a překlad vstupních podmínek),
- klasifikace textu podle typu rizika (např. sledování chování, sdílení dat, odpovědnost),
- rozšíření vstupů o uživatelské hodnocení nebo zpětnou vazbu,
- automatické generování doporučení nebo alternativních formulací.

Systém byl navržen tak, aby se změnám tohoto typu přizpůsobil, a to jak z pohledu technického řešení (architektura, formát vstupu), tak i z pohledu uživatelského použití (možnost doplnit vlastní data, jednoduché ovládání).

17 PREZENTACE VÝSLEDKŮ A VYHODNOCENÍ ŘEŠENÍ

V této kapitole jsou detailně vyhodnoceny výsledky analýz smluvních podmínek provedených pomocí jazykového modelu ChatGPT. Cílem bylo posoudit, jak si model poradí s různými vstupy, a to jak ze známých databází (např. ToS;DR), tak z vlastních textů zadaných uživatelem. Kromě toho byla testována i schopnost modelu reagovat na různé formulace promptů, tedy zadání, podle kterého se má výstup generovat. Zohledněna byla jak srozumitelnost a relevance výstupů, tak i možnosti přizpůsobení výstupu konkrétním potřebám uživatele.

17.1 Použití výchozího promptu

Ve většině testovacích scénářů byl použit jednotný výchozí prompt:

- Na základě těchto bodů vygeneruj shrnutí pro uživatele:
 - a. Hlavní výhody a nevýhody podmínek
 - b. Rizika nebo varování
 - c. Celkové hodnocení a doporučení
 - d. Vytvoř také vlastní hodnocení pomocí školní známky

Tento prompt měl za cíl vygenerovat ucelené shrnutí zaměřené na praktické využití, tedy upozornit uživatele na silné a slabé stránky podmínek, přidat doporučení a uzavřít výstup jednoduchým známkováním, které je snadno srozumitelné.

17.2 Výstupy založené na výchozím promptu

Model byl testován na několika službách: GitHub, Facebook a YouTube. Ve všech případech byl pomocí API načten seznam tvrzení z ToS;DR a následně analyzován.

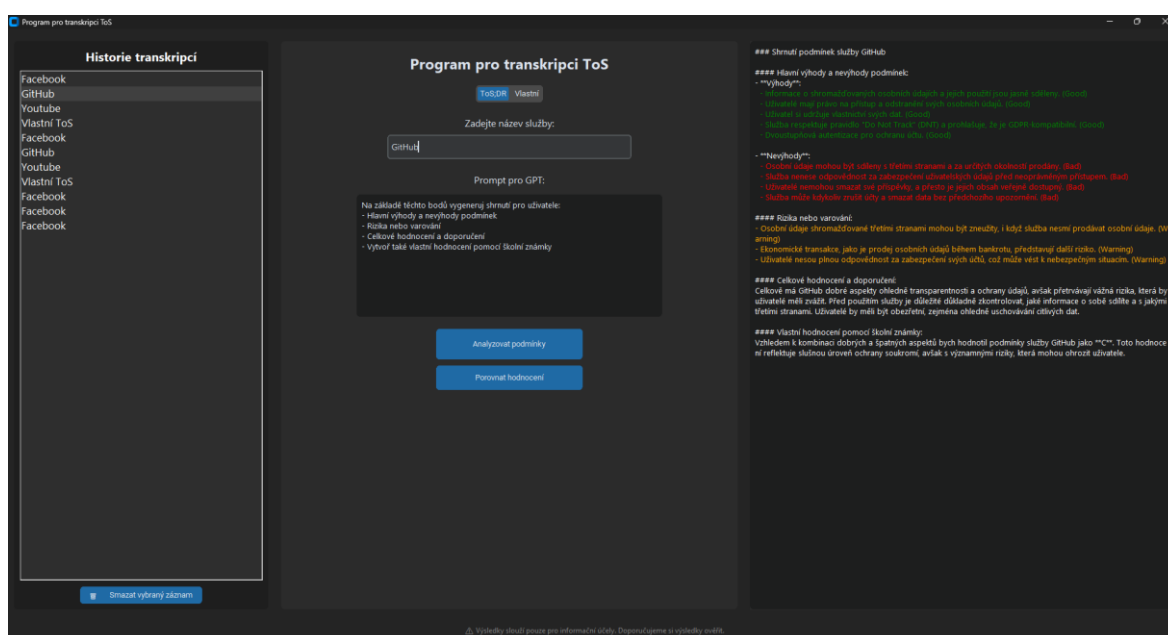
17.2.1 Github

Model při analýze GitHubu vyzdvihl několik pozitiv. Mezi nejvýznamnější výhody patřilo jasné vymezení způsobu nakládání s osobními údaji, možnost exportu dat, podpora „Do Not Track“ a dvoufaktorová autentizace. všechny tyto body byly označeny zeleně jako „good“, tedy pozitivní.

Na druhé straně model označil jako problematické zejména možnost sdílení údajů s třetími stranami, absenci odpovědnosti za zabezpečení údajů a složitost při smazání účtu. Tyto byly označeny červeně („bad“), což značí potenciální riziko pro uživatele.

Oranžově („warning“) pak model vyznačil rizika spojená s ekonomickými transakcemi (například prodej dat během bankrotu), s nedostatečnou ochranou účtů a s možnostmi zneužití údajů třetími stranami.

Výsledné hodnocení podmínek služby bylo „C“, tedy střední úroveň, která představuje kombinaci přínosů a hrozeb, se silným důrazem na nutnost obezřetnosti při používání služby.



Obrázek 5 Výstup analýzy služby Github s využitím výchozího promptu (zdroj: vlastní)

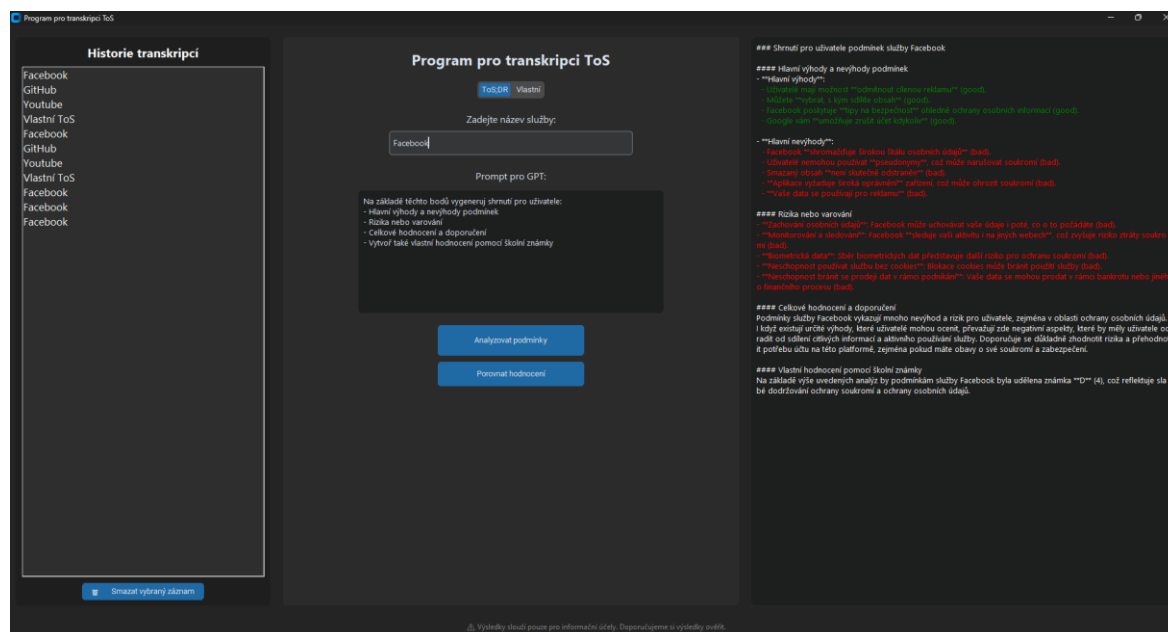
17.2.2 Facebook

Facebook byl hodnocen jako nejproblematictější z analyzovaných služeb. Pozitivně (zeleně) byly vyhodnoceny pouze možnosti jako „odmítnout cílenou reklamu“, „možnost výběru sdílení obsahu“ nebo informace o možnosti zrušit účet. I u těchto bodů však bylo poznamenáno, že jde spíše o základní funkce.

Negativa (červeně) byla velmi rozsáhlá: zahrnovala široký sběr osobních údajů, sledování napříč weby, nutnost používat skutečné jméno, nejasné podmínky mazání dat, biometrická data, a dokonce i možnost obchodování s daty v rámci firemních transakcí.

Rizika byla podrobně popsána a označena rovněž červeně, zejména kvůli vysoké míře sledování a zpracování citlivých údajů bez dostatečné transparentnosti.

Celkové hodnocení bylo vyjádřeno známkou „D“, což odpovídá vysoké míře rizika a nízké úrovni ochrany uživatele.



Obrázek 6 Analýza výstupu služby Facebook pomocí výchozího promptu (zdroj: vlastní)

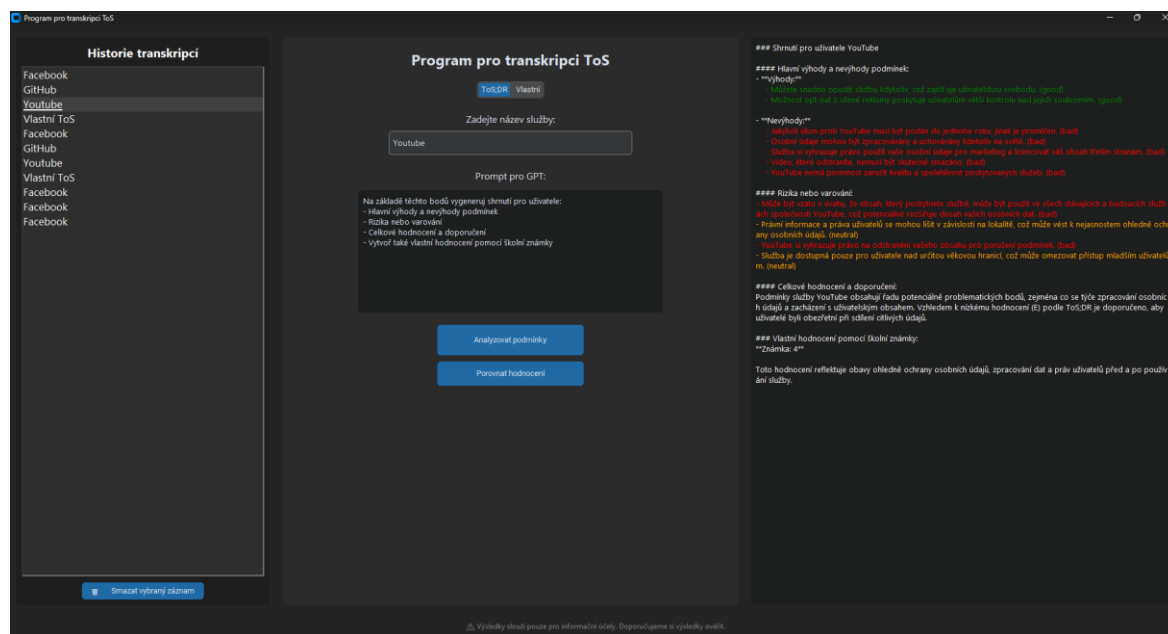
17.2.3 Youtube

U služby YouTube byly mezi pozitivní (zeleně označená) zařazeny především možnost kdykoliv opustit platformu a odhlásit se z cílené reklamy, což přispívá k ochraně soukromí a uživatelské svobodě.

Mezi negativa (červeně) model zařadil například fakt, že jakékoliv žaloby musí být podány do jednoho roku, že data mohou být zpracovávána mimo EU, a že služba si ponechává trvalá práva na uživatelský obsah. Také upozornil na to, že YouTube odmítá odpovědnost za služby třetích stran.

V části rizik byly některé body označeny oranžově jako varování (neutral/warning), například právní nejistota ohledně uživatelských práv nebo omezení dostupnosti služby mladším uživatelům v závislosti na lokalitě.

Hodnocení podmínek bylo stanoveno na „4“ (známka školní stupnice), což odráží přítomnost několika závažných problémů s ochranou údajů a s právem na vlastní obsah.



Obrázek 7 Výstup analýzy služby Youtube pomocí výchozího promptu (zdroj: vlastní)

17.3 Alternativní prompt a jeho vliv na výstup

Pro ověření vlivu formulace zadání (promptu) na kvalitu výstupu modelu byl připraven alternativní prompt, zaměřený na právní analýzu smluvních podmínek jako celku. Jeho znění bylo následující:

„Zhodnot' následující podmínky služby z právního hlediska jako celek. Neprováděj hodnocení jednotlivých bodů samostatně a nevypisuj žádný seznam bodů s klasifikací. Místo toho vytvoř celkovou právní analýzu, která odpoví na následující otázky:

- *Jaká jsou potenciální rizika pro ochranu soukromí uživatele?*
- *Obsahují podmínky nevyvážená ustanovení nebo právní nepřesnosti?*
- *Jaký je přístup služby k odpovědnosti a právům uživatele?*
- *Doporučil bys právní úpravy nebo změny v textu?*

Výsledek uveď pouze jako strukturované shrnutí, ne jako bodové hodnocení. “

Tento prompt měl za cíl nasměrovat model ke komplexnější a právně orientované interpretaci podmínek užívání. Na rozdíl od výchozího promptu, který vedl k výstupům ve formě seznamů výhod, nevýhod a doporučení, alternativní verze umožnila hlubší a systematickou analýzu celého dokumentu. Výsledné texty byly delší, preciznější a

vykazovaly vyšší míru odbornosti, například v identifikaci nevyvážených ustanovení nebo rozboru právních důsledků formulací týkajících se odpovědnosti, přenosu dat nebo souhlasu.

17.4 Pozorované rozdíly mezi prompty

Aby bylo možné objektivně posoudit vliv rozdílné formulace promptu, byly porovnány výstupy modelu získané pomocí výchozího (uživatelsky orientovaného) a alternativního (právně orientovaného) zadání. Výchozí prompt kladl důraz na přehlednost a srozumitelnost pro běžného uživatele, tedy shrnutí výhod, nevýhod, rizik a celkové doporučení ve formě strukturovaného seznamu. Alternativní prompt byl formulován tak, aby výstup představoval celkovou právní analýzu bez seznamů, bez bodového hodnocení, ale s důrazem na odpovědi na konkrétní právní otázky a strukturované shrnutí celého dokumentu.

Výsledky ukázaly, že změna promptu výrazně ovlivňuje charakter a hloubku odpovědi. Model se při použití alternativního zadání soustředil na jazykovou přesnost, právní kvalifikace, asymetrii mezi uživatelem a službou a celkové postavení uživatele v dokumentu. Identifikoval i skryté právní problémy, jako je nejasné vymezení odpovědnosti, nedostatečná ochrana dat nebo jednostranná oprávnění poskytovatele.

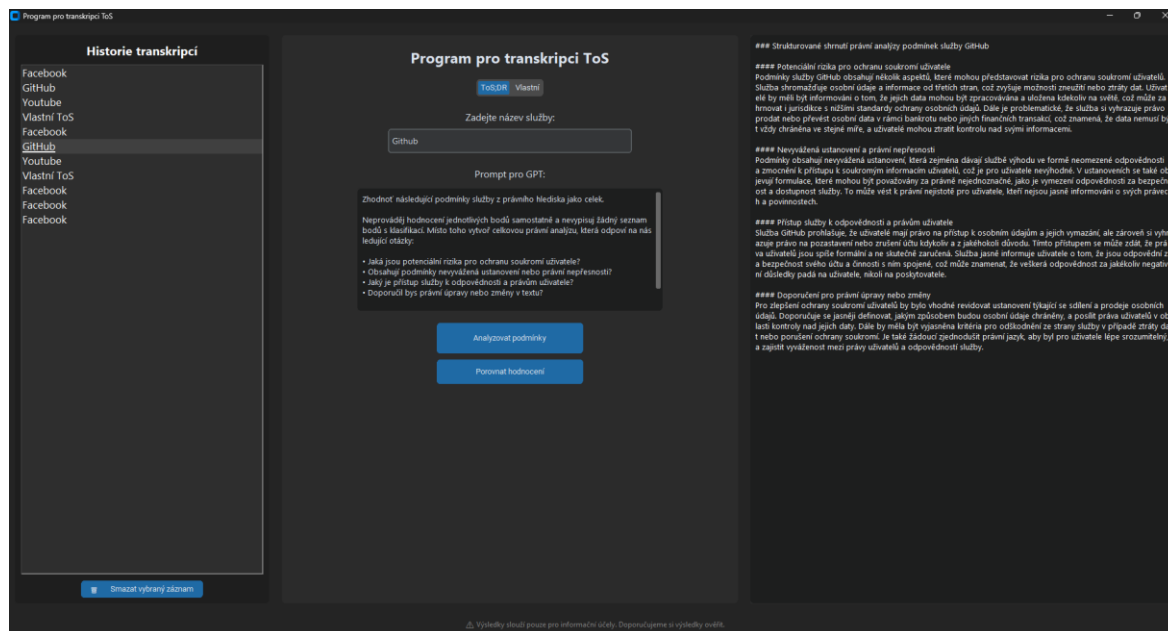
Naopak u výchozího promptu model generoval odpovědi více přístupné pro laického uživatele, což je výhodné pro rychlé orientační použití. Alternativní zadání je tak vhodnější v případech, kdy je požadována odbornější interpretace, například při interním právním posouzení nebo při přípravě revize dokumentu.

17.4.1 Github

Právní analýza podmínek služby GitHub odhalila několik oblastí, které mohou představovat riziko pro uživatele z hlediska ochrany soukromí. Především jde o možnost předávání údajů třetím stranám, riziko zneužití při prodeji dat nebo jejich zpracování mimo EU a absenci jasně vymezené odpovědnosti poskytovatele. Model poukázal na to, že některé formulace týkající se práv uživatelů, například na smazání dat či na přístup k údajům, nejsou dostatečně závazné a přenášejí odpovědnost plně na uživatele.

Za problematické lze považovat také některé právní nepřesnosti, především v oblasti zřeknutí se odpovědnosti ze strany služby. GitHub v podmínkách uvádí, že si vyhrazuje právo ukončit účet kdykoliv a z jakéhokoliv důvodu, což oslabuje právní jistotu uživatele. Přístup k právům uživatele je podle modelu nevyvážený, podmínky působí jednostranně a spíše chrání poskytovatele před odpovědností.

Model doporučil, aby GitHub přehodnotil zejména části týkající se sdílení dat a vyjasnil pravidla pro ukončení účtu. Také navrhl přesnější vymezení odpovědnosti v případech porušení podmínek a zavedení mechanismů pro efektivní řešení sporů.



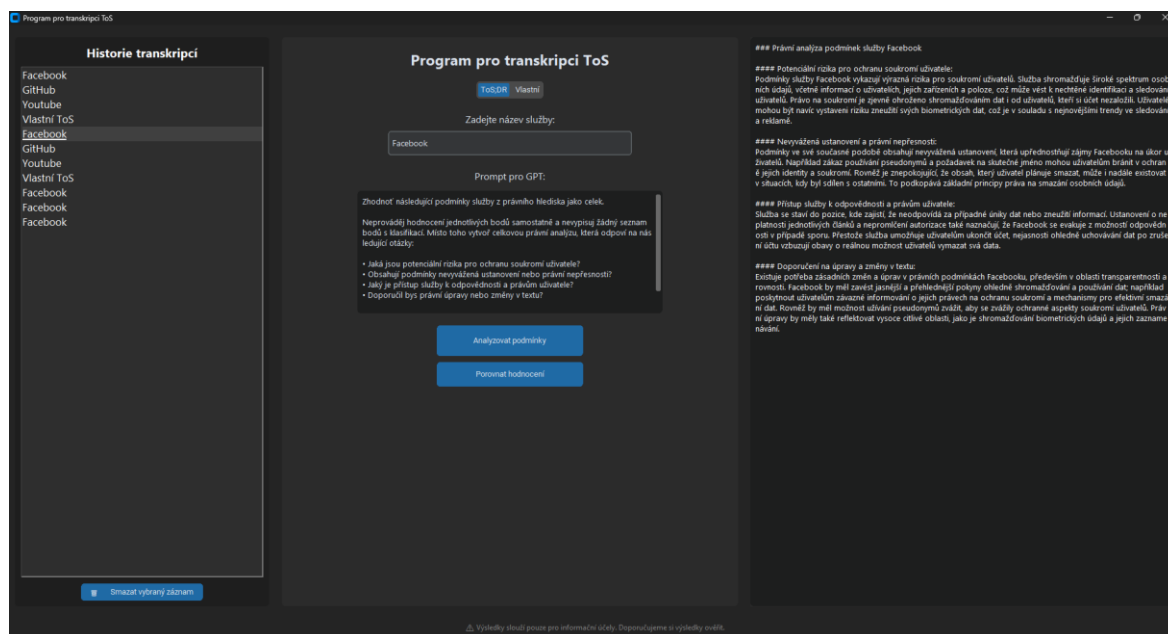
Obrázek 8 GitHub analýza – alternativní prompt (zdroj: vlastní)

17.4.2 Facebook

Podmínky služby Facebook byly označeny jako právně problematické zejména v oblasti ochrany soukromí a rovnováhy práv. Model identifikoval vysokou míru rizika, které plyne z rozsáhlého sběru dat, včetně údajů o poloze a chování mimo samotnou platformu. Zejména znepokojující je fakt, že Facebook vytváří tzv. „shadow profiles“ i o lidech, kteří účet nemají.

Největší právní slabiny se podle analýzy týkají nevyvážených ustanovení, která nahrávají zájmům platformy, například povinnost uvádět skutečné jméno, omezené možnosti smazání dat nebo přenášení odpovědnosti výhradně na uživatele. Také ustanovení o odpovědnosti platformy jsou neurčitá, bez jasné definice právní zodpovědnosti za případné zneužití dat.

Model navrhl několik konkrétních právních úprav, zejména umožnění používání pseudonymů, zavedení přesných pravidel pro výmaz dat, jasnou definici ochrany citlivých údajů a zajištění transparentního informování uživatelů o jejich právech. Zásadní je také posílení kontroly nad tím, jakým způsobem Facebook zachází s informacemi uživatelů mimo vlastní platformu.



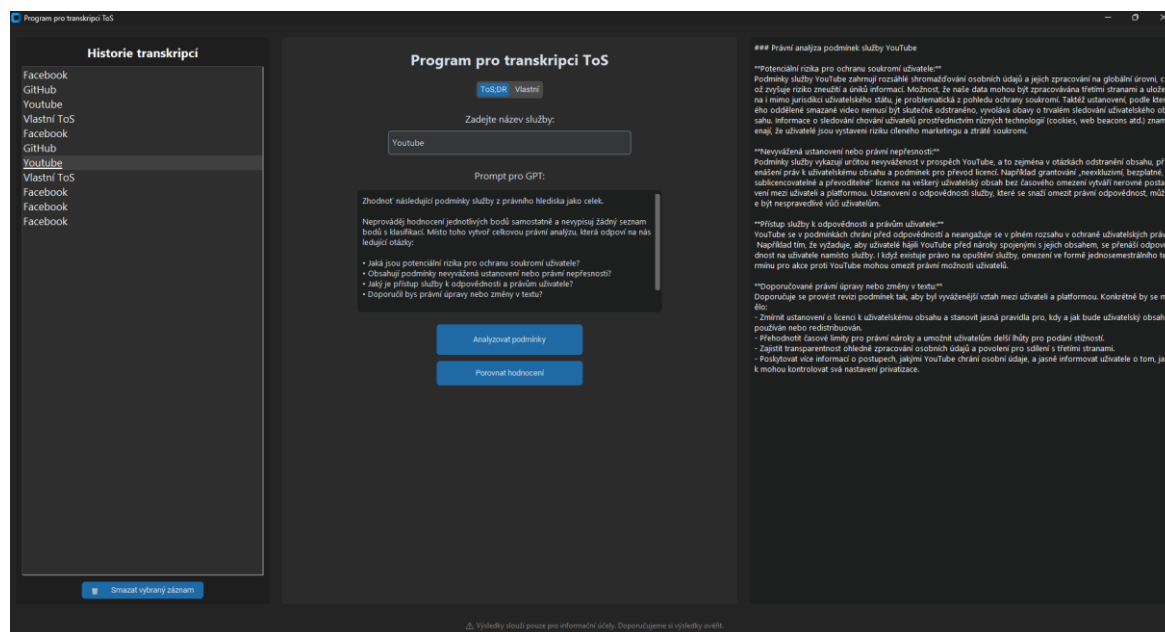
Obrázek 9 Facebook analýza – alternativní prompt (zdroj: vlastní)

17.4.3 Youtube

U YouTube byly hlavní právní rizika spojena s globálním přenosem dat a rozsáhlým sledováním uživatelů. Model upozornil na to, že zpracování dat probíhá na základě obecně formulovaných ustanovení, která nevymezují dostatečně právní rámec pro přenos údajů do jiných jurisdikcí. Rizikové je i uchovávání obsahu po smazání a licenční podmínky, které dávají YouTube široké právo s uživatelským obsahem nakládat.

Podmínky služby navíc obsahují vágní formulace týkající se odpovědnosti např. že YouTube nenese odpovědnost za obsah, ale zároveň si ponechává právo na jeho odstranění. Tento přístup může být právně sporný a oslabuje postavení uživatele. Chybí také přesné lhůty pro podání stížnosti nebo požadavky na transparentnost zpracování dat.

Model doporučil několik zásadních právních úprav: jasné definování licencí a práv k obsahu, rozšíření lhůt pro nápravu nebo stížnosti, zajištění kontroly nad sdílením dat a přesnější jazyk v části o odpovědnosti. Celkově bylo doporučeno posílit právní ochranu uživatelů, především v otázkách odpovědnosti a spravedlivého přístupu.



Obrázek 10 Youtube analýza – alternativní prompt (zdroj: vlastní)

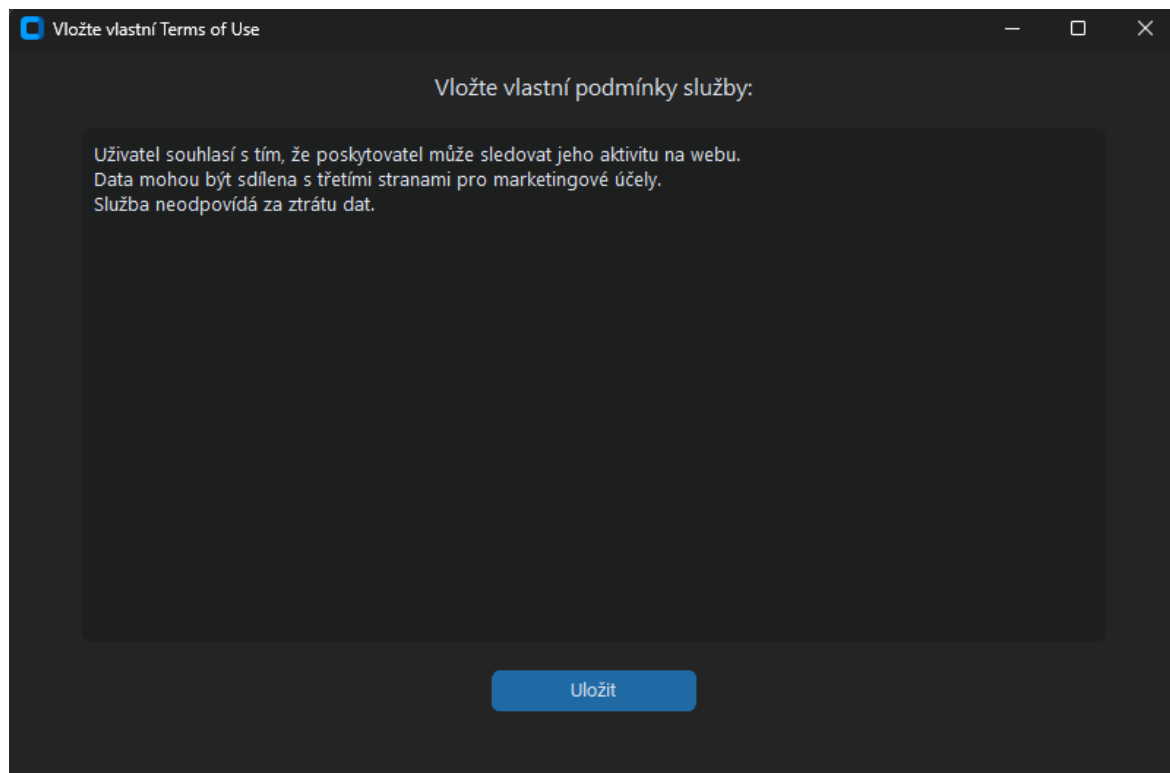
17.5 Analýza vlastního dokumentu a přizpůsobení výstupu

Součástí testování byla také možnost analýzy podmínek užití, které byly zadány přímo uživatelem bez využití veřejných databází. Tento scénář simuluje reálnou situaci, kdy má uživatel k dispozici vlastní smluvní dokument nebo výňatek z podmínek, který chce podrobit automatickému zhodnocení. V aplikaci lze tyto dokumenty zadat dvěma způsoby, buď ručním vložením textu, nebo nahráním souboru ve formátu PDF. Díky této funkci je možné analyzovat i složitější dokumenty v jejich původní podobě bez nutnosti manuálního přepisování.

V rámci testování byl do aplikace zadán vlastní text prostřednictvím textového editoru (viz Obrázek 7) a následně analyzován pomocí výchozího promptu, který generuje výhody, nevýhody, rizika, celkové hodnocení a doporučení.

Ve výstupu (viz Obrázek 8) model vyzdvihl jednu výraznou výhodu, konkrétně možnost kdykoliv zrušit účet, kterou označil za pozitivní z hlediska kontroly uživatele nad vlastní účastí ve službě. Zároveň však identifikoval závažná rizika, zejména v oblasti ochrany soukromí a osobních údajů. Kriticky byla hodnocena skutečnost, že poskytovatel může sledovat aktivitu uživatele na webu a sdílet data s třetími stranami pro marketingové účely, aniž by bylo zřejmé, jaká opatření jsou přijata na jejich ochranu.

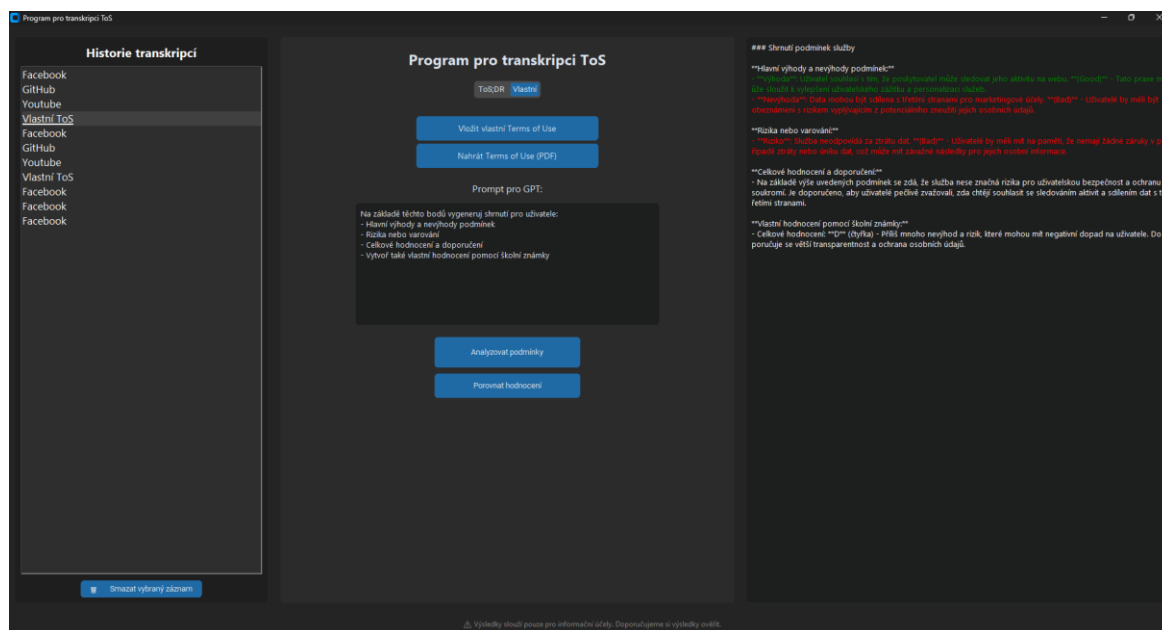
Jako hlavní nedostatek bylo označeno i to, že poskytovatel nenese odpovědnost za ztrátu dat, což může mít negativní dopad na důvěru uživatelů a potenciálně ohrozit jejich data v případě technických nebo bezpečnostních selhání.



Obrázek 11 Ukázka vlastních Terms of Use (zdroj: vlastní)

Výsledné hodnocení modelu bylo „D“, což odpovídá podmínkám s převahou nevýhod a rizik. V doporučení model vyzval k větší transparentnosti, lepší ochraně dat a přehlednějšímu vymezení odpovědnosti mezi uživatelem a poskytovatelem.

Tento výsledek potvrzuje, že je aplikace schopna efektivně pracovat i s vlastním, neznámým textem, včetně dokumentů ve formátu PDF, a upozornit na potenciálně problémové aspekty podmínek, které by jinak zůstaly skryty. Tato vlastnost činí řešení vhodným nejen pro analýzu známých služeb, ale také pro vnitřní audity firemních dokumentů nebo revizi právních textů před jejich zveřejněním.



Obrázek 12 Analýza vlastních Terms of Use (zdroj: vlastní)

17.6 Analýza vlastních ToS pomocí kreativního promptu

Pro ověření flexibility jazykového modelu a jeho schopnosti interpretovat podmínky v různém kontextu byl zadán kreativní prompt:

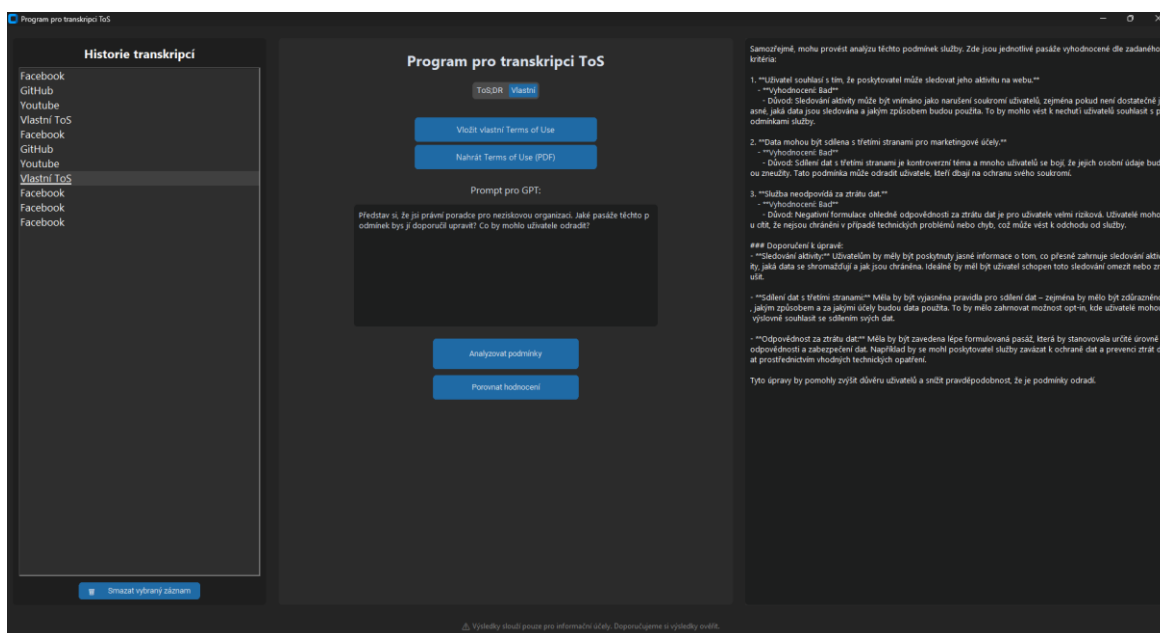
„Představ si, že jsi právní poradce pro neziskovou organizaci. Jaké pasáže těchto podmínek bys jí doporučil upravit? Co by mohlo uživatele odradit?“

Na rozdíl od standardního shrnutí model tentokrát vystupoval jako konzultant a analyzoval konkrétní části textu z hlediska jejich formulace a potenciálního negativního dopadu na uživatele. Výstup byl strukturován jako seznam doporučení s důrazem na úpravy jazykových formulací a zvýšení transparentnosti.

Model identifikoval následující problematické oblasti:

- Sledování aktivity uživatele: Doporučil upřesnit, jaké konkrétní údaje jsou sledovány, a zdůraznit možnost omezení sledování ze strany uživatele.
- Sdílení dat s třetími stranami: Navrhl zavést jasná pravidla, která by specifikovala účel sdílení, právní rámec a případně nabídla možnost opt-out.
- Odpovědnost za ztrátu dat: Upozornil na potřebu zavést jazyk, který by zavazoval poskytovatele služeb k ochraně dat a k přijetí konkrétních technických opatření.

Každá pasáž byla modelovým výstupem ohodnocena z hlediska důvěryhodnosti. Doporučené změny měly vést ke zvýšení důvěry uživatele, snížení obav z nedostatečné ochrany a lepšímu souladu s běžnými standardy ochrany soukromí. Tímto způsobem model demonstroval využitelnost nejen pro analýzu existujících dokumentů, ale také jako nástroj pro jejich cílenou právní revizi a vylepšení.



Obrázek 13 Vlastní Terms of Use – kreativní prompt (zdroj: vlastní)

17.7 Porovnání výstupů s oficiálním shrnutím z ToS;DR

Pro ověření přesnosti a relevance výstupů generovaných jazykovým modelem ChatGPT bylo provedeno přímé porovnání s oficiálními shrnutími z projektu ToS;DR. Tento komunitně spravovaný nástroj poskytuje ručně ověřená hodnocení konkrétních bodů z podmínek užívání řady internetových služeb. Každý bod je opatřen verdiktem (např. „dobrý“, „špatný“, „neutrální“) a tematicky zařazen (např. soukromí, bezpečnost, odpovědnost, vlastnictví dat). Díky barevnému označení bylo možné přímo vizuálně ověřit shodu nebo nesoulad s výstupy z GPT.

17.8 Postup porovnání

Pro každou službu (GitHub, Facebook, YouTube) byly pomocí veřejného API rozhraní ToS;DR načteny všechny dostupné hodnotitelné body („points“). Tyto body byly následně zobrazeny vedle odpovědi ChatGPT v přehledném okně a porovnány po jednotlivých tvrzeních. Cílem bylo zjistit, zda model identifikuje podobná témata a klasifikuje je ve shodě

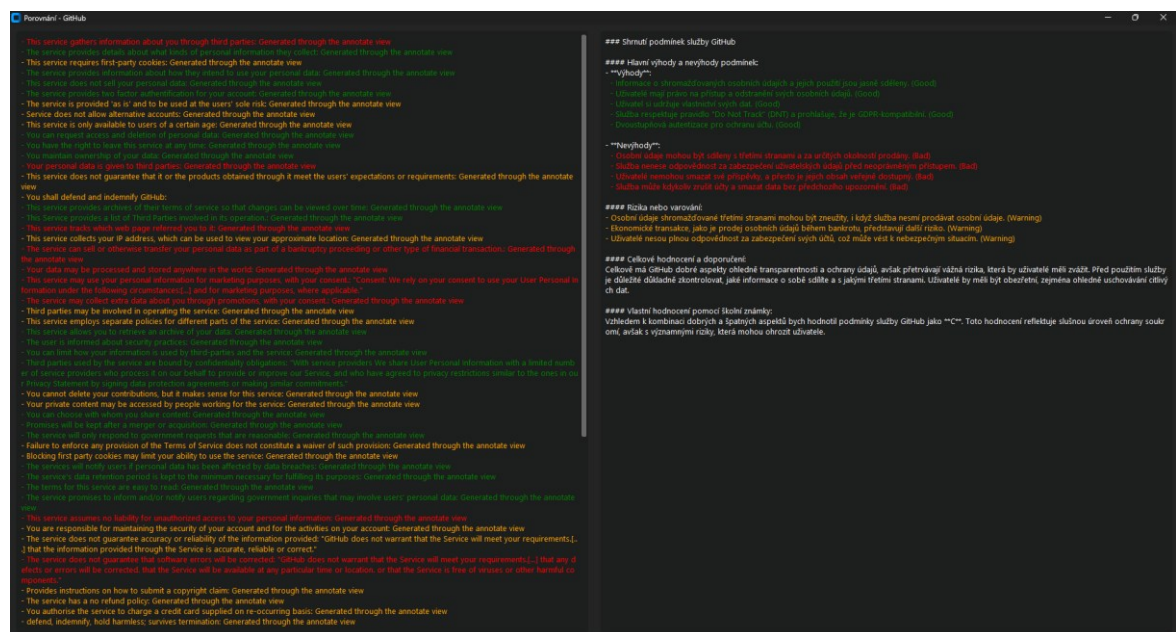
s oficiálním verdiktem. Porovnání bylo provedeno pro výstupy vytvořené pomocí výchozího promptu.

17.8.1 Github

GPT výstup správně zachytil následující prvky shodné s hodnocením ToS;DR:

- Upozornění na možnost sdílení dat se třetími stranami a ztrátu kontroly nad údaji (shoda s bodem "Third-party data sharing", hodnoceným jako „bad“).
- Uvedení možnosti exportu dat uživatelem (shoda s bodem „You can export your data“, hodnoceným jako „good“).
- Negativní vymezení odpovědnosti služby za ztrátu dat nebo škody (odpovídá verdiktu „Limited liability“, hodnocenému jako „bad“).
- Zmínka o sledovacích technologiích nebo cookies, i když méně explicitně.

Celkově model reflektoval všechny klíčové prvky identifikované v ToS;DR, přičemž některé body byly rozvedeny detailněji. Barevné rozlišení odpovídalo převážně původní klasifikaci.



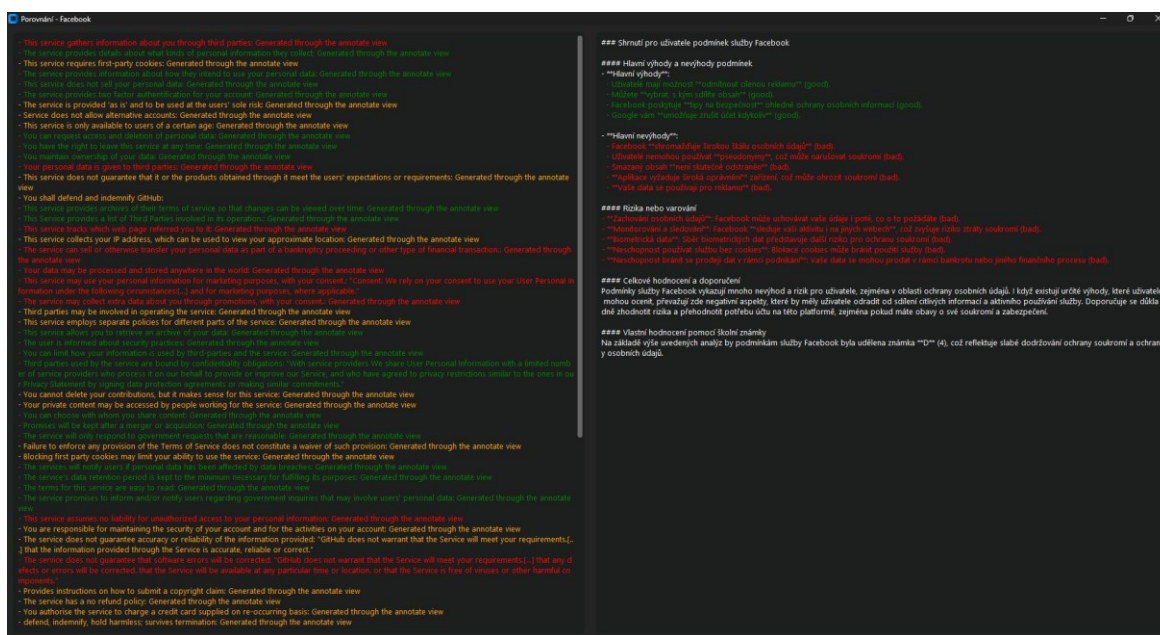
Obrázek 14 Porovnání bodů ToS;DR s výstupem modelu u službě Github (zdroj: vlastní)

17.8.2 Facebook

Výstupy modelu byly velmi dobře v souladu s oficiálním hodnocením ToS;DR:

- ChatGPT výslovně upozornil na rozsah shromažďovaných osobních údajů, sledování napříč službami, použití cookies a profilování. Tyto skutečnosti odpovídají několika verdiktům označeným jako „bad“ (např. „Tracking“, „Profiling“, „No deletion of data“).
- Dále upozornil na absenci práva na anonymitu a nutnost použití skutečného jména, což odpovídá negativnímu bodu v ToS;DR.
- Výstup zahrnoval i body, které v ToS;DR existují, ale jsou formulovány stručněji, například nemožnost skutečně odstranit data nebo omezit jejich použití.

Model tedy nejen vystihl hlavní rizika, ale přidal i praktický kontext k důsledkům, což zvyšuje uživatelskou srozumitelnost.



Obrázek 15 Porovnání bodů ToS;DR s výstupem modelu u služby Facebook (zdroj: vlastní)

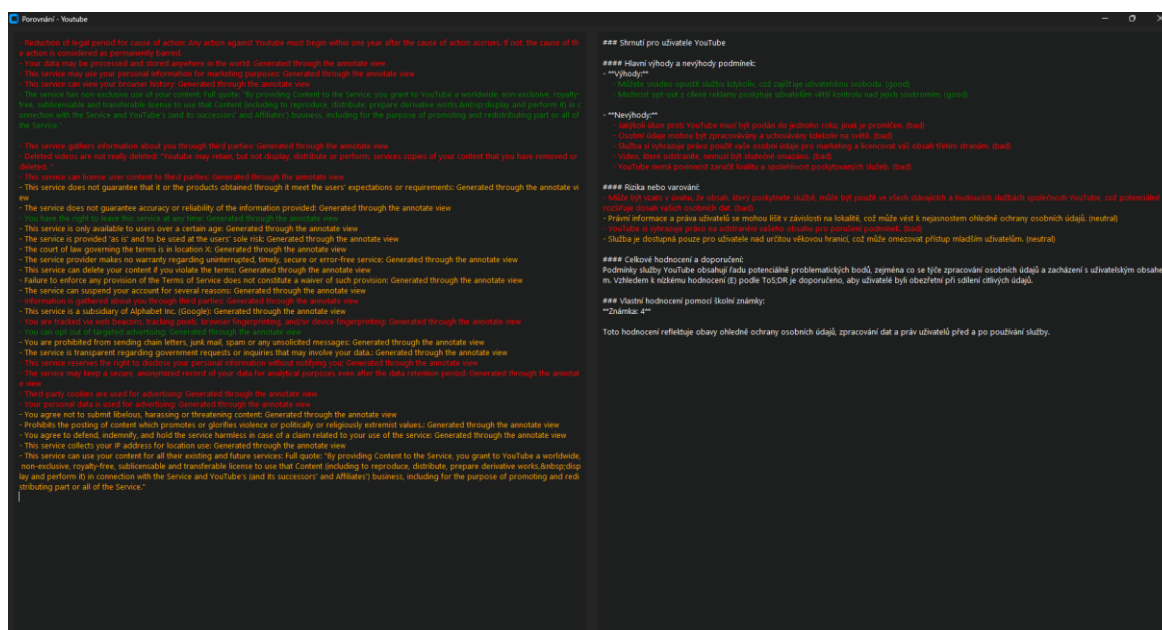
17.8.3 Youtube

Také v případě služby YouTube byla potvrzena vysoká shoda s body uvedenými v ToS;DR:

- GPT správně uvedl, že služba může využívat osobní údaje pro reklamní účely. To odpovídá několika negativním bodům (např. „Personal data used for ads“, „Cookies“).

- Bylo také identifikováno, že YouTube si ponechává smazaný obsah a může ho i nadále používat „Deleted content not really deleted“.
- Výstup rovněž obsahoval varování ohledně širokých licencí na uživatelský obsah, což odpovídá bodům typu „Broad license grant“.

Výstup modelu tak nejen odpovídal shrnutí z ToS;DR, ale v některých oblastech šel i více do hloubky. Upozornil na rizika, která jsou v ToS;DR pouze naznačena, čímž se ukázal jako užitečný nástroj pro doplňující analýzu.



Obrázek 16 Porovnání bodů ToS;DR s výstupem modelu u služby Youtube (zdroj: vlastní)

17.9 Časová proměnlivost výstupů

V rámci testování bylo cílem ověřit, zda jazykový model ChatGPT generuje konzistentní výstupy i při opakovaném spuštění se stejným vstupem v různou denní dobu. Za tímto účelem byla služba Facebook analyzována ve třech různých časech v 9:00, 15:00 a 19:00, přičemž byl zachován naprosto totožný vstup a prompt.

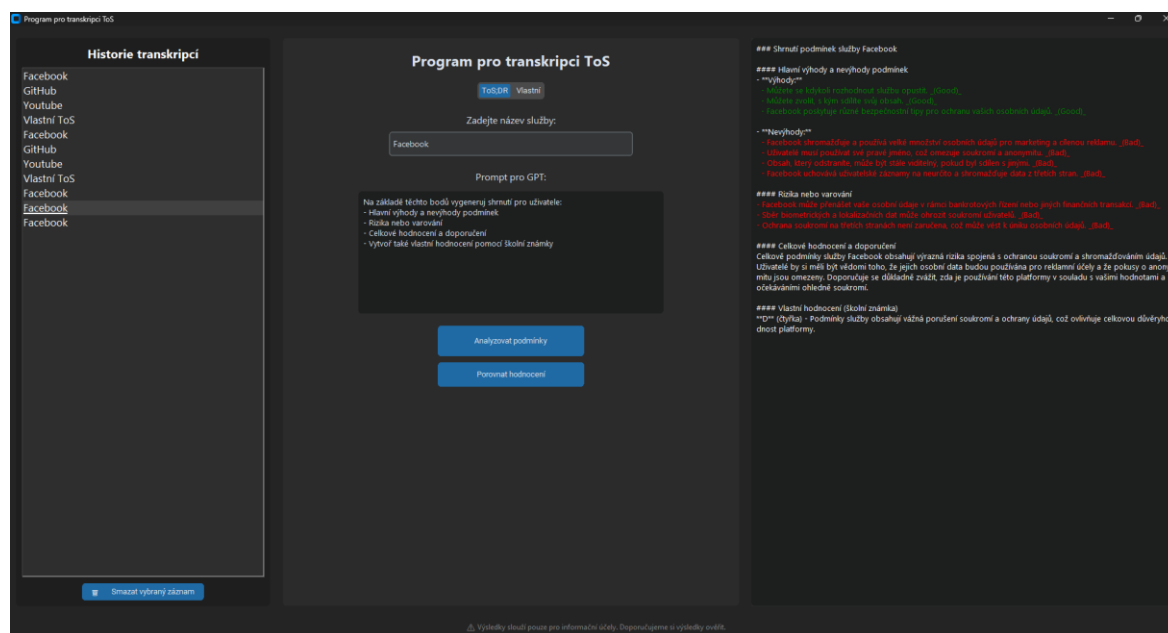
17.9.1 Výsledky analýzy

Ve všech třech případech byla rozpoznána stejná základní struktura výstupu – model správně identifikoval výhody (např. možnost zrušení účtu nebo opt-out z cílené reklamy), nevýhody (např. sledování uživatelů, obchodování s daty, omezení anonymita) a rizika (např. sdílení dat s třetími stranami, nedostatečná ochrana citlivých údajů). Důležité je, že ve všech časech byla podmínkám udělena známka „D“.

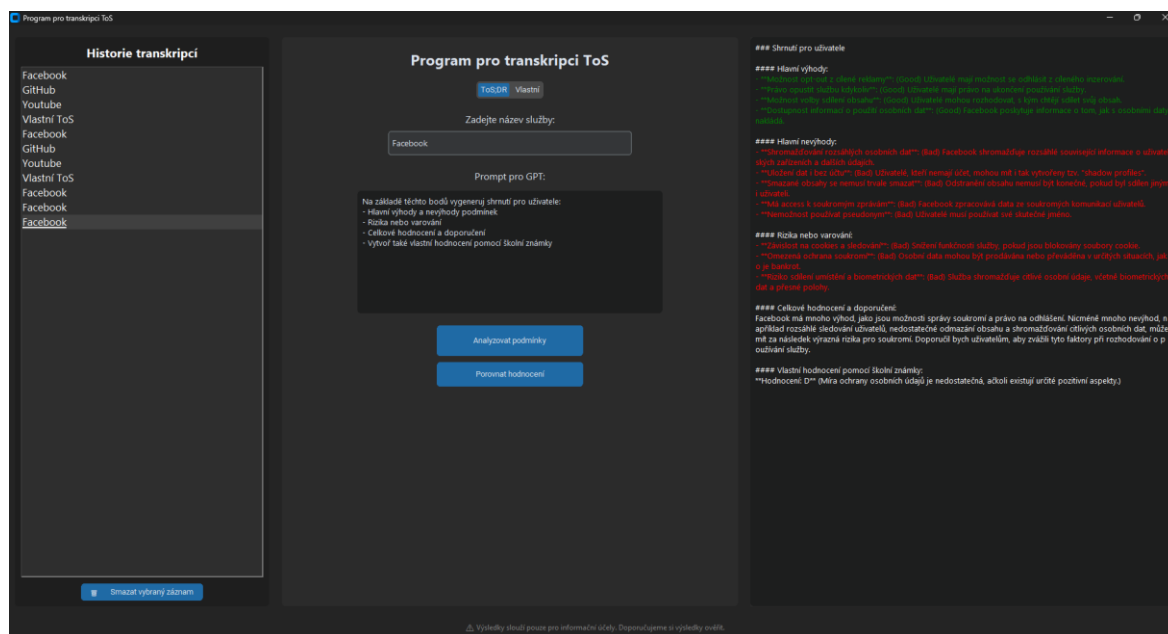
Přesto došlo k určitým jazykovým a stylistickým odchylkám:

- 9:00 (Obrázek 16): Výstup byl nejstručnější, spíše informativní než vysvětlující. Obsahoval méně detailů a komentářů, rychle shrnul klíčové body.
- 15:00 (Obrázek 17): Text byl obsáhlejší, s výraznějším důrazem na formulace a opakování. Obsahoval více podrobností a rozvedené formulace negativ.
- 19:00 (Obrázek 18): Výstup byl nejvyváženější, jazykově ucelený. Rizika byla lépe vysvětlena a celkové doporučení působilo profesionálněji.

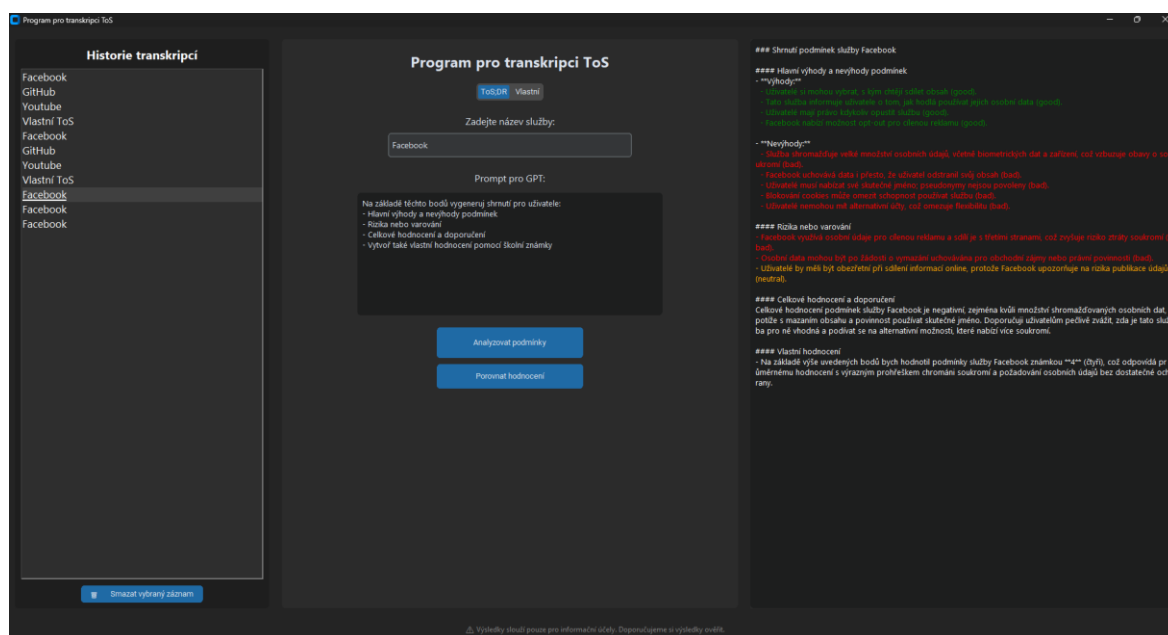
Výsledky ukazují, že jazykový model zachovává základní konzistenci v hodnocení i obsahu, ale jednotlivé výstupy se mohou jemně lišit ve stylu, rozsahu a výběru formulací. Tyto rozdíly však nejsou nejspíše způsobeny denní dobou, ale vycházejí z náhodnosti generativního procesu modelu. I při opakovaném zadání stejných vstupů generuje model mírně odlišné varianty, ale zanechává hlavní myšlenky.



Obrázek 17 Analýza služby Facebook – 9:00 (zdroj: vlastní)



Obrázek 18 Analýza služby Facebook – 15:00 (zdroj: vlastní)



Obrázek 19 Analýza služby Facebook – 19:00 (zdroj: vlastní)

17.10 Shrnutí výsledků

V rámci testování a vyhodnocování řešení byly analyzovány různé přístupy, typy vstupů a varianty promptů s cílem zjistit, jak dobře si velký jazykový model ChatGPT poradí s interpretací právních dokumentů, konkrétně s podmínkami užívání online služeb. Testování bylo provedeno na třech úrovních: s využitím veřejných dat z ToS;DR, s vlastními texty vloženými uživatelem a s různě formulovanými prompty ovlivňujícími charakter výstupu.

Při použití výchozího promptu model generoval přehledné shrnutí ve formě výhod, nevýhod, rizik a závěrečného hodnocení známkou. Tato forma se ukázala jako vhodná pro rychlou orientaci běžných uživatelů, kteří si chtějí udělat jasnou představu o kvalitě podmínek služby bez nutnosti číst celý dokument. Výstupy byly navíc barevně rozlišeny podle rizikovitosti jednotlivých bodů, což usnadňuje jejich pochopení.

Alternativní prompt, který modelu přidělil roli právního poradce, vedl k podstatně hlubší analýze. Namísto seznamu klasifikovaných bodů model nabídl komentované návrhy na úpravy konkrétních pasáží, upozorňoval na právní slabiny, doporučoval posílit transparentnost nebo zpřesnit jazykové formulace. Tento přístup prokázal využitelnost modelu i v expertnějším prostředí, např. při revizi podmínek před publikací.

Důležitou součástí testování byla i práce s vlastním vstupem zadaným uživatelem, ať už ručně, nebo nahráním PDF. I zde model správně identifikoval klíčová rizika (např. sledování aktivity, sdílení dat s třetími stranami, absence odpovědnosti) a poskytl relevantní zpětnou vazbu. Výsledné hodnocení v těchto případech odpovídalo závažnosti zjištěných nedostatků a navržená zlepšení měla formu konkrétních doporučení k přeformulování textu.

Při porovnání výstupů modelu s oficiálními hodnoceními z ToS;DR se potvrdilo, že GPT ve většině případů správně zachycuje hlavní body, například upozorňuje na sběr osobních údajů, sdílení s třetími stranami, odpovědnost za data nebo právní omezení. Model často použil vlastní formulace, ale významově se shodoval s verdikty z ToS;DR. V některých případech dokonce rozvedl problematiku hlouběji, např. doplněním důsledků pro uživatele nebo návrhy řešení.

Zajímavým doplňkem byla i časová analýza výstupů, která ukázala, že výsledky modelu mohou být mírně odlišné i při identickém vstupu, pokud je požadavek zadán v jiný čas. Výstupy se lišily především v rozsahu a formální úrovni vyjadřování. Obsahově však zůstávaly klíčové myšlenky, hodnocení i klasifikace bodů podobné. To potvrzuje, že hlavní faktory ovlivňující výsledky jsou náhodnost generativního procesu a jazykové variace, a ne denní doba.

Celkově lze říct, že ChatGPT je velmi schopným nástrojem pro analýzu smluvních podmínek. Nabízí kombinaci přístupnosti pro normální uživatele i možnosti hlubší právní interpretace. Díky možnosti práce s různými typy vstupů, od veřejně dostupných dat po vlastní dokumenty, a díky otevřenosti k různým promptům představuje flexibilní a praktické

řešení pro hodnocení podmínek užívání online služeb, a to i při opakovaném použití v různém čase.

18 SPUŠTĚNÍ APLIKACE

Aplikaci je možné spustit v libovolném vývojovém prostředí, které podporuje Python, doporučeno je například prostředí PyCharm. Nejprve je nutné mít na počítači nainstalovaný Python, ideálně ve verzi 3.10 nebo novější. Po instalaci stačí otevřít složku s aplikací (soubor `main.py`) jako nový projekt v PyCharm.

Po otevření projektu může být potřeba nastavit Python interpreter, pokud není nastaven automaticky, lze v prostředí jednoduše zvolit nainstalovanou verzi Pythonu nebo vytvořit nový virtuální interpreter. Po otevření souboru `main.py` se v kódu u jednotlivých importů zobrazí žárovka nebo červené podtržení. Kliknutím na žárovku a výběrem možnosti „Install package“ se příslušná knihovna automaticky nainstaluje.

Jakmile jsou všechny knihovny úspěšně nainstalovány, lze aplikaci spustit kliknutím pravým tlačítkem na soubor `main.py` a výběrem volby „Run“. Aplikace se následně otevře v grafickém okně a je připravena k použití. Pro správné fungování je třeba mít v kódu vložený platný OpenAI API klíč, který se nachází na začátku souboru v proměnné `api_key`.

ZÁVĚR

Tato práce se zaměřila na využití umělé inteligence, konkrétně velkých jazykových modelů, pro interpretaci a hodnocení smluvních podmínek online služeb. V rámci zadání bylo potřeba:

- vypracovat literární rešerši k dané problematice,
- navrhnout vhodný formát dat pro zpracování,
- zpracovat data z webové stránky tosdr.org,
- navrhnout vlastní řešení zpracování dat pomocí ChatGPT API,
- analyzovat možnosti využití pro další webové stránky a podmínky užití,
- a nakonec vhodně prezentovat výsledky práce.

Literární rešerše shrnula relevantní oblasti jako fungování velkých jazykových modelů, principy zpracování přirozeného jazyka, existující přístupy k analýze Terms of Use a dostupné zdroje dat (ToS;DR, PrivacySpy). Byl navržen flexibilní formát dat pro reprezentaci bodů hodnocení, který umožňuje jednotné zpracování jak oficiálních výstupů, tak generovaných odpovědí. Dále byla vytvořena aplikace, která kombinuje uživatelskou přívětivost s pokročilou analýzou – umožňuje načítání dat z API tosdr.org i vlastní analýzu pomocí ChatGPT (včetně podpory PDF dokumentů a vlastních promptů). Výsledky byly porovnávány s oficiálními body a výstupy byly přehledně vizualizovány barevným rozlišením rizikovosti. Součástí práce je i zhodnocení přesnosti modelu a jeho použitelnosti v praxi.

V rámci vývoje bylo nutné řešit technická omezení, jako je práce s omezenou délkou vstupu modelu, integrace různých zdrojů dat a návrh efektivního a přehledného uživatelského rozhraní. Výsledkem je funkční a rozšiřitelný nástroj, který lze využít jak pro osobní orientaci v podmínkách užití, tak jako podpora právní či akademické analýzy.

Do budoucna lze uvažovat o rozšíření aplikace o další datové zdroje, lokalizaci do více jazyků, pokročilejší hodnocení rizik či integraci s dalšími nástroji pro sledování právní transparentnosti online služeb.

SEZNAM POUŽITÉ LITERATURY

Seznam použité literatury musí být vždy vypracován podle použité citační normy v rámci závěrečné práce. Následující ukázky seznamu použité literatury vychází z citační normy ČSN ISO 690:2022, kde se řazení seznamu řídí zvolenou metodou odkazování.

Příklad č. 1 – metoda číselného odkazu, řazení dle výskytu v textu

- [1] STRYKER, Cole. *What is artificial intelligence (AI)?* Online. IBM. 2024, vol. 2024, s. 1-11. Dostupné z: <https://www.ibm.com/think/topics/artificial-intelligence>. [cit. 2025-03-13].
- [2] *Jednoduše*. Universum. Praha: Euromedia Group, 2023. ISBN 978-80-242-9293-9. [cit. 2025-03-13].
- [3] VONDRÁK, Ivo. *Umělá inteligence*. Olomouc: Univerzita Palackého, 1991. ISBN 80-706-7013-4. [cit. 2025-03-14].
- [4] *What is the history of artificial intelligence (AI)?* Online. Tableau. 2025. Dostupné z: <https://www.tableau.com/data-insights/ai/history>. [cit. 2025-03-13].
- [5] *The Turing Test: What Is It, What Can Pass It, and Limitations*. Online. Investopedia. 2024. Dostupné z: <https://www.investopedia.com/terms/t/turing-test.asp>. [cit. 2025-03-14].
- [6] MAŘÍK, Vladimír; ŠTĚPÁNKOVÁ, Olga a LAŽANSKÝ, Jiří. *Umělá inteligence*. Praha: Academia, 2013. ISBN 978-80-200-2276-9. [cit. 2025-03-14].
- [7] *What is NLP (natural language processing)?* Online. IBM. 2024. Dostupné z: <https://www.ibm.com/think/topics/natural-language-processing>. [cit. 2025-03-15].
- [8] *What is natural language processing (NLP)?* Online. CloudFlare. © 2025. Dostupné z: <https://www.cloudflare.com/learning/ai/natural-language-processing-nlp/>. [cit. 2025-03-15].
- [9] CHOWDHARY, K. R. *Fundamentals of artificial intelligence*. New Delhi: Springer, [2020]. ISBN 978-81-322-3970-3.
- [10] JURAFSKY, Daniel a MARTIN, James H. *Speech and Language Processing*. Online. 3rd ed draft. Stanford: Stanford University, 2025. Dostupné z: https://web.stanford.edu/~jurafsky/slp3/ed3book_Jan25.pdf. [cit. 2025-03-15].

- [11] RYBKA, Michal. Jak se vlastně umělá inteligence učí? Online. *Alza*. 2024, roč. 2024, s. 1-3. Dostupné z: <https://www.alza.cz/jak-se-vlastne-umela-inteligence-uci>. [cit. 2025-03-13].
- [12] PERRY, Christian. *How to Train AI: Step-by-Step Guide*. Online. Undetectable. 2024. Dostupné z: <https://undetectable.ai/blog/how-to-train-ai/>. [cit. 2025-03-15].
- [13] FLINDERS, Mesh a SMALLEY, Ian. What is AI inference? Online. *IBM*. 2024, vol. 2024, s. 1-6. Dostupné z: <https://www.ibm.com/think/topics/ai-inference>. [cit. 2025-03-15].
- [14] *What are large language models (LLMs)?* Online. *IBM*. 2023. Dostupné z: <https://www.ibm.com/think/topics/large-language-models>. [cit. 2025-03-16].
- [15] *How generative AI and LLMs work*. Online. *Microsoft*. 2024. Dostupné z: <https://learn.microsoft.com/en-us/dotnet/ai/conceptual/how-genai-and-llms-work>. [cit. 2025-03-16].
- [16] *What is ChatGPT?* Online. *TechTarget*. Copyright 1999 - 2025. Dostupné z: <https://www.techtarget.com/whatis/definition/ChatGPT>. [cit. 2025-03-16].
- [17] *ChatGPT: Everything you need to know about OpenAI's GPT-4 tool*. Online. *Science Focus*. © 2024. Dostupné z: <https://www.sciencefocus.com/future-technology/gpt-3>. [cit. 2025-03-16].
- [18] *15 Benefits of ChatGPT (+8 Disadvantages)*. Online. In: *SemRush*. © 2008–2025. Dostupné z: <https://www.semrush.com/blog/benefits-chatgpt/>. [cit. 2025-03-16].
- [19] *What are the Advantages of ChatGPT?* Online. *Parangat*. © 2025. Dostupné z: <https://www.parangat.com/what-are-the-advantages-of-chatgpt/>. [cit. 2025-03-16].
- [20] *ASR Transcription Software*. Online. In: *Verbit*. © 2025. Dostupné z: <https://verbit.ai/ai-technology/asr-and-the-next-generation-of-transcription/#head-3>. [cit. 2025-04-08].
- [21] *Terms of Use*. Online. *TermsFeed*. © 2012 - 2025. Dostupné z: <https://www.termsfeed.com/dictionary/terms-of-use-definition/>. [cit. 2025-03-17].
- [22] *What are the Terms of Use, and why do you need them?* Online. *Linkedin*. © 2025. Dostupné z: <https://www.linkedin.com/pulse/what-terms-use-why-do-you-need-them-avitar-group>. [cit. 2025-03-17].

- [23] *What is a Terms of Use Agreement and How To Make One*. Online. Ownr. © 2025. Dostupné z: <https://www.ownr.co/blog/terms-of-use-agreement/>. [cit. 2025-03-17].
- [24] *About*. Online. PrivacySpy. © 2025. Dostupné z: <https://privacyspy.org/about/>. [cit. 2025-04-08].
- [25] Facebook. Online. Privacyspy. 2021. Dostupné z: <https://privacyspy.org/product/facebook/>. [cit. 2025-05-08].
- [26] *About ToS;DR*. Online. ToS;DR. 2012. Dostupné z: <https://tosdr.org/en/about>. [cit. 2025-04-08].
- [27] *How Data Collection is Done for AI Model Training*. Online. Powerling. 2024, vol. 2024, s. 1. Dostupné z: <https://www.powerling.com/blog/how-data-collection-is-done-for-ai-model-training>. [cit. 2025-03-15].
- [28] *What is Web Scraping and How to Use It?* Online. GeeksforGeeks. 2025. Dostupné z: <https://www.geeksforgeeks.org/what-is-web-scraping-and-how-to-use-it/>. [cit. 2025-03-15].
- [29] *Web scraping: k čemu slouží a kdo ho provádí?* Online. Coderslab. © 2012-2025. Dostupné z: <https://coderslab.cz/cz/blog/co-je-webscraping>. [cit. 2025-03-15].
- [30] *BeautifulSoup*. Online. Crummy. 2025. Dostupné z: <https://www.crummy.com/software/BeautifulSoup/>. [cit. 2025-03-15].
- [31] *Beautiful Soup: Build a Web Scraper With Python*. Online. Real Python. 2024. Dostupné z: <https://realpython.com/beautiful-soup-web-scraper-python/>. [cit. 2025-03-15].
- [32] *Web Scraping vs. API: Which Is Best for Your Project?* Online. Zenrows. 2025, 3. března 2025. Dostupné z: <https://www.zenrows.com/blog/web-scraping-vs-api#what-is-api-scraping>. [cit. 2025-03-15].
- [33] PECINOVSKÝ, Rudolf. *Python: kompletní příručka jazyka pro verzi 3.11*. Knihovna programátora. Praha: Grada Publishing, 2023. ISBN 978-80-271-3891-3. Dostupné také z: <https://www.grada.cz/>. [cit. 2025-03-17].
- [34] *What Is Python Used For? A Beginner's Guide*. Online. Coursera Inc. © 2025. Dostupné z: <https://www.coursera.org/articles/what-is-python-used-for-a-beginners-guide-to-using-python>. [cit. 2025-03-17].

SEZNAM OBRÁZKŮ

Obrázek 1 Stavba neuronové sítě [2, s. 77].....	18
Obrázek 2 Ukázka hodnocení na stránce Privacyspy [33] (vlastní zpracování – screenshot)	35
Obrázek 3 Schéma zpracování dotazu pomocí API ToS;DR (zdroj: vlastní).....	48
Obrázek 4 Ukázka vzhledu aplikace (zdroj: vlastní)	56
Obrázek 5 Výstup analýzy služby Github s využitím výchozího promptu (zdroj: vlastní).....	61
Obrázek 6 Analýza výstupu služby Facebook pomocí výchozího promptu (zdroj: vlastní).....	62
Obrázek 7 Výstup analýzy služby Youtube pomocí výchozího promptu (zdroj: vlastní).....	63
Obrázek 8 GitHub analýza – alternativní prompt (zdroj: vlastní)	65
Obrázek 9 Facebook analýza – alternativní prompt (zdroj: vlastní).....	66
Obrázek 10 Youtube analýza – alternativní prompt (zdroj: vlastní).....	67
Obrázek 11 Ukázka vlastních Terms of Use (zdroj: vlastní)	68
Obrázek 12 Analýza vlastních Terms of Use (zdroj: vlastní).....	69
Obrázek 13 Vlastní Terms of Use – kreativní prompt (zdroj: vlastní)	70
Obrázek 14 Porovnání bodů ToS;DR s výstupem modelu u služby Github (zdroj: vlastní)	71
Obrázek 15 Porovnání bodů ToS;DR s výstupem modelu u služby Facebook (zdroj: vlastní)	72
Obrázek 16 Porovnání bodů ToS;DR s výstupem modelu u služby Youtube (zdroj: vlastní)	73
Obrázek 17 Analýza služby Facebook – 9:00 (zdroj: vlastní).....	74
Obrázek 18 Analýza služby Facebook – 15:00 (zdroj: vlastní).....	75
Obrázek 19 Analýza služby Facebook – 19:00 (zdroj: vlastní).....	75

SEZNAM TABULEK

Tabulka 1 API vs Webscraping	40
Tabulka 2 Výhody navrhovaného řešení	52

SEZNAM POUŽITÝCH SYMBOLŮ A ZKRATEK

AI	Artificial Intelligence
API	Application Programming Interface
ASR	Automatic Speech Recognition
CSV	Comma Separated Values
EU	Evropská unie
GDPR	General Data Protection Regulation
GPT	Generative Pre-trained Transformer
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
ID	Identifier
JSON	JavaScript Object Notation
KNN	K-nearest neighbors
LLM	Large Language Model
NLP	Natural-Language Processing
PDF	Portable Document Format
REST API	Representational State Transfer API
SVG	Scalable Vector Graphics
SVM	Support vector machines
ToS;DR	Terms of Service; Didn't Read
URL	Uniform Resource Locator
XML	Extensible Markup Language

SEZNAM PŘÍLOH

Příloha P I: main.py

Příloha P II: history.json

PŘÍLOHA P I: MAIN.PY

Struktura přílohy:

Příloha obsahuje Python skripty (verze 3.x) potřebné ke spuštění desktopové aplikace pro analýzu smluvních podmínek pomocí OpenAI GPT a ToS;DR API. Zahrnuje hlavní soubor main.py, moduly pro práci s PDF, správu výstupů a uživatelské rozhraní v CustomTkinter.

PŘÍLOHA P II: HISTORY.JSON

Struktura přílohy:

Příloha obsahuje historii provedených transkripcí a analýz smluvních podmínek prováděných pomocí navržené aplikace. Tyto záznamy zahrnují vstupní texty, výsledky modelu, klasifikace a metadata o jednotlivých požadavcích. Údaje sloužily jako podklad pro vyhodnocení výstupů a byly využity při tvorbě ukázkových screenshotů v praktické části bakalářské práce.