Kvalita a přesnost detekce textů vytvořených umělou inteligencí

Daniel Rafaj

Bakalářská práce
2024

Univerzita Tomáše Bati ve Zlíně
Fakulta aplikované informatiky

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

Akademický rok: 2023/2024

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

(projektu, uměleckého díla, uměleckého výkonu)

Jméno a příjmení:	**Daniel Rafaj**
Osobní číslo:	**A20390**
Studijní program:	**B0613A140020 Softwarové inženýrství**
Forma studia:	**Prezenční**
Téma práce:	**Kvalita a přesnost detekce textů vytvořených umělou inteligencí**
Téma práce anglicky:	**Quality and Accuracy of Artificial-Intelligence-Based Text Detectors**

Zásady pro vypracování

1. Sestavte přehled a charakteristiky volně dostupných nástrojů využívajících umělou inteligenci využitelných pro vytváření podvržených akademických textů.
2. Uveďte přehled a charakteristiky dostupných nástrojů pro odhalování podvržených akademických textů.
3. S využitím doporučené literatury podrobně popište zjištěné nedostatky v kvalitě a přesnosti detekce akademických textů vytvořených umělou inteligencí.
4. Vyberte a reprodukujte některý ze zjištěných závažných nedostatků.
5. Navrhněte možná dílčí řešení předmětného problému.

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

Seznam doporučené literatury:

1. KASÍK, Pavel. *Jak odhalit text psaný počítačem: můžete to zkusit, ale je to téměř nemožné*. Online. In: Seznam Zprávy. May 20 2023. Dostupné z: https://www.seznamzpravy.cz/clanek/tech-ai-umela-inteligence-jak-odhalit-text-psany-pocitacem-muzete-to-zkusit-ale-je-to-temer-nemozne-230894. [cit. 2023-10-30].
2. SADASIVAN, Vinu Sankar; KUMAR, Aounon; BALASUBRAMANIAN, Sriram; WANG, Wenxiao a FEIZI, Soheil. *Can AI-Generated Text be Reliably Detected?* Online. 28 Jun 2023. Dostupné z: https://arxiv.org/abs/2303.11156. [cit. 2023--10-27].
3. WEBER-WULFF, Debora; ANOHINA-NAUMECA, Alla; BJELOBABA, Sonja; FOLTÝNEK, Tomáš; GUERRERO-DIB, Jean et al. *Testing of Detection Tools for AI-Generated Text*. Online. 10 Jul 2023. Dostupné z: https://arxiv.org/abs/2306.15666. [cit. 2023-10-27].
4. WILLIAMS, Tom. *GPT-4's launch 'another step change' for AI and higher education*. Online. In: THE – Times Higher Education. March 23, 2023. Dostupné z: https://www.timeshighereducation.com/news/gpt-4s-launch-another-step--change-ai-and-higher-education. [cit. 2023-10-27].
5. WILLIAMS, Tom. *AI text detectors 'biased against non-native English speakers'*. Online. In: THE – Times Higher Education. July 11, 2023. Dostupné z: https://www.timeshighereducation.com/news/ai-text-detectors-biased-against--non-native-english-speakers. [cit. 2023-10-27].
6. WILLIAMS, Tom. *AI text detectors aren't working. Is regulation the answer?* Online. In: THE – Times Higher Education. August 9, 2023. Dostupné z: https://www.timeshighereducation.com/news/ai-text-detectors-arent-working-regulation-answer. [cit. 2023-10-27].

Vedoucí bakalářské práce: **doc. Ing. Libor Pekař, Ph.D.**
Ústav automatizace a řídicí techniky

Datum zadání bakalářské práce: **5. listopadu 2023**
Termín odevzdání bakalářské práce: **13. května 2024**

UNIVERZITA TOMÁŠE BATI VE ZLÍNĚ
Fakulta aplikované informatiky
DĚKAN
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 5. ledna 2024

Prohlašuji, že

- beru na vědomí, že odevzdáním bakalářské práce souhlasím se zveřejněním své práce podle zákona č. 111/1998 Sb. o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších právních předpisů, bez ohledu na výsledek obhajoby;
- beru na vědomí, že bakalářská práce bude uložena v elektronické podobě v univerzitním informačním systému dostupná k prezenčnímu nahlédnutí, že jeden výtisk bakalářské práce bude uložen v příruční knihovně Fakulty aplikované informatiky Univerzity Tomáše Bati ve Zlíně;
- byl/a jsem seznámen/a s tím, že 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;
- beru na vědomí, že podle § 60 odst. 1 autorského zákona má UTB ve Zlíně právo na uzavření licenční smlouvy o užití školního díla v rozsahu § 12 odst. 4 autorského zákona;
- beru na vědomí, že podle § 60 odst. 2 a 3 autorského zákona mohu užít své dílo – bakalářskou práci nebo poskytnout licenci k jejímu využití jen připouští-li tak licenční smlouva uzavřená mezi mnou a Univerzitou Tomáše Bati ve Zlíně s tím, že vyrovnání případného přiměřeného příspěvku na úhradu nákladů, které byly Univerzitou Tomáše Bati ve Zlíně na vytvoření díla vynaloženy (až do jejich skutečné výše) bude rovněž předmětem této licenční smlouvy;
- beru na vědomí, že pokud bylo k vypracování bakalářské práce využito softwaru poskytnutého Univerzitou Tomáše Bati ve Zlíně nebo jinými subjekty pouze ke studijním a výzkumným účelům (tedy pouze k nekomerčnímu využití), nelze výsledky bakalářské práce využít ke komerčním účelům;
- beru na vědomí, že 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 samostatně a použitou literaturu jsem citoval. V případě publikace výsledků budu uveden jako spoluautor.
- že odevzdaná verze bakalářské práce a verze elektronická nahraná do IS/STAG jsou totožné.

Ve Zlíně, dne 9.5.2024

Daniel Rafaj, v.r.
podpis studenta

ABSTRAKT

Tato bakalářská práce se zaměřuje na veřejně dostupné nástroje využívající umělé inteligence, které lze využít k vytvoření podvržených akademických textů, ale také i na nástroje pro detekci těchto textů. V práci jsou popsány techniky vytváření a odhalování textů pomocí jazykových modelů a zároveň se testují jednotlivé detektory, a také útoky na ně. Cílem této práce je zjistit, zda je možné v dnešní době možné spolehlivě detekovat text vytvořený umělou inteligencí a navrhnout řešení, které by zvýšilo přesnost detekce.

Klíčová slova: umělá inteligence, jazykový model

ABSTRACT

This bachelor's thesis focuses on publicly available tools utilizing artificial intelligence that can be used to generate forged academic texts, as well as tools for detecting such texts. The thesis describes techniques for creating and uncovering texts using language models, while also testing individual detectors and attacks against them. The aim of this work is to determine whether it is currently possible to reliably detect text generated by artificial intelligence and to propose solutions to enhance detection accuracy.

Keywords: artificial intelligence, language model

Prohlašuji, že odevzdaná verze bakalářské práce a verze elektronická nahraná do IS/STAG jsou totožné.

Prohlašuji, že při tvorbě této práce jsem použil nástroj generativního modelu AI [ChatGPT; https://chat.openai.com/] za účelem vytváření textu, který byl následně použit pro testování detektorů textů umělé inteligence, přehlednější formulace textu a překládání textu v případech, kdy jsem nevěděl, jak dané slovo přeložit a zároveň Google překladač nebyl schopen dané slovo rozumně přeložit. Po použití tohoto nástroje jsem proved kontrolu obsahu a přebírám za něj plnou zodpovědnost.

OBSAH

ÚVOD

Na úvod byl chtěl poznamenat, že se podvrženým textem v této práci rozumí textu, jež byl vytvořen umělou inteligencí a byl následně použit při akademické činnosti za cílem získat nečestnou výhodu.

V dnešní době je umělá inteligence jedním z nejvíce diskutovaných témat a vyskytuje se v mnoha oblastech našeho každodenního života, aniž bychom si toho byli vědomi. Ve světě internetu jsou nám při prohlížení webu zobrazovány reklamy, které nejvíce vyhovují naším zájmům a potřebám, nebo při zadání jakéhokoliv dotazu ve vyhledávači jsou nám nabízeny nejvíce relevantní výsledky hledání. Jelikož je trénovatelná pro naše vlastní požadavky, tak se stává použitelnou v obrovské části průmyslu, jako je například zdravotnictví, hospodářství, doprava, výroba a mnoho dalších.

Studenti začali využívat nástrojů jako je ChatGPT pro ulehčení akademických povinností, čímž bylo zřetelné, že bude potřeba regulovat použití takových nástrojů. Proto se začali vyvíjet detektory, které měli na starost odhalovat AI text, ale rychle se zjistilo, že spolehlivá detekce není tak jednoduše realizovatelná.

Začátek teoretické části popisuje stručnou historii AI a následně je popsáno její rozdělení. Dále je rozebrán jazykový model a princip generování textu. Následně jsou sepsány charakteristiky nástrojů pro generování a detekci AI textu a jakých technik využívají. V neposlední řadě jsou uvedeny zjištěné nedostatky při detekci AI textu a také možné útoky na detektory.

Praktická část práce obsahuje testování 5 různých detektorů na různých variacích textu. Výsledky testování jsou poté porovnány s výsledky jiné práce a je zjištěno, zda se výsledky shodují a ke konci kapitoly je uvedeno, jak vylepšit přesnost detekce AI textu.

Cílem této práce je zjistit, zda je v dnešní době možné spolehlivě detekovat text vygenerovaný umělou inteligencí, vytvořit přehled přesností detektorů a zároveň seznámit čtenáře s umělou inteligencí a jakých technik využívá pro generování a detekci textu.

I. TEORETICKÁ ČÁST

1 UMĚLÁ INTELIGENCE

Umělá inteligence (AI) je rapidně vyvíjející se technologie, která se snaží simulovat lidskou inteligenci pomocí strojů. Umělá inteligence zahrnuje různé podoblasti, včetně strojového učení (ML) a hlubokého učení (DL), které systémům umožňují učit se a přizpůsobovat se novým způsobem pomocí tréninkových dat. Lze ji aplikovat v mnoha odvětvích, jako je např. zdravotnictví, finance a doprava, ale je možné ji také aplikovat téměř ve všech způsobech, jak počítače v dnešní době používáme. [1]

1.1 Historie

První zmínka termínu „umělá inteligence“ se uskutečnila v roce 1956 John McCarthym při průběhu akademické konference, která se zabývala stejným tématem, ale samotná otázka, zda stroje skutečně dokážou myslet, začala mnohem dříve. [2]

Už od roku 380 př. n. l. začali mnozí matematici, filozofové a profesoři uvažovat o číselných systémech, počítacích strojích a mechanických technikách, které pomohly sestrojit budoucí koncept mechanizovaného lidského myšlení nelidských bytostí. [3]

V roce 1921 mělo vědeckofantastické drama Karla Čapka, R.U.R., světovou premiéru, ve které byl poprvé zmíněn termín „robot“, kterým nazval uměle vyrobené lidi v továrně. [3]

V roce 1949 vědec Edmund Berkeley vydal knihu „Giant Brains: Or Machines That Think“, která poznamenala, že stroje jsou čím dál tím více vyvinuté. Dále je v knize uvedené srovnávání strojů s lidským mozkem, kde přirovnal hardware a dráty k lidskému masu a nervům, přičemž poznamenal, že schopnost strojů je srovnatelná se schopností lidské mysli a uvedl, že „stroj proto může myslet“. [3]

Druhá polovina 20. století se osvědčila jako období, kdy se mnoho pokroků v oblasti umělé inteligence povedlo uskutečnit s nárůstem výzkumných poznatků v oblasti AI od různých počítačových vědců.

Roku 1950 publikoval Alan Turing spis „Computing Machinery and Intelligence“, ve kterém nejdříve klade jednoduchou otázku, a to zda dokážou stroje myslet. Turing poté navrhl metodu pro hodnocení, zda stroje mohou myslet, která se stala známou jako Turingův test. Tento test, byl předložen jako jednoduchý test, který lze použít k prokázání, zda stroje dokážou myslet nebo ne. Turingův test používá jednoduchý pragmatický přístup, který říká, že pokud nelze rozpoznat počítač od inteligentního člověka, tak stroj dokáže myslet. [2]

Jestli stroje opravdu dokážou myslet, bylo stále velkou otázkou, a proto v roce 1980 Americký filozof John Searle napsal článek pojmenovaný „The Chinese Room Argument“, který se také stal jedním z nejznámějších argumentů současné filozofie. Searle si představuje, že je sám v místnosti a sleduje počítačový program pro reakci na čínské znaky, které mu jsou pod dveřmi předány. Searle nezná jediný čínský znak a přesto tím, že dodržuje program pro manipulaci se symboly a číslicemi stejně jako počítač, posílá vhodné řetězce čínských znaků zpět pod dveřmi, což vede lidi, co jsou venku za dveřmi k mylnému předpokladu, že v místnosti je čínsky hovořící osoba. Z daného argumentu vyplývá, že počítač dokáže produkovat validní odpovědi při použití celé knihovny pravidel a tabulek prohledávání, čímž je mu umožněno, aby vypadal, že jazyku opravdu rozumí, i když ve skutečnosti není schopen produkovat žádné reálné pochopení. [2]

Do konce 20. století byly prováděny podstatné pokroky, ale největší růst této technologie je značně vidět ve 21. století, kde dané téma začalo být jedním z nejvíce hovořených, ale kvůli rapidnímu pokroku se také stává jedním z více kontroverzních a začíná být na něho také kladena otázka etiky.

1.2 Typy umělé inteligence

Umělou inteligenci lze klasifikovat do několika typů na základě schopností, funkcionalit a technologií.

1.2.1 Typy na základě schopností

1.2.1.1 Úzká umělá inteligence

Úzká umělá inteligence (ANI) vykonává úzce vymezené úlohy a dále se neučí. Působí v omezeném předdefinovaném rozsahu nebo sadě kontextů, což znamená, že je nepoužitelná v jiných úlohách, než na které byla vytvořena. Často se využívá k automatizace opakovaných pracovních úloh. Příkladem využití je úzké umělé inteligence jsou hlasoví asistenti jako Alexa nebo Siri. [4]

1.2.1.2 Obecná umělá inteligence

Obecná umělá inteligence (AGI) dosud není realizována. V ideální formě by se podobala chování člověku a dokázala by velmi efektivně řešit jakýkoliv úkol a zároveň se samostatně zdokonalovat a rozvíjet své schopnosti. [4]

1.2.1.3 Superinteligentní umělá inteligence

Superinteligentní umělá inteligence (ASI) je stejně jako u předchozího typu dosud nerealizována a je pouze v hypotetickém stavu. Tento typ by dokázal překračovat lidskou inteligenci a dosahovat nejlepších výkonů ve všech oblastech. Je zřejmé, že by tento typ mohl přinést neskutečné množství výhod pro celý svět, ale zároveň by mohlo dojít k mnohým rizikům. [4]

1.2.2 Typy na základě funkcionalit

1.2.2.1 Reaktivní stroje

Reaktivní stroje jsou systémy umělé inteligence, které nemají žádnou paměť a jsou specifické pro úlohu, což znamená, že vstup vždy poskytuje stejný výstup. Reaktivní stroje mohou reagovat pouze na omezenou kombinaci vstupů a jsou nejzákladnějším typem umělé inteligence. Jelikož nemají paměť, tak nemůžou stavět na předchozích znalostech ani provádět složitější úkoly. [5]

V praxi jsou reaktivní stroje užitečné pro provádění základních autonomních funkcí, jako je například filtrování spamu z e-mailové schránky, chatboti, nebo i AI do her. [5]

1.2.2.2 Limitovaná paměť

Umělá inteligence s omezenou pamětí může ukládat minulá data a používat tato data k předpovědím. To znamená, že aktivně buduje svou vlastní krátkodobou znalostní bázi a plní úkoly založené na těchto znalostech. [6]

Jádrem omezené paměti umělé inteligence je hluboké učení, které napodobuje funkcionalitu neuronů v lidském mozku. To umožňuje stroji sbírat data ze zkušeností a učit se z nich, což mu pomáhá zlepšovat přesnost jeho akcí v průběhu času. Příklad využití AI s limitovanou pamětí jsou samořídící automobily. [6]

1.2.2.3 Teorie mysli

Teorie mysli je prvním ze dvou typů umělé intelligence, které ještě nebyly realizovány. Termínem se chápe schopnost lidí číst emoce druhých a předvídat budoucí akce na základě těchto informací, takže pokud někdy bude vytvořena, tak bude schopna právě těchto věcí. Teorie mysli ještě nebyla plně realizována a představuje další významný budoucí milník ve vývoji umělé inteligence. [6]

Mohla přinést spoustu pozitivních změn do technologického světa, ale také představuje svá vlastní rizika.

1.2.2.4 Umělá inteligence se sebevědomím

Sebevědomá umělá inteligence popisuje umělou inteligenci, která by vnímala a chápala svou vlastní přítomnost a dokonce měla své vlastní myšlenky a emoce. Tato umělá inteligence je fází za teorií mysli a je jedním z konečných cílů ve vývoji umělé inteligence. [7]

1.2.3 Typy na základě technologií

1.2.3.1 Strojové učení

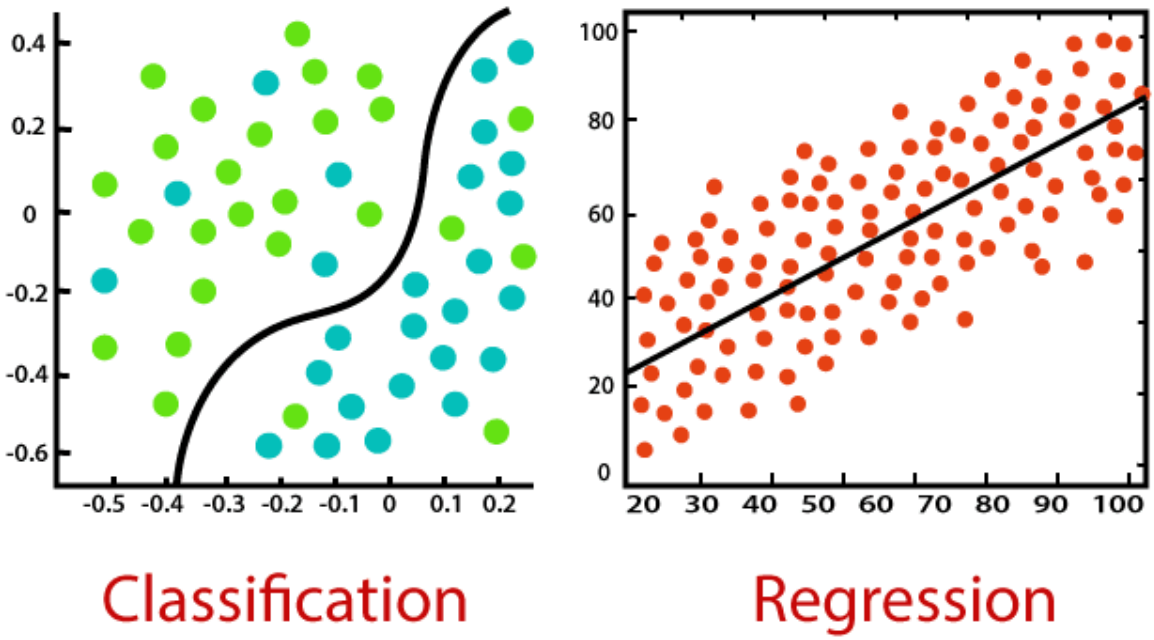
Strojové učení je odvětví umělé inteligence, ve kterém je hlavním zaměřením vytvořit takový systém, který se dokáže sám zlepšovat pomocí učení se z dat a zkušeností, a pomocí nich tvořit predikce. Na rozdíl od tradičního programování, kde jsou předdefinovány instrukce programu, jsou systému strojového učení předány ukázky a úloha pro vypracování. Jelikož nemá předem-definovaný postup, podle kterého by došel k výsledku úlohy, tak algoritmus zjišťuje vzorky na datech, podle kterých se učí a následně zjišťuje řešení úlohy. [8]

Proces učení jde rozdělit na 3 hlavní části, kde první částí je rozhodovací proces, který vyprodukuje odhad vzoru na základě vstupních dat. Druhou částí je chybná funkce, která má na starost vynést předpověď modelu. Třetí a poslední částí je optimalizační proces, ve kterém algoritmus iteračně opakuje úpravy vah tak, aby se snížila odchylka mezi známým příkladem a odhadem modelu, dokud není dosáhnuto požadované úrovně přesnosti. Strojové učení lze také rozdělit na 3 hlavní typy. [9]

1.2.3.1.1 Učení pod dohledem

Učení pod dohledem využívá označených dat, kde každý vstup má odpovídající požadovaný výstup. Zaměřuje se na zjišťování vztahů a vzorků mezi vstupem a výstupem. [10]

Existují 2 typy algoritmů učení pod dohledem, klasifikace a regrese. Při klasifikaci dochází k hledání funkci, která pomáhá rozřadit vstupní data do předem-definovaných tříd na základě jejich vlastností. Využívá se k předpovídání diskrétních výsledků na rozdíl od regrese, u které se předpovídají spojité výsledky. [10]



Obrázek 1. Klasifikace a regrese [10]

1.2.3.1.2 Učení bez dohledu

Učení bez dohledu zahrnuje trénování modelu na souboru dat, který není tříděný. Model je ponechán, aby sám našel vzorky a vztahy v datech bez dalšího zásahu člověkem. Tento proces lze rozdělit do 3 hlávních kroků. Prvním krokem je shlukování, při kterém dochází k řazení prvků podobajících se vlastností do stejné skupiny. Druhým krokem je sdružování, které je použito pro nalezení vztahů mezi vstupními proměnnými. [11]

1.2.3.1.3 Učení s posilou

Učení s posilou je metoda, pomocí které se zjišťuje optimální chování agenta v neznámém prostředí, aby byla získána maximální odměna. To znamená, že se jedná o metodu pokusu a omylu. Využívá se algoritmu, který je schopen učit se z jednotlivých výsledků a je schopen rozhodnout, jakou následující činnost provést. [12]

1.2.3.2 Hluboké učení

Hluboké učení je odvětví strojového učení, které se skládá z neuronové sítě se třemi nebo více skrytými vrstvami. Skládá se ze vstupní vrstvy, přes kterou data vstupují. Dále jsou využity skryté vrstvy, které mají za účet zpracovávat a přenášet data do dalších vrstev. Jako poslední je vrstva výstupní, ze které data ke konci vystupují. [13]

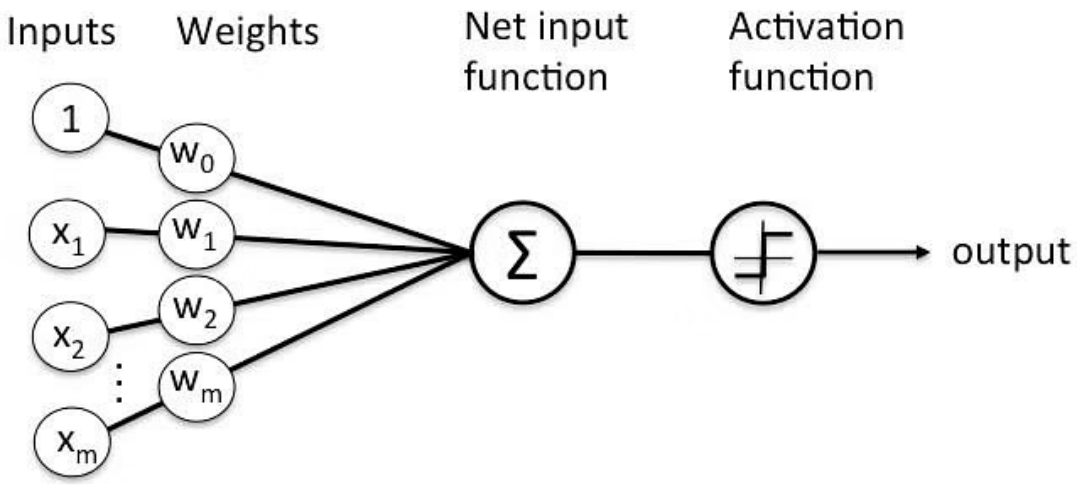
1.2.3.2.1 Neuronová síť

Neuronová síť (NN) je model strojového učení navržený tak, aby napodoboval funkci a strukturu lidského mozku. Neuronové sítě jsou tvořeny kolekcí procesních jednotek nazývaných uzly, které si navzájem předávají data stejným způsobem, jakým si v lidském mozku neurony předávají elektrické impulsy. Pokud neuronová síť obsahuje 2 nebo více skryté vrstvy, tak se hovoří o tzv. „hluboké neuronové síti“, jinak se mluví o tzv. „mělké neuronové síti“. [14]

Každý uzel provádí určitý druh zpracování na vstupu, který obdrží od předchozího uzlu, nebo z vstupní vrstvy, jedná-li se o první vrstvu. V podstatě každý uzel obsahuje matematický vzorec, kde každá proměnná ve vzorci má jinou váhu. Pokud výstup tohoto matematického vzorce překročí určitou prahovou hodnotu, uzel předá data další vrstvě v neuronové síti, jinak žádná data předána nejsou. [14]

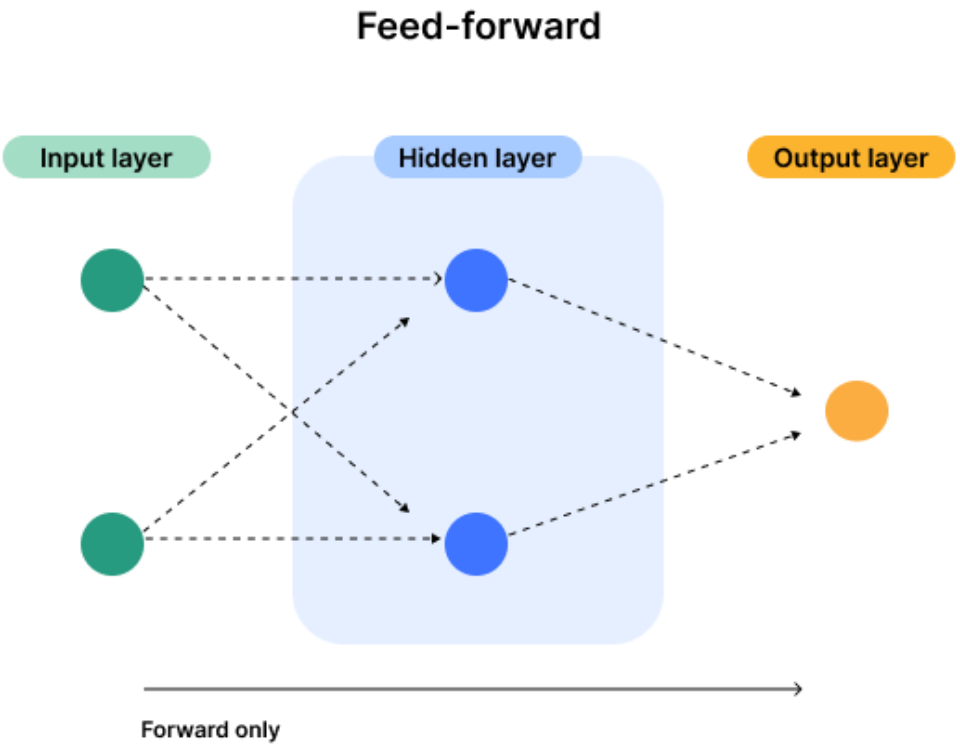
Existuje mnoho typů neuronových sítí, které jsou již k dispozici, ale zároveň existuje nemalé množství neuronových sítí, kterou jsou zatím pouze ve fázi vývoje. Jelikož existuje mnoho typů neuronových sítí, tak se zaměřím na ty nejznámější a nejvíce využívané:

- Perceptron. Tento model představuje jeden z nejzákladnějších a nejstarších modelů neuronů a funguje jako základní stavební blok neuronové sítě, který se využívá k identifikaci charakteristik ve vstupních datech. Nejdříve jsou data přivedena na vstupní vrstvu, na které jsou poté aplikovány váhy, které reprezentují sílu propojení mezi vstupním a výstupním neuronem. Dále se na vstupní vrstvě využívá sklonu nebo tendenci, díky které dokáže perceptron lépe pracovat s komplexnějšími vzorky. Jakmile je proveden součet součinů vstupních hodnot a jejich vah, tak jsou data vyslána do aktivační funkce, která nám dává konečný výsledek. Pokud je k součtu vah přičten sklon a výsledek je větší než 0, tak perceptron vynese na výstup 1, jinak vynese na výstup 0. Perceptron se obvykle používá pro lineárně oddělitelná data, kde se učí klasifikovat vstupy do dvou kategorií na základě rozhodovací hranice. Jelikož je perceptron učícím se algoritmem, který řadí data do dvou kategorií, tak je možné ho nazvat binárním klasifikátorem. [15]



Obrázek 2. Perceptron [15]

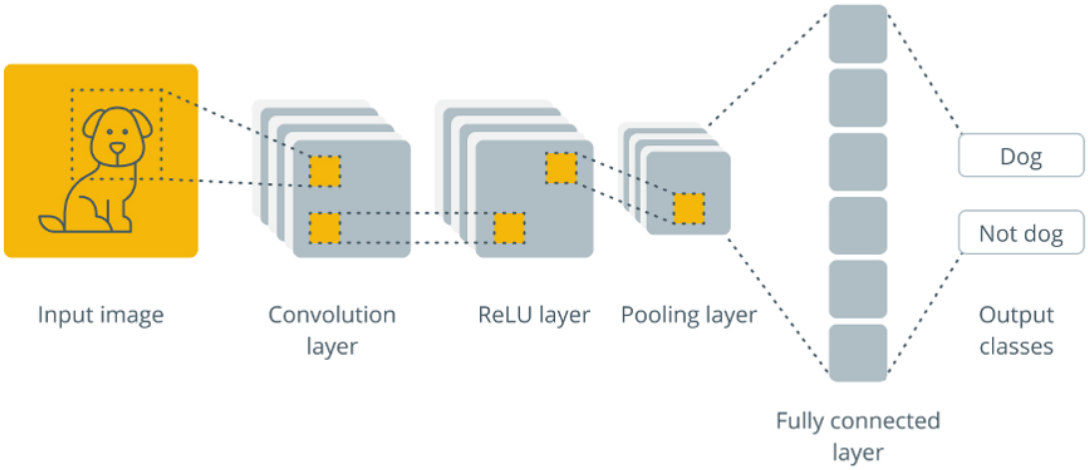
- Dopředné neuronové sítě (FNN). Jedná se o jednoduchou formu neuronových sítí, kde vstupní data putují pouze jedním směrem. Neuronová síť má vstupní, skryté a výstupní vrstvy. Každá vrstva je složena z neuronů, které jsou propojeny váhami a pracuje ve 2 fázích. První fází je fáze dopředná, ve které jsou vstupní data přiváděna do sítě a následně se šíří sítí dále. V každé skryté vrstvě se vypočítá vážený součet vstupů s tím, že tento proces pokračuje, dokud není dosaženo výstupní vrstvy a je provedena předpověď. Druhou fází je zpětná propagace, která zjišťuje rozdíl mezi předvídaným výstupem a skutečným výstupem. Jakmile je chyba vypočítána, tak je vyslána zpět na vstupní vrstvu, s tím, že jsou upraveny váhy tak, aby se zjištěná chyba minimalizovala. [16]



Obrázek 3. Dopředná neuronová síť [14]

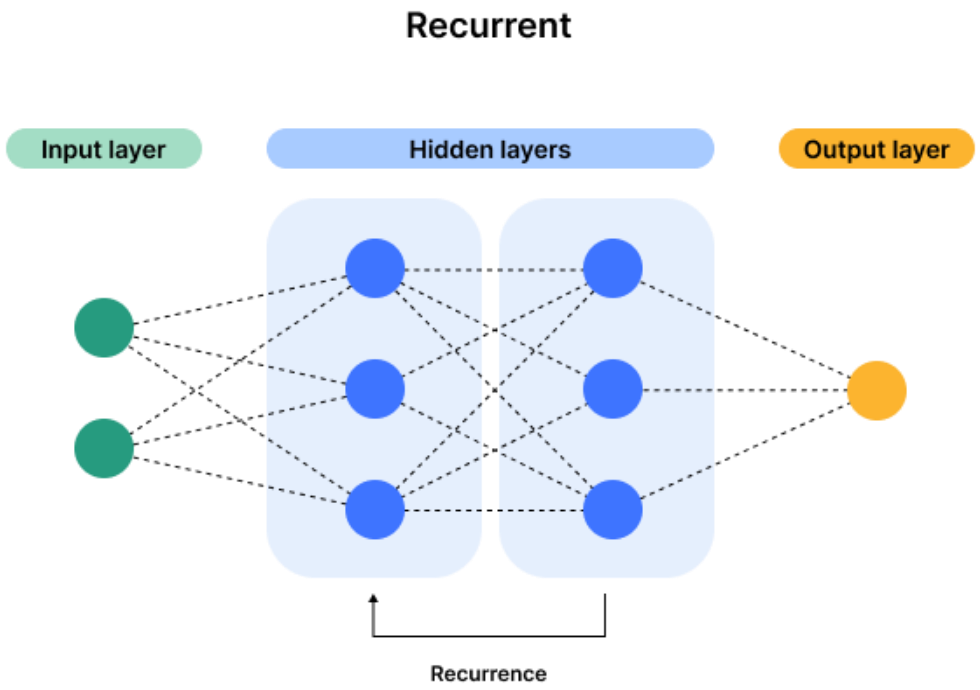
- Konvoluční neuronové sítě (CNN). Tento typ neuronové sítě se odlišuje od ostatních neuronových sítí tím, že dokážou velmi efektivně zpracovávat obraz nebo zvuk. Síť obvykle zpracovává obrázky jako data na vstupní vrstvě a obsahuje šířku, výšku a hloubku obrázku. Dále se data přesouvají do konvoluční vrstvy, na které jsou následně aplikovány filtry. Tyto filtry jsou malé matice, nejčastěji o rozměrech 3x3. Filtr dále jezdí po vstupních datech a vypočítává bodový součin mezi váhou filtru a odpovídajícím vstupním polem. Výstup této vrstvy se označuje jako mapa prvků. Další vrstvou je vrstva aktivační, ve které se provádí aktivační funkce, která se přidá k výstupu předchozí vrstvy a tím se dosahuje nelineárnosti v síti. Následně jsou přenesena do sdružovací vrstvy, kde je hlavním úkolem zmenšit velikost objemu dat, což v následku zrychluje výpočet a snižuje nároky na paměť. Dva běžné druhy sdružovacích vrstev jsou maximální a průměrné. Jednou z posledních vrstev je vrstva zplošťovací, která má za úkol sloučit výsledné mapy prvků do jednorozměrného vektoru, který umožňuje mapy předat do zcela propojené vrstvy, ve které probíhá konečná kategorizace nebo regrese. Poslední vrstvou je vrstva výstupní, ve které se

přivádí výstup ze zcela propojené vrstvy do logistické funkce pro klasifikační úlohy. [17; 18]



Obrázek 4. Konvoluční neuronová síť [19]

- Rekurentní neuronové sítě (RNN). Jedná se o neuronovou síť, ve které je výstup z předchozího kroku přiváděn zpět to vstupní vrstvy aktuálního kroku. V tradičních neuronových sítích jsou všechny vstupy a výstupy na sobě nezávislé, ale v případech, kdy je nutné předpovídat další slovo věty, jsou potřebná předchozí slova, a proto je potřeba si je pamatovat. Nejdůležitější vlastností této sítě je její skrytý stav, který si pamatuje některé informace o sekvenci. Používá stejné parametry pro každý vstup, protože provádí stejnou úlohu na všech vstupech nebo skrytých vrstvách k vytvoření výstupu. [20]



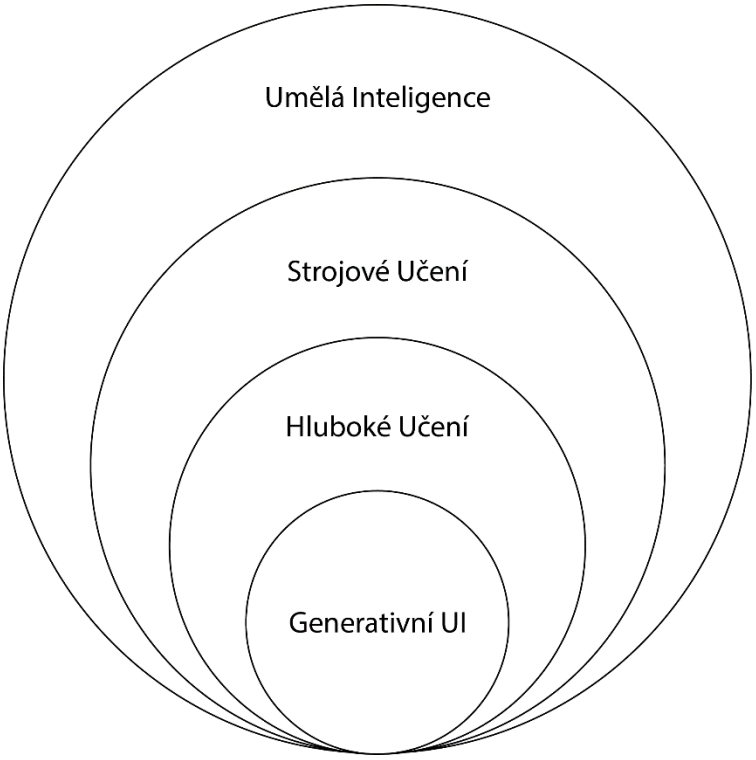
Obrázek 5. Rekurentní neuronová síť [14]

1.2.3.3 Zpracování přirozeného jazyka

Zpracování přirozeného jazyka (NLP) umožňuje počítačům porozumět psanému nebo mluvenému lidskému jazyku. Zahrnuje mnoho různých technik pro interpretaci lidského jazyka jako je například pomocí strojového učení nebo pravidel. Základní princip je ten, že je věta rozdělena na několik malých sekvencí, pomocí kterých se následně snaží pochopit vztahy mezi nimi, pomocí kterých vytvoří sémantický význam. [21]

1.2.3.4 Počítačové vidění

Počítačové vidění (CV) využívá neuronových sítí pro zpracování vstupního obrazu, čímž umožňuje počítačům interpretovat a analyzovat obrázky nebo videa. Funguje víceméně stejným způsobem, jako lidský zrak. Využití CV je například detekce obličeje nebo sledování objektu. [22]



Obrázek 6. Přehled odvětví umělé inteligence [Autor]

1.3 Velké jazykové modely

Velké jazykové modely (LLM) jsou velmi rozsáhlé modely hlubokého učení, které jsou předem trénovány na obrovském množství dat. S pomocí architekturou transformátorů jim je umožněno rozpoznávat, překládat, předvídat nebo generovat text nebo jiný obsah. [23]

Často se často hovoří o tom, že velké jazykové modely jsou to samé co generativní umělá inteligence. Generativní umělá inteligence je termín, který označuje modely umělé inteligence, které mají schopnost generovat jakýkoliv obsah jako je například text nebo kód, ale i videa, obrázky nebo hudbu, zatímco velké jazykové modely pouze spadají do téhle kategorie.

1.3.1 Princip jazykových modelů

Nejprve označíme sadu slovní zásoby jako V a jazykový model jako L_θ. Jazykový model L_θ zpracovává sekvenci tokenů, neboli dotazů, která je označena jako vstup $h := \{h_1, h_2, ..., h_N\}$; $h_i \in V$. Token je zpracováván na vstupu jazykového modelu, ze kterého je následně předpovězen první výstupní token s_0, který opět náleží V. Pro získání následujícího výstupního tokenu s_1 je nutné vzít vstup h a minulý výstupní token s_0 a zadat tuto dvojici jako nový vstupní dotaz $[h, s_0]$. Tento proces se opakuje pro všechny časové kroky. Každý token, který

je generovaný v konkrétním časovém kroku, je označen jako t-token, kde *t* značí číslo časového kroku, při kterém byl token vytvořen, tzn., že token s_0 by měl hodnotu $t = 0$, jelikož byl vytvořen v prvním kroku. Vstupní dotaz pro t-token můžeme definovat jako $[h, s_{1:t-1}]$, kde h značí počáteční vstup, který byl vložen jazykovému modelu na vstup, a $s_{1:t-1}$ značí sekvenci tokenů, které byly vygenerovány jazykovým modelem až po token t_{-1}, neboli všechny tokeny před aktuálním t-tokenem. Pro vytvoření dalšího tokenu pro daný dotaz následně jazykový model vynese na výstup logit vektor ℓt, který v sobě drží hodnoty pravděpodobnosti přiřazené každému z tokenů v sadě slovní zásoby. Zároveň je tento vektor V-dimenzionální, což znamená, že počet dimenzí se rovná počtu tokenů slovní zásoby. Matematický zápis pro logit vektor by tedy vypadal jako $\ell t = L_\theta([h, s_{1:t-1}])$, kde parametrizovaný jazykový model L_θ přivede dotaz $[h, s_{1:t-1}]$ na vstup, po kterém je následně vygenerovaný logit vektor ℓ_t. Poté je využita softmax funkce, která má za účel vzít vektor čísel a převést ho na distribuci pravděpodobností. Toto je provedeno tak, že je každý prvek vstupního vektoru umocněn a poté vydělen součtem všech umocněných hodnot. Tímto získáme pouze hodnoty mezi čísly 0 a 1, což je umožňuje využívat jako hodnoty pravděpodobnosti. Každá hodnota pravděpodobnosti tokenu odpovídá pravděpodobnosti, že právě tento token bude využit další v pořadí. Tento proces lze definovat jako $P_{L\theta}(s_t = \cdot|[h, s_{1:t-1}])$, kde $P_{L\theta}$ značí distribuci pravděpodobnosti možných následujících tokenů s_t pro daný kontext $[h, s_{1:t-1}]$. Pravděpodobnost vzorkovacího tokenu $s_i \in V$ se značí následovně. [24]

$$pt(i) = \frac{\exp(\ell t(i))}{\sum_{v \in V}(\ell t(v))}$$

Nakonec je následující token s_t vygenerován vzorkováním z distribuce pravděpodobnosti p_t. [24]

1.3.2 Transformátorový model

Transformátorový model je nejběžnější architekturou velkého jazykového modelu, ve kterém jsou nejdříve vstupní data zpracovány a pak prochází množstvím vrstev obsahující dopředné neuronové sítě a mechanismus sebevnímání. Sebevnímání je to, co umožňuje modelu transformátoru vzít v úvahu různé části sekvence nebo celý kontext věty a vytvářet předpovědi. [25]

Princip transformátorového modelu začíná tím, že se vstupní text přeloží do číselné podoby, která značí sémantický význam všech tokenů ve vstupu. Následovně je provedeno poziční kódování, které obsahuje specifické vzorky, které dokážou kódovat informaci o pozici. Dále

je využita softmax funkce pro výpočet vah pozornosti v mechanismu sebevědomí, po které jsou vrstvy normalizovány a jsou provedena zbytková spojení. Výstup mechanismu sebevědomí je předán dopředným vrstvám, které provedou nelineární transformace. Často transformátory obsahují na sobě několik vrstev, což modelu umožňuje získat vlastnosti dat. [25]

1.3.3 Architektura velkých jazykových modelů

Velké jazykové modely se skládají z více vrstev neuronové sítě. Vložená vrstva, dopředí vrstva, opakující se vrstva a mechanismus pozornosti spolupracují při zpracování vstupního textu a generování výstupního obsahu. [25]

- Vložená vrstva. Má za úkol převést každé slovo ve vstupním textu na vysoce rozměrnou vektorovou reprezentaci, které pomáhají modelu porozumět kontextu. [25]
- Dopředí vrstva. Obsahuje několik plně propojených vrstev, které aplikují nelineární transformace na vložení vstupu. Tyto vrstvy pomáhají modelu naučit se abstrakce vyšší úrovně ze vstupního textu. [25]
- Opakující se vrstva. Je navržena tak, aby postupně interpretovala informace ze vstupního textu. Tato vrstva udržuje skrytý stav, který se aktualizuje v každém časovém kroku, což umožňuje modelu zachytit závislosti mezi slovy ve větě. [25]
- Mechanismus pozornosti. Umožňuje modelu zaměřit se selektivně na různé části vstupního textu. Tento mechanismus pomáhá modelu věnovat se nejdůležitějším částem vstupního textu a vytvářet přesnější předpovědi. [25]

1.3.4 Trénování velkých jazykových modelů

Trénováním je porozuměno proces výuky velkých jazykových modelů, aby dokázali rozumět lidskému jazyku a zároveň ho generovat. Trénování se skládá z 3 hlavních fází:

- Předtrénink. V této fázi jsou transformátory trénovány na velkém množství nezpracovaných textových dat, kde internet je primárním zdrojem dat. Trénování se provádí pomocí technik učení bez dozoru, které k označení dat nevyžaduje lidskou činnost. Cílem předtréninku je naučit se statistické vzorce jazyka. Nejmodernější strategií pro dosažení lepší přesnosti transformátorů je zvětšení modelu, kterého lze dosáhnout zvýšením počtu parametrů, nebo také zvětšení velikosti trénovacích dat. [26; 23]
- Jemné ladění. V této fázi je použit už předtrénovaný model, na kterém se následně provádí další trénování na menších, specifických sadách dat, aby se zdokonalily jejich schopnosti a zlepšil výkon pro konkrétní úlohu. To znamená, že klíčem k

efektivnímu jemnému ladění jsou další data a trénování. Další datové sady poskytují nové nezpracované informace, zatímco trénování pomáhá uvést tato data do kontextu a pochopit, jak spojit otázky s nejrelevantnějšími a nejvhodnějšími odpověďmi. Trénování je často vykonáváno se zpětnou vazbou nebo systémem hodnocení, který hodnotí reakce umělé inteligence a vede systém k lepším výsledkům. [26; 23]

- Ladění pomocí dotazů. V této fázi jsou vybrány nejefektivnější dotazy, které jsou následně poskytnuty modelu umělé inteligence jako kontext specifický pro úkol. Jedná se o způsob identifikace častějších nebo důležitějších otázek a výcviku umělé inteligence tak, aby efektivněji reagovala na tyto běžné výzvy. Výhodou ladění pomocí dotazů je, že lze tímto způsobem skromněji trénovat modely, aniž by byla přidávána další data, což vede k významné úspoře času a nákladů. [26; 23]

Je důležité poznamenat, že se většinou provádí pouze buď jemné ladění, nebo ladění pomocí dotazů, ale je také možné fáze kombinovat dle potřeby.

2 NÁSTROJE PRO VYTVÁŘENÍ PODVRŽENÝCH TEXTŮ

Rychlý růst nástrojů poháněných umělou inteligencí poskytuje velký potenciál pro společnost a pro vzdělávání. Umělá inteligence dokáže například automatizovat mnoho úkolů, které jsou všední nebo únavné. Nástroje založené na umělé inteligenci s sebou mohou také přinést problémy, se kterými se původně nepočítalo. Generativní umělá inteligence umožňuje, aby si student mohl nechat vytvořit text téměř okamžitě za minimální nebo žádné náklady. Jeden z argumentů, že k takové situaci nebude často docházet, byl, že umělá inteligence nedokáže vytvářet texty, které by se zabývaly stejným tématem, ale byly zároveň odlišně napsané.

Toto tvrzení, bylo nejspíše pravdivé dříve, ale s tím, jak je umělá inteligence v dnešní době vyvinuta, je možné, aby třída 25 studentů dostala 25 různých variant odpovědí na určité téma během pár vteřin. Ovšem se nemusí jednat pouze o nástroje generující texty, ale třeba i nástroje pro parafrázování.

Jedna z hlavních technologií používaných ke generování textu pomocí umělé inteligence je zpracování přirozeného jazyka. Tyto algoritmy umožňují počítačům rozumět různým aspektům lidského jazyka, včetně gramatiky, syntaxe a významu. Díky tréninku na rozsáhlých datových sadách obsahujících knihy, články a webové stránky se modely umělé inteligence naučí identifikovat základní prvky jazyka a využívat je k tvorbě smysluplných vět nebo odstavců. [27]

Pokročilejší modely využívají technik jako je například analýza sentimentu, která umožňuje umělé inteligenci porozumět emocionálnímu stavu textu, což pomáhá umělé inteligenci identifikovat hlavní témata a myšlenky textu. U těchto modelů umělá inteligence dokáže pochytit jemné rozdíly kontextu, a díky tomu jsou schopné generovat text, který se zdá skoro nerozlišitelný od textu psaného člověkem. Tohoto je dosáhnuto díky tokenizaci, která rozebere text na malé části. [28]

Při vytváření textu pomocí nástrojů umělé inteligence můžeme dojít k několika omezením:

- Příliš velká závislost na trénovacích datech. Modely generující text se až příliš opírají o data, na kterých byly trénovány. Pokud jsou trénovací data omezená, zkreslená nebo nedokážou zahrnout celou škálu jazykových variant, může to vést ke zkreslení výsledku nebo třeba k nedostatku rozdílnosti textu. [27]

- Řešení nečekaných situací. Modely generující text mohou mít potíže, pokud se setkají s neobvyklými či vzácnými scénáři, které nebyly dobře zohledněny v trénovacích datech. Takové situace mohou vést k vytváření chybných odpovědí.
- Omezená schopnost chápat kontext. Modely pro generující text často mají obtíže s pochopením širšího kontextu. Vytvářejí text na základě vzorů v trénovacích datech, aniž by skutečně porozuměly významu slov. Tento nedostatek porozumění může mít za následek nepřesnosti, nejasnosti nebo nesmyslné výstupy. [27]

V dnešní době existuje velké množství nástrojů využívající umělé inteligence pro vytváření textů, z nichž jsou mnohé zpoplatněny, proto se budu zabývat pouze těmi, které zpoplatněné nijak nejsou, nebo těmi, které alespoň nabízí bezplatné zkušební období, jelikož jsou právě tyto nástroje těmi, které jsou nejvíce lidmi využívány.

2.1 ChatGPT

GPT-3.5 je bezplatný jazykový model založený na neuronových sítích, který dokáže porozumět a generovat přirozený jazyk nebo kód. Jedná se o jazykový model společnosti OpenAI, který využívá hluboké učení k produkci textu podobného lidskému. Jedná se o jemně vyladěnou verzi GPT-3. Tento model získal významnou pozornost a uznání díky své nepřekonatelné schopnosti porozumět a generovat text podobný lidskému. [29]

GPT -3 Obsahuje 175 miliard parametrů, čímž se stává jedním z nejrozsáhlejších jazykových modelů, které byly kdy vytvořeny v době svého uvedení na trh s tím, že GPT-3.5 má 2x tolik parametrů. Pro srovnání počtu parametrů s ostatními jazykovými modely můžeme využít modelu GPT-4, který má zhruba 1.5 bilionu parametrů a je aktuálně nejrozsáhlejším jazykovým modelem na světě, zatímco model GPT-2 měl kolem 1.5 miliard parametrů. GPT-3.5 dodává konzistentně přesné a relevantní výsledky, zvyšuje standardy pro jazykové modely a prokazuje se jako inovátor v oblasti umělé inteligence. [30]

Charakteristiky GPT-3.5:

1. Porozumění jazyku a vytváření textu – model využívá zpracování přirozeného jazyka k tomu, aby model dokázal porozumět každodennímu jazyku lidí, díky kterému dokáže rozpoznat věci jako sarkasmus nebo slovní hříčky. Dále tento model využívá algoritmy hlubokého učení k pochopení složitých vzorců a vztahů v jazykových datech, což mu umožňuje generovat sofistikovanější odpovědi. [29]

2. Obsáhlá slovní zásoba – Jelikož byl ChatGPT trénován na databázích o velikosti 570 GB, tak má chatbot neskutečně obsáhlou slovní zásobu. [29]
3. Pochopení kontextu – ChatGPT bere v potaz předchozí sekvence zpráv při interakci s uživatelem, pro poskytnutí co nejpřesnějších odpovědí. [29]
4. Schopnost pracovat s více jazyky - Jelikož je model trénován na textových datech 95 různých jazyků, dokáže díky tomu zpracovávat požadavky a generovat odpovědi ve všech různých jazycích. [29]
5. Kreativita – ChatGPT má schopnost kreativně používat lidský jazyk. Model může uživatelům například psát text písní. [29]
6. Schopnost sebezdokolování – Model se stále zlepšuje s postupem času, jelikož se dokáže samostatně vylepšovat díky zpětné vazbě od uživatelů. [29]

Na rozdíl od ostatních chatbotů, kterými se budu zabývat následovně, má GPT-3.5 spíše limitovanou funkcionalitu co se týče uživatelského rozhraní. Uživateli je umožněno komunikovat s chatbotem pouze pomocí manuálního zadávání textu, kde žádné šablony neexistují, a není ani možné přiložit soubor do konverzace.

2.2 Copy AI

Tento nástroj je postaven na velkém jazykovém modelu GPT-3 společnosti OpenAI a je navržen tak, aby pomáhal uživatelům s psaním textů. Nabízí různé nástroje a šablony, které pomáhají začít psát text, a je k dispozici v mnoha jazycích, což ho činí ideálním pro začátečníky, kteří se chtějí seznámit s generativní umělou inteligencí. Má bezplatnou verzi, která je avšak limitována vytvořením pouze 2000 tisíce slov za měsíc. [31]

Charakteristiky nástroje Copy AI:

1. Podpora mnoha jazyků – Přestože není model Copy AI trénovaný na tak rozmanitém množství jazyků jako GPT-3.5, tak je stále schopen pracovat s 25 různými jazyky.
2. Vestavěná kontrola plagiátorství - Tento nástroj analyzuje text vytvořený pomocí Copy AI a porovnává je s již existujícím obsahem v databázi, aby odhalil podobnosti nebo případy plagiátorství. Pokud nástroj identifikuje možné plagiátorství, uživatelé mohou upravit svůj text, aby se ujistili, že jejich obsah je originální.
3. Automatizace pracovního postupu – Jedná se o proces, který využívá umělou inteligenci k provádění a správě zdlouhavých nebo rutinních úkolů. To může zahrnovat různé aktivity, jako je manipulace s daty, jejich zpracování nebo jejich analýza.

Automatizace umožňuje zvýšit efektivitu a produktivitu tím, že tyto opakující se úkoly přenechá počítačům, zatímco se člověk může místo toho soustředit na jiné úkoly.

4. Hlas značky – Tato funkce umožňuje vytvářet obsah v různých hlasových podobách značky, což zajistí, že text vždy osloví cílovou skupinu. Tento nástroj pomáhá udržet konzistentní tón komunikace, ať už jde o různé autory, kanály nebo publikum. Pro využití této technologie stačí nahrát původní obsah dané značky, který Copy AI analyzuje a následně této značce vytvoří přizpůsobený hlas.
5. Předdefinované dotazy – Copy AI umožňuje využití několika desítek předem definovaných dotazů. Jediné co stačí udělat je kliknout na tlačítko „procházet dotazy“, jeden dotaz vybrat a manuálně zadat téma, kterým se má dotaz zabývat.

2.3 Writesonic

Writesonic je pravděpodobně jedním z nejrozsáhlejších nástrojů využívající umělou inteligenci z hlediska funkcionality a zároveň využívá velkého jazykového modelu GPT-4 společnosti OpenAI.

Charakteristiky nástroje Writesonic:

1. Podobně jako u nástroje Copy AI, Writesonic podporuje 24 různých jazyků
2. Generování článků – Tento nástroj je spíše zaměřený pro generování článků a blogů, ale zároveň se dá vygenerovaný text využit pro akademické účely, jelikož nástroj umožňuje uživateli vytvořit fakticky přesné články, které mohou obsahovat až 5000 slov. Tato funkce stojí určitý počet kreditů podle toho, jak rozsáhlou odpověď si uživatel vybere. Každému uživateli je darováno 50 kreditu při založení účtu.
3. Chatsonic – jedná se o chatbot založený na ChatGPT. Bezplatná verze využívá ChatGPT-3.5, zatímco placená verze využívá ChatGPT-4. Je zde možné nahrát textové dokumenty, obrázky, nebo i audiové soubory. Před odesláním dotazu je uživateli doporučeno nechat dotaz upravit pomocí AI, aby dotaz vyhovoval co nejlépe sadě dat, na kterých byl model trénován.
4. Parafráze obsahu – nástroj umožňuje zaměnit text zatímco se zanechá význam. Parafrázovaný text je poté možné vložit do editoru, kde můžeme nechat AI text rozšiřovat.
5. Pro využití jakéhokoliv nástroje je vyžadováno přihlášení, buď vytvořeným účtem přimo ná stránce Writesonic, nebo účtem Google.

3 NÁSTROJE PRO ODHALOVÁNÍ PODVRŽENÝCH TEXTŮ

Detektory umělé inteligence jsou vybaveny sadou trénovacích dat, která obvykle obsahuje jak texty vytvořené člověkem, tak texty vytvořené umělou inteligencí. Detektory tyto texty analyzují, aby dokázali identifikovat nejdůležitější charakteristiky pro rozpoznání textů vytvořených umělou inteligencí.

Dvě hlavní charakteristiky, na které se zaměřují detektory umělé inteligence, jsou:

- Zmatenost. Je to ukazatel, který vyjadřuje, jak moc je model překvapen konkrétním vstupem na základě svého tréninkového souboru dat. Nižší hodnota zmatenosti naznačuje, že je model méně překvapený a díky tomu je schopen lépe předpovídat vstup. [32]

  Představme si, že máme větu „Jsem unavený, jdu si lehnout do…“. Jazykový model s velkou zmateností není schopen předpovídat vstup s velkou spolehlivostí, proto je možné a velice pravděpodobné, že by nabídl slovo, které by bylo v daném kontextu nevhodné, jako je například „televize“. Naopak jazykový model s malou zmateností by s největší pravděpodobností nabídl slovo „postel“, čímž by byla predikce správná.

- Proměnlivost. Jedná se o variaci délky a struktury vět. Obsah vytvořený umělou inteligencí má tendenci být méně proměnlivý a mít menší sklon k výkyvům v délce a struktuře vět ve srovnání s textem napsaným člověkem. Jeden významný znak jazykových modelů je ten, že veškerý text, který vytvoří, se konzistentně podobá textu psaného umělou inteligencí. Zatímco by člověk mohl omylem napsat větu, která vypadá jako věta vytvořená umělou inteligencí, lidé mají tendenci měnit styl skládání vět nebo i slovní zásobu velmi často v průběhu psaní textu. Na druhou stranu, jazykové modely používají stejná pravidla pořád dokola k výběru dalšího slova ve větě, což vede k nízké proměnlivosti. To znamená, že vyšší proměnlivost obvykle naznačuje, že se dokument více podobá textu vytvořenému člověkem. [32]

3.1 Princip detektorů AI textu

Pro detekci textu vytvořeného umělou inteligencí se využívá několika technik a tyto techniky lze zároveň kategorizovat 2 způsoby. První kategorií jsou detektory připravené, které proaktivně zapojují model pro detekci už během generování textu. Zde se často využívá metod vodoznaku nebo metod založených na vyhledávání. Druhou kategorií jsou detektory

pozdější, které provádí detekci až poté, co je text vygenerovaný. Zde patří například metoda zero-shot. [24]

3.1.1 Detekce pomocí vodoznaku

Princip této techniky spočívá v tom, že se model učí vkládat specifický signál nebo identifikátor do generovaného obsahu, čímž vzniká vodoznak. Tento proces obvykle zahrnuje provedení drobných změn v modelu během trénovací fáze, jako jsou změny vah modelu. Poté co model dokončí fázi trénování a je následně nasazen, dokážou specializované algoritmy detekovat přítomnost vodoznaku, který byl skrytě vložen do textu při generování, a tím zjišťovat, zda jsou v textu například vzácné textová spojení, které by mohly napovídat tomu, že byl text vygenerován umělou inteligencí. [33]

3.1.2 Detekce pomocí vyhledávání

Tato metoda vyžaduje přístup k databázi konkrétního modelu a provádí se obnovení dříve vygenerované sekvence, což slouží jako obrana proti útokům parafrázováním. Poskytovatel API nejprve uloží každou sekvenci vygenerovanou pomocí jejich LLM do databáze, do které je poté nabízen uživatelům přístup pomocí API. Uživatelé tak mohou zadávat text jako dotaz přes poskytnuté rozhraní, které následně prohledává celou databázi dříve vygenerovaných textů a snaží se najít sekvenci, která přibližně odpovídá obsahu vstupního dotazu. [24]

Dále je vypočítáno skóre pravděpodobnosti mezi vzorky v databázi a danou sekvencí. Pokud je hodnota pravděpodobnosti vysoká, tak to napovídá tomu, že je daná sekvence vytvořená pomocí stejného LM. Naopak u textu psaného člověkem je předpokládaná hodnota pravděpodobnosti menší. Tuto techniku detekce lze považovat jako jednu z nejspolehlivějších technik detekce AI textů, jelikož je do určité míry odolná vůči útokům parafrázování díky tomu, že lze textové sekvence stále porovnávat na základě sémantických vlastností. [24]

Přestože je daná technika často schopna správně detekovat AI text, tak není dokonalá a pokud je využito rekurzivního parafrázování, při kterém dojde k záměně textu 5x, tak je přesnost detekce snížena o 75%. Tento rapidní pokles může být způsobem 2 důvody. Prvním možným důvodem je to, že opakované parafrázování upravilo text natolik, že byl odstraněn původní význam. Jako druhý možný důvod se uvádí, že byl původní význam zachován a opravdu byl pouze změněn text natolik, že detektor nebyl schopen dále text spolehlivě detekovat. Aktuálně se neví, jaký z těchto důvodů je ten, který způsobuje rapidní pokles v přesnosti detekce. [24]

3.1.3 Zero-shot detekce

Tento způsob detekce nevyžaduje žádný přístup ke vzorkům textu generovaného pomocí AI nebo psaného člověkem, jelikož tato metoda se spoléhá na to, že pokud je text vygenerovaný pomocí AI, tak bude obsahovat nějakou informaci, kterou bude možné detekovat a následně označit. Pro detekci je možné využívat předem trénovaného LM, který se ale může kompletně lišit od modelu, pomocí kterého byl text generován, což je extrémní výhodou oproti předešlým technikám detekce, jelikož je tímto umožněno detektor využívat pro vícero případů. [24]

3.2 ZeroGPT

Pro určení původu textu mohou uživatelé pomocí tohoto nástroje vložit své texty k analýze, kterou složité algoritmy provedou prostřednictvím technologie DeepAnalyse. Nástroj uvádí, že dosahuje přesnosti vyšší než 98% a dokáže rozpoznat texty vytvořené umělou inteligencí v libovolném jazyce. [34]

Charakteristiky ZeroGPT:

1. Detektor nemá stanovený minimální počet znaků nebo slov pro detekci, ale zdá se, že dokáže správně detekovat, až když je na vstupu kolem 50 odlišných slov.
2. Maximální počet znaků je stanoven na 15000.
3. AI ZeroChat-4 & 5 - Jedná se o chatbota podobající se například ChatGPT.
4. Sumarizace - Tato funkce umožňuje za pár vteřin zkrátit libovolnou větu nebo odstavec a dostat výsledný význam. Maximální délka textu pro, jež chcete zkrátit, je 1500 slov. Dále je možné vybrat výstupní délku zkráceného textu.
5. Parafrázování - Tento nástroj dovoluje člověku přepsat veškerý text, s tím, že se zachová originální význam. Maximální délka vstupního textu pro parafrázování je 300 slov.
6. Kontrola gramatiky - Nástroj pro kontrolu gramatiky a pravopisu dokáže zkoumat anglickou gramatiku a vylepšit ji. Detekuje gramatické chyby, pravopis a interpunkci ve psaném obsahu.
7. Počítadlo slov - Jedná se o proces, při kterém se zajišťuje přesnost počtu slov, znaků, ale také i vět. Nástroj se automaticky aktualizuje během psaní nebo vkládání textu, což je užitečné pro dodržování slovních limitů v rámci úkolů nebo článku.

8. Generátor citací – Nástroj umožňuje uživateli vygenerovat určitý styl citace po tom, co zadá všechny potřebné informace o zdroji.
9. Všechny funkce až na jednu lze využít bez jakéhokoliv přihlášení. Pokud chce uživatel využít chatbotu, bude se muset přihlásit buď účtem, který si na této stránce vytvořil, nebo účtem Google.

3.3 GPTZero

GPTZero při detekci textu prochází internetový archiv aby zjistil, zda už text neexistuje. Také využívá klasifikační model, který sleduje text po jednotlivých větách. Nástroj uvádí vysokou přesnost detekce textů vytvořených modelem GPT-4. [35]

Charakteristiky nástroje GPTZero:

1. Minimální počet znaků pro detekci textu je 250.
2. Maximální počet znaků pro detekci je 5000.
3. Uživateli bez účtu je umožněno využít pouze 7 detekcí denně. Pokud si ale uživatel bezplatně založí účet nebo se přihlásí již existujícím Google účtem, daný limit se zvýší na 100 detekcí denně.
4. Deep scan – Nástroj umožňuje provést hlubší detekci textu, přičemž u každé věty zobrazí hodnotu pravděpodobnosti výskytu AI textu.
5. Nástroj také umožňuje vložení dokumentu pro detekci.

3.4 Writer

Charakteristiky nástroje Writer:

1. Minimální počet slov pro detekci je 50.
2. Maximální počet slov pro detekci je 5000.
3. Není stanoven žádný denní limit pro detekci.
4. Po uživateli není vyžadováno žádné přihlášení.
5. Nástroj umožňuje vložení URL adresy, kterou následně analyzuje a obsah identifikovaný jako textový řetězec poté vloží do pole pro detekci.

3.5 Scribbr

Vývojáři nástroje tvrdí, že je nástroj schopen detekovat většinu AI textu. Provedli svou vlastní studii zabývající se přesností detekce AI textů různých nástrojů pro detekci, kde zjistili, že bezplatná verze nástroje Scribbr má 78% přesnost detekce AI textu. [36]

Charakteristiky nástroje Scribbr:

1. Nástroj vyžaduje minimálně 25 slov pro zahájení detekce.
2. Maximální počet slov pro detekci je 500.
3. Tento nástroj také neobsahuje žádné omezení, co se týče denního počtu detekcí.
4. Obsahuje nástroje pro sumarizaci, kontrolu plagiátorství a parafrázování.
5. Nástroj nevyžaduje přihlašení pro využití jakékoliv ze zmíněných funkcionalit.

3.6 DetectGPT

Bezplatné verzi nástroje je umožňěno využívat pouze detektoru AI textu, zatímco všechny ostatní funkce jsou dostupné pouze pro placenou verzi.

Charakteristiky nástroje DetectGPT:

1. Minimální počet znaků je dán 250.
2. Maximální počet znaje stanoven 5000.
3. Nástroj vyžaduje přihlášení buď účtem založený přímo na stránce DetectGPT nebo účtem Google. Uživateli je umožněno zkontrolovat 3000 slov v bezplatné verzi nástroje předtím. Jakmile uživatel vypotřebuje všech 3000 slov, je po něm vyžadováno upgradu na placenou verzi.
4. Je možné vložit URL adresu nebo textový dokument

4 ZJIŠTĚNÉ NEDOSTATKY PŘI ODHALOVÁNÍ TEXTŮ UMĚLÉ INTELIGENCE

Je zjištěn značný pokles ve spolehlivosti detekce AI textu při využití některých z následujících útoků.

4.1 Parafrázování nástrojem

Vstupní textový dotaz můžeme označit jako h, který je zpracováván jazykovým modelem L_θ generující výstupní sekvenci $s = \{s_1, s_2, ..., s_N\}$ o délce N. Model použitý pro útok parafrází označíme G_ϕ. [24]

První typ útoku spočívá v úpravě výstupní sekvence s tak, aby byl získána upravená výstupní sekvence s'. Toho je dosaženo pomocí záměny určitých tokenů v sekvenci s náhradními slovy s tím, že se zachová sémantický význam. Pokud chceme nahradit token s_k ve výstupní sekvenci s, tak je modelu G_ϕ předložen dotaz, aby vygeneroval kandidáty pro náhradu se stejným sémantickým významem jako token s_k. Tyto kandidáty označíme jako $\tilde{s}_k$. Hodnota, která specifikuje maximální dovolené změny, které lze provést při procesu generování nebo úpravy textu před tím, než začne ztrácet původní význam, se označuje ϵ. Následně lze definovat minimalizační framework je definován následovně. [24]

$$s' = \arg\min_{s'} D(s') \; ; \; s'_k \in \{s_k\} \cup \tilde{s}_k,$$

$$\sum_{i=1}^{N} I[s_i \neq s'_i] \leq \epsilon N$$

kde pomocí D je označen detektor, na který je prováděn útok, a pomocí I je označena funkce indikátoru. [24]

Druhý typ útoku spočívá v úpravě vstupního dotazu h tak, aby se posunula distribuce textu generovaného jazykem a byl následně získán upravený vstupní dotaz h', čímž se získává znovu upravená výstupní sekvence s'. Tento proces se provádí pomocí připojení dalšího naučitelného dotazu h_p k originálnímu dotazu h, což nám vytvoří nový dotaz $h' = [h, h_p]$. Pro nalezení vhodného dodatečného dotazu h_p se detektor nechá několikrát dotazovat vstupními dotazy $\{h_1, h_2, ..., h_m\}$, kde m určuje počet dotazů. Vyhledávací funkce se snaží nalézt dodatečný dotaz h_p tak, aby průměrná míra detekce byla minimalizována pro všechny následně vygenerované výstupy, a je definována následovně. [24]

$$\arg\min_{h_p} \frac{1}{n} \sum_{i=1} I\left[D\left(L_\theta([h_i, h_p])\right) \geq \delta\right]$$

4.2 Parafrázování člověkem

Pro útok může sám člověk parafrázovat text označený vodoznakem, aby se vyhnul detekci. Je zjištěno, že parafrázování člověkem dokáže vyprodukovat lepší výsledky, než když je využito parafrázování nástrojem z čeho vyplývá, že text musí být značně delší, pokud bylo na něm bylo provedeno parafrázování člověkem a chceme dále detekovat text označen vodoznakem. [24]

4.3 Generativní útok

Tento způsob útoku spočívá v tom, že útočník zadá modelu, aby použil šifru při generování odpovědi, čímž se kompletně změní sekvence textů, ale zároveň je tato upravená sekvence snadno vrácena útočníkem do původního stavu. [24]

Příklad toho, jak takový útok může vypadat, je například při použití emotikonů. Útočník zadá modelu, aby použil libovolný emotikon po každém vygenerovaném tokenu. Při odstranění těchto emotikonů by došlo k randomizování červeného seznamu následujících tokenů, čimž je možné se úspěšně vyhnout detekci využívající vodoznak. Dalšími příklady útoku jsou například dotazy, aby byl v odpovědi zaměněn veškerý výskyt písmena „s“ písmenem „z“, nebo aby byla odpověď vygenerována pomocí base64 kódování.[24]

Jednou z možných realizovatelných obran proti tomuto útoku je zahrnout všechny zmíněné dotazy už při procesu jemného ladění, aby model věděl, že na takové dotazy nemá odpovídat. [24]

4.4 Copy-paste útok

Při této metodě jsou do dlouhého textu psaného člověkem vkládány kousky AI textu. Je možné upravovat 2 parametry, podle kterých lze nastavovat síla útoku. Prvním parametrem je počet vložených sekvencí AI textu a druhým je podíl dokumentu, který je tvořen AI textem. Jelikož způsoby detekce mají potíže při zpracování textu, který je upraven tímto způsobem, tak je nutné do modelu zahrnout techniku okenního schéma, která umožňuje detekovat menší sekvence textu, které by jinak mohly být přehlédnuty. [24]

4.5 Spoofing útok

Technika útoku je založena na tom, že útočník záměrně vytváří textové úseky takovým stylem, aby byly detekovány a označeny jako texty generované pomocí AI, čímž zhoršuje reputaci vlastníkovi daného modelu. [24]

Při použití tohoto útoku na modely s vodoznakem se nejprve vypočítá odhad zeleného seznamu pro některé z nejvíce používaných slov modelu. Tento odhad slouží jak náhrada skutečného seznamu povolených slov. Opakuje se velký počet dotazování modelu a je pozorován výskyt párových tokenů na výstupu pro generování nezávislého odhadu zeleného seznamu pro každý kontext. Poté, co je zjištěn odhad pro zelené seznamy, může útočník generovat text, který bude následně detekován pomocí vodoznaku a označen jako vytvořený AI díky tomu, že bude do textu vkládat tyto zjištěné tokenové sekvence. [24]

Daný útok lze využít i u detektorů využívající vyhledávání. Nejdříve je lidský text proveden nástrojem parafrázování, a jelikož jsou všechny dřívější sekvence parafrázování uloženy do databáze, tak je tento text taky do databáze uložen, což by způsobovalo to, že by mohl lidský text dosáhnout vysoké podobnosti s parafrázovanými texty v databázi, čímž by detektor falešně označil text jako vytvořený pomocí AI. [24]

II. PRAKTICKÁ ČÁST

5 TESTOVÁNÍ NÁSTROJŮ PRO ODHALOVÁNÍ AI TEXTŮ

V této kapitole budu testovat jednotlivé nástroje pro detekci a to tak, že každému nástroji budou připraveny texty, které byly vytvořené různými způsoby o lišících se délkách. Ke konci této kapitoly bude také vytvořen přehled všech nástrojů s jejich přesností, jakou dokázali nebo nedokázali podvržený text detekovat.

Jako první budu na všech nástrojích testovat čistě vygenerovaný text umělou inteligencí, který by měl jít teoreticky nejjednodušeji odhalit.

5.1 Detekce AI textu

Pro úspěšné vytvoření benchmarku je potřeba vytvořit stupnici přesnosti klasifikace pro texty psané člověkem a texty vygenerované pomocí AI. Pokud text předložený nástroji pro detekci AI textu je vytvořený člověkem, bude se jednat o text negativní. Naopak bude-li nástroji předložen text vygenerovaný pomocí AI, tak budeme hovořit o textu pozitivním.

Tabulka 1. Stupnice přesnosti klasifikace pro texty psané člověkem.

Nástroj klasifikoval lidský text jako	Jedná se o detekci	Zkratka
(80-100%) lidský	Pravdivě negativní	PN
(60-79%) lidský	Částečně pravdivě negativní	ČPN
(40-59%) lidský	Nejasné	NE
(20-39%) lidský	Částečně falešně pozitivní	ČFP
(0-19%) lidský	Falešně pozitivní	FP

Tabulka 2. Stupnice přesnosti klasifikace pro texty generované AI.

Nástroj klasifikoval AI text jako	Jedná se o detekci	Zkratka
(80-100%) lidský	Falešně negativní	FN
(60-79%) lidský	Částečně falešně negativní	ČFN
(40-59%) lidský	Nejasné	NE
(20-39%) lidský	Částečně pravdivě pozitivní	ČPP

(0-19%) lidský	Pravdivě pozitivní	PP

Následně je potřeba daný text vygenerovat, a pro to využiju nástroje ChatGPT, jelikož je nejvyužívanějším z této kategorie nástrojů. Také budu muset používat anglické texty, protože většina nástrojů je specializována právě pro detekci anglických textů a texty ostatních jazyků není ani zdaleka schopna tak přesně rozpoznávat.

Tabulka 3. Generování 5 vět pomocí ChatGPT.

Autor	Generate 5 sentences about why it's important to educate yourself.
ChatGPT	Educating yourself is crucial for personal growth and development. By learning continuously, you expand your knowledge and skills, enabling you to adapt to various situations effectively. Education empowers you to make informed decisions and navigate through life's challenges with confidence. Furthermore, ongoing learning fosters innovation and creativity, opening doors to new opportunities and possibilities. Ultimately, investing in self-education is an investment in your future success and fulfillment.

Následně je text předložen všem zmíněným nástrojům pro detekci.

Tabulka 4. Výsledky detekce 5 vět vytvořených pomocí AI.

Nástroj	Zjištěná hodnota detekce	Výsledek
ZeroGPT	100% AI text	PP
GPTZero	100% AI text	PP
Writer	86% lidský text	FN
Scribbr	100% AI text	PP
DetectGPT	100% AI text	PP

Z daných výsledků jde vidět, že 4 z 5 nástrojů dokázali text správně klasifikovat. Následnovně si nechám vygenerovat AI text, který bude tentokrát dlouhý 10 vět.

Tabulka 5. Generování 10 vět pomocí ChatGPT.

Autor	Generate 10 sentences about why it's important to educate yourself.
ChatGPT	Educating yourself is essential for both personal and professional advancement. Continuous learning broadens your perspective and understanding of the world around you. It equips you with the knowledge and skills necessary to navigate life's complexities and challenges successfully. Furthermore, self-education fosters intellectual curiosity and critical thinking abilities, enabling you to analyze information critically and make well-informed decisions. By investing in your education, you increase your employability and career prospects, positioning yourself for greater opportunities and advancement. Moreover, ongoing learning enhances your confidence and self-esteem, empowering you to pursue your goals with determination and resilience. Education also plays a vital role in personal growth, allowing you to discover new interests, talents, and passions. Additionally, educated individuals tend to lead healthier lifestyles, make better financial decisions, and contribute positively to their communities. Ultimately, lifelong learning is a journey of self-discovery and empowerment, enriching your life in countless ways.

Následně je opět provedena detekce textu všemi 5 nástroji.

Tabulka 6. Výsledky detekce 10 vět vytvořených pomocí AI.

Nástroj	Zjištěná hodnota detekce	Výsledek
ZeroGPT	100% AI text	PP
GPTZero	100% AI text	PP
Writer	84% lidský text	FN
Scribbr	100% AI text	PP
DetectGPT	88% AI text	PP

Jak lze vidět, tak se výsledek ve většině případu vůbec nezměnil, anebo se změnil jen o neznačnou část.

5.2 Detekce AI textu s využitím nástroje pro parafrázování

Nástroj pro parafrázování jsem zvolil opět ChatGPT, jelikož ze všech nástrojů, co dokáží parafrázovat, měl nejvíce přesvědčivý text. Jako text pro parafrázování jsem zvolil předešle vygenerovaný text pomocí ChatGPT. Jako text byl vybrán původně vygenerovaný text.

Tabulka 7. Parafrázování 5 vět pomocí ChatGPT.

Autor	Paraphrase the last 5 sentences about why it's important to educate yourself.
ChatGPT	Continuously educating oneself is essential for personal advancement. Through ongoing learning, individuals enhance their competencies and understanding, equipping them to tackle diverse circumstances adeptly. Education empowers individuals to make well-informed choices and confront life's obstacles with assurance. Additionally, continual learning nurtures creativity and ingenuity, unlocking avenues for growth and exploration. Ultimately, dedicating resources to self-education is an investment in future achievements and contentment.

Text je předložen detektorům textu.

Tabulka 8. Výsledky detekce 5 parafrázovaných vět nástrojem.

Nástroj	Zjištěná hodnota detekce	Výsledek
ZeroGPT	100% lidský text	FN
GPTZero	100% AI text	PP
Writer	93% lidský text	FN
Scribbr	100% AI text	PP
DetectGPT	79% lidský text	ČFN

Z těchto výsledů lze značně vidět, že jednoduchá záměna textu pomocí AI stačí k tomu, aby některé z detektorů přestali detekovat AI text s přesností. Následovně necháme parafrázovat předchozí text, který obsahoval 10 vět, opět pomocí nástroje ChatGPT.

Tabulka 9. Parafrázování 10 vět pomocí ChatGPT.

Autor	Paraphrase the last 10 sentences about why it's important to educate yourself.
ChatGPT	Self-education is crucial for both personal and professional progress. Continuous learning expands your viewpoint and comprehension of the world. It provides you with the knowledge and abilities necessary to navigate life's complexities effectively. Moreover, self-directed learning cultivates intellectual curiosity and enhances critical thinking skills, enabling you to assess information critically and make informed decisions. By investing in your educational journey, you enhance your employability and career opportunities, positioning yourself for advancement. Furthermore, ongoing learning boosts your self-assurance and self-respect, enabling you to pursue your objectives with determination and resilience. Education also plays a pivotal role in personal development, facilitating the discovery of new interests, talents, and passions. Additionally, educated individuals typically lead healthier lives, make wiser financial choices, and contribute positively to their communities. Ultimately, lifelong learning is a voyage of self-exploration and empowerment, enriching your life in myriad ways.

Následně je text představen jednotlivým detektorům.

Tabulka 10. Výsledky detekce 10 parafrázovaných vět nástrojem.

Nástroj	Zjištěná hodnota detekce	Výsledek
ZeroGPT	100% AI text	PP
GPTZero	100% AI text	PP
Writer	86% lidský text	FN
Scribbr	100% AI text	PP
DetectGPT	78% AI text	ČPP

Daný výsledek naznačuje, že kratší délka parafrázovaného textu je v některých případech těžší detekovat, zatímco čím je text delší, tak má detektor větší možnost pozorovat nízkou proměnlivost textu.

5.3 Detekce AI textu parafrázovaného člověkem

V této části provedu vlastní parafrázování AI textu, který mi byl předešle vygenerován.

Tabulka 11. Parafrázování 5 vět člověkem.

Educating ourselves is what allows us to grow and develop. By studying, you obtain more knowledge and skills, which makes you ready to adapt to various situations. Thanks to education, you're able to make reasonable decisions, making it easier to navigate throughout your life. Continuous learning builds up innovation and creativity, opening doors to opportunities. At the end, investing in self-education can be viewed as an investment in your future success and also happiness.

Text je předložen detektorům.

Tabulka 12. Výsledky detekce 5 vět parafrázovaných člověkem.

Nástroj	Zjištěná hodnota detekce	Výsledek
ZeroGPT	100% lidský text	FN
GPTZero	92% AI text	PP
Writer	100% lidský text	FN
Scribbr	100% AI text	PP
DetectGPT	54% lidský text	NE

Pokud porovnáme dané výsledky s výsledky, kdy jsme pro 5 vět použili nástroj pro parafrázování, tak zjišťujeme, že jsme dostali z velké části stejné hodnoty, jen u detektoru DetectGPT došlo k poklesu hodnoty.

Následně provedu vlastní parafrázování stejného AI textu, tentokrát o 10 větách.

Tabulka 13. Parafrázování 10 vět člověkem.

Educating ourselves is what allows us to grow and develop. By studying, you obtain more knowledge and skills, which makes you ready to adapt to lot of new situations. Thanks to education, you're able to make reasonable decisions, which helps with navigating in your life. Continuous learning can help build up creativity, opening doors to opportunities you

wouldn't have otherwise. Investing in self-education can be viewed as an investment in your future success and also happiness. Continuous learning also helps with your confidence and self-esteem, making you able to pursue your goals with determination. Education is also very important when it comes to personal growth, allowing you to discover new interests, talents, or even people. Educated individuals tend to lead healthier lifestyles or make better financial decisions, thanks to their reasonable decisions. At the end of the day, a lot of learning in our lives comes down to self-discovery, which benefits our lives in many ways.

Text je předložen detektorům.

Tabulka 14. Výsledky detekce 10 vět parafrázovaných člověkem.

Nástroj	Zjištěná hodnota detekce	Výsledek
ZeroGPT	83% lidský text	FN
GPTZero	55% lidský text	NE
Writer	100% lidský text	FN
Scribbr	73% AI text	ČPP
DetectGPT	90% AI text	PP

Ve srovnání s výsledky, kdy se pro parafrázování 10 vět použil nástroj, vidíme značný pokles v přesnosti detekce AI textu.

5.4 Detekce textu vytvořeného člověkem

Pro testování použiji text z wikipedie, který se zabývá Evropskou unií, s tím, že se jeho délka podobá textům předešlým.

Tabulka 15. Úryvek anglického textu z wikipedie.

The European Union has seven principal decision-making bodies, its institutions: the European Parliament, the European Council, the Council of the European Union, the European Commission, the Court of Justice of the European Union, the European Central Bank and the European Court of Auditors. Competence in scrutinising and amending legislation

is shared between the Council of the European Union and the European Parliament, while executive tasks are performed by the European Commission and in a limited capacity by the European Council (not to be confused with the aforementioned Council of the European Union). The monetary policy of the eurozone is determined by the European Central Bank. The interpretation and the application of EU law and the treaties are ensured by the Court of Justice of the European Union. The EU budget is scrutinised by the European Court of Auditors. There are also a number of ancillary bodies which advise the EU or operate in a specific area.

Text je představen detektorům.

Tabulka 16. Výsledky detekce lidského textu.

Nástroj	Zjištěná hodnota detekce	Výsledek
ZeroGPT	100% AI text	FP
GPTZero	89% lidský text	PN
Writer	93% lidský text	PN
Scribbr	99% lidský text	PN
DetectGPT	73% lidský text	ČPN

Z daných výsledků je možné usoudit, že pokud se jedná o čistě lidský text, tak je většina detektorů v tomto případě schopna opravdu detekovat, že se jedná pouze o lidský text.

5.5 Detekce textu vytvořeného člověkem za použití překladače

V této části znovu použiju článku z wikipedie, který se zabývá starověkým Římem, a jeho délka opět odpovídá předešlým textům. Tentokrát bude vybraný úryvek textu v češtině, ale bude poté použit Google překladač pro přeložení textu do angličtiny.

Tabulka 17. Úryvek českého textu z wikipedie.

V době svého největšího rozsahu, za císaře Trajána, se moc Říma rozprostírala do všech zemí podél Středozemního moře, dále do Galie, velké části Británie a do oblasti Černého moře. Řím vládl nad většinou zemí tehdy známého světa (*orbis terrārum*). Římem

založená říše posloužila jako nástroj šíření klasické kultury, umění a obchodu do všech jím podmaněných krajů. Tehdejší životní úrovně a počtu obyvatelstva mělo být v Evropě a Africe znovu dosaženo až o mnoho staletí později. Ve východní části impéria došlo k promísení latinské kultury s helénistickými a orientálními elementy. Naproti tomu západ byl zcela romanizován. Římané navíc neměli vliv pouze na území jimi bezprostředně ovládaná, ale i na okolní země, které se nacházely mimo jejich kontrolu.

Následně je daný text přeložen Google překladačem a dostáváme ho v následující podobě:

Tabulka 18. Přeložený český úryvek do angličtiny pomocí překladače.

At its greatest extent, under the emperor Trajan, the power of Rome extended to all the countries along the Mediterranean Sea, and also to Gaul, much of Britain, and the Black Sea region. Rome ruled over most of the countries of the then known world (orbis terrārum). The empire founded by Rome served as a tool for the spread of classical culture, art and trade to all the regions it conquered. The living standards and population numbers of that time were to be reached again in Europe and Africa many centuries later. In the eastern part of the empire there was a mixing of Latin culture with Hellenistic and Oriental elements. In contrast, the west was completely Romanized. In addition, the Romans had influence not only on the territories directly controlled by them, but also on the surrounding countries that were outside their control.

Následně je provedena detekce nástroji.

Tabulka 19. Výsledky detekce lidského přeloženého textu.

Nástroj	Zjištěná hodnota detekce	Výsledek
ZeroGPT	100% lidský text	PN
GPTZero	99% lidský text	PN
Writer	91% lidský text	PN
Scribbr	76% lidský text	ČPN
DetectGPT	100% AI text	FP

Výsledky ukazují, že jsou v tomto případě detektory schopny správně detekovat lidský text, který byl přeložen.

5.6 Přehled výsledků

Jelikož také z velké části záleží na tom, o jaký typ textu se jedná, například jestli se jedná o dopis nebo informativní text, nebo jestli je dané téma aktuální nebo 100 roků staré, tak by dosavadní zjištěné výsledky nebyly dostačující k posouzení, zda dokáží nástroje spolehlivě detekovat AI text. Proto jsem provedl minulé testování znovu, ale na 9 různých textech, abych dostal spolehlivější výsledek, díky kterému budu moct posoudit přesnost těchto nástrojů. To znamená, že bylo provedeno celkově dalších 45 testů, při kterých byly vybrány co nejodlišnější relevantní texty, což nám dělá celkově 53 provedených testů. Jediný rozdíl v následujícím testování je ten, že nebyly vždy použity 2 odlišné délky textů u prvních 3 detekcí, ale byla použita jen jedna, která se vždy lišila.

Kdybychom dostali pouze celé výsledky, kterými jsou PN, PP, FN nebo FP, a neměli bychom žádné částečné výsledky, mohli bychom využít následujícího vzorce pro zjištění přesnosti využitých detektorů:

$$\text{PŘESNOST} = \frac{\text{PN} + \text{PP}}{\text{PN} + \text{PP} + \text{FN} + \text{FP}}$$

Jelikož naše výsledky obsahují i částečné výsledky, tak musíme vybrat z některých postupů výpočtu přesnosti, které řeší i částečné výsledky. Využiji binární postupu, ve kterém jsou všechny částečně pravdivé výsledky brány jako nepravdivé. Vzorec daného postupu vypadá následovně:

$$\text{PŘESNOST\_BIN} = \frac{\text{PN} + \text{PP}}{\text{PN} + \text{ČPN} + \text{PP} + \text{ČPP} + \text{FN} + \text{ČFN} + \text{FP} + \text{ČFP} + \text{NE}}$$

Protože je šířka tabulky omezená, tak jsou zkráceny následující označení:

- AI – Čistě AI text.
- AI_P – AI text s využitím nástroje parafrázování.
- AI_PČ – AI text parafrázovaný člověkem.
- L – Čistě lidský text.
- LPŘ – Lidský text s využitím nástroje pro překlad.

V každé buňce je maximální číslo 10, jelikož každý nástroj testoval právě 10 textů, a toto číslo uvádí počet textů, které byly správně identifikovány. Dále je uvedený celkový počet

správně klasifikovaných textů, přesnost, a v posledním sloupci také pořadí, v jakém jsou výsledně hodnoceny všechny nástroje odhalování podle jejich přesnosti detekce.

Tabulka 20. Výsledek binárního hodnocení

Nástroj	AI	AI_P	AI_PČ	L	LPŘ	Celkem	Přesnost	Pořadí
ZeroGPT	10	6	5	6	7	34	68%	4.
GPTZero	10	7	6	10	8	41	82%	1.
Writer	7	5	4	9	7	32	64%	5.
Scribbr	10	8	6	7	6	37	74%	3.
DetectGPT	9	7	5	9	8	38	76%	2.
Průměr	92%	66%	56%	88%	76%			

Pokud porovnám zjištěné výsledky s výsledky práce “Testing of Detection Tools for AI-Generated Text” [37], tak lze vidět, že mé výsledky jsou značně pozitivnější. U nástrojů, které by využity v obou pracích, má druhá práce následující zjištěné přesnosti:

- ZeroGPT – 59%
- GPTZero – 54%
- Writer – 50%
- DetectGPT – 46%

Tento značný rozdíl může být způsoben odlišným způsobem testování nebo používáním značně odlišného textu. Nicméně je z našich výsledků poznat, že pokud se testují detektory na datech, které nejsou úplně nové a už existují nějakou dobu na internetu, tak jsou schopny z větší přesností texty detekovat. Jelikož není přesnost dokonalá, tak stále dochází k případům, kdy jsou texty vytvořené lidmi falešně klasifikovány jako AI texty. V této práci nebyl testován obří počet textů na detektorech a byly využívány texty, které už byly pravděpodobně jednou, ne-li vícekrát testovány. Proto rozdíl mezi mými výsledky a výsledky zmíněné druhé práce dávají smysl, jelikož v dané práci mohlo být využito obrovského množství textu, který nemusel být ani veřejně v té době publikovaný.

5.7 Možné řešení pro zvýšení přesnosti detekce

V dnešní době není možné spolehlivě detekovat text vytvořený umělou inteligencí, jelikož jsou dnešní detektory háklivé na řadu útoků. Všechny existující návrhy využívají už existujících metod detekce a dále na nich budují frameworky, ale stále nejsou ani zdaleka schopny 100% přesnosti detekce.

Ghosal a výzkumníci [24] přišli na návrh, při kterém by bylo využito sémantického vodoznaku. Tato technika vodoznaku by se na rozdíl od všech ostatních metod vodoznaku, které v dnešní době máme, lišila tím, že by sledovala význam konkrétního textu místo toho, aby pouze sledovala jeho formu. Jelikož jsou dnešní techniky vodoznaku náchylné na různé útoky, které mění text, ale zanechávají sémantický význam, tak by bylo potřeba využití právě metody, která by pracovala na vysoké úrovni sémantiky. Ovšem není zřejmé, zda je v dnešní době možné dosáhnout takového řešení, a spíše se návrh stává tématem pro budoucí výzkumy jako všechny ostatní návrhy.

ZÁVĚR

V teoretické části byl nejdřív představen úvod do umělé inteligence, ve které byly popsány její různé typy a odvětví. Byly popsány velké jazykové modely a jejich princip generování textu. Dále byly uvedeny charakteristiky nástrojů pro generování a detekci AI textů, kde byly mimo jiné popsány techniky detekce. V neposlední řadě byly zjištěny nedostatky, které způsobují nepřesnost při klasifikaci textu.

V praktické části bylo provedeno testování 5 velmi často používaných bezplatných detektorů, na kterých byly testovány anglické texty čistě vytvořené pomocí AI, vytvořené pomocí AI a následně parafrázované pomocí AI nástroje, vytvořené pomocí AI a následně parafrázované člověkem, vytvořené člověkem a jako poslední byly testovány texty vytvořené člověkem, které nebyly anglické, ale byly následně přeloženy do angličtiny. Z tohoto testování byl vytvořen přehled výsledků, který byl následně srovnán s výsledky jiné práce pro porovnání. Poslední kapitola této části se zabývá možnými řešeními, které by mohly vést k přesnější detekci AI textu.

V průběhu shromažďování informací k této práci jsem zjistil, že v dnešní době není možné spolehlivě detekovat AI text. Původně bylo v plánu vymyslet své vlastní řešení, jak vyřešit tento problém, ale jelikož i samotní vývojáři od společnosti OpenAI uznali, že v dnešní době není možné spolehlivě text detekovat, tak jsem rychle zjistil, že by mé úsilí bylo marné. Jedná se o boj mezi nástroji pro generování a detekci textu, kde nástroje pro generování textu mají velký náskok a zatím nejsou žádné náznaky, že by se situace měla v blízké době změnit.

SEZNAM POUŽITÉ LITERATURY

[1] Coursera Staff. *What Is Artificial Intelligence? Definition, Uses, and Types*. Online. In: Coursera Articles. Dostupné z: https://www.coursera.org/articles/what-is-artificial-intelligence. [cit. 2024-3-3].

[2] SMITH, Chris; MCGUIRE, Brian; HUANG, Ting; YANG, Gary. *The history of Artificial Intelligence*. Online. Dec 2006. Dostupné z: https://courses.cs.washington.edu/courses/csep590/06au/projects/history-ai.pdf. [cit. 2024-3-3].

[3] REYNOSO, Rebecca. *A Complete History of Artificial Intelligence*. Online. In: G2 Articles. May 25 2021. Dostupné z: https://www.g2.com/articles/history-of-artificial-intelligence. [cit. 2024-3-3].

[4] Simplilearn. *Types of Artificial Intelligence You Should Know in 2024.* Online. In: Simplilearn Articles. Dostupné z: https://www.simplilearn.com/tutorials/artificial-intelligence-tutorial/types-of-artificial-intelligence. [cit. 2024-3-3].

[5] SHTIA, Hussein. *Reactive Machines AI: The Foundation of Modern Artificial Intelligence*. Online. In: LinkedIn Posts. Jan 6 2024. Dostupné z: https://www.linkedin.com/pulse/reactive-machines-ai-foundation-modern-artificial-hussein-shtia-sowaf. [cit. 2024-3-4].

[6] Coursera Staff. *4 Types of AI: Getting to Know Artificial Intelligence.* Online. In: Coursera Articles. Dostupné z: https://www.coursera.org/articles/types-of-ai. [cit. 2024-3-4].

[7] JORGE, *Henrique. Self-Awareness in Artificial Intelligence*. Online. In: Medium News. Aug 15 2023. Dostupné z: https://henriquejorge.medium.com/self-awareness-in-artificial-intelligence-9a7e214b584. [cit. 2024-3-4].

[8] CRABTREE, Matt. *What is Machine Learning? Definiton, Types, Tools & More*. Online. In: Datacamp Blog. Dostupné z: https://www.datacamp.com/blog/what-is-machine-learning. [cit. 2024-3-4].

[9] *What is machine learning (ML)?*. Online. In: IBM Topics. Dostupné z: https://www.ibm.com/topics/machine-learning. [cit. 2024-3-4].

[10] ALL, Moez. *Supervised Machine Learning*. Online. In: Datacamp Blog. Aug 2022. Dostupné z: https://www.datacamp.com/blog/supervised-machine-learning. [cit. 2024-3-4].

[11] *Unsupervised Machine Learning*. Online. In: Learning Platform Javatpoint. Dostupné z: https://www.javatpoint.com/unsupervised-machine-learning. [cit. 2024-3-4].

[12] BAJAJ, Prateek. *Reinforcement learning.* Online. In: GeeksforGeeks Articles. Dostupné z: https://www.geeksforgeeks.org/what-is-reinforcement-learning/. [cit. 2024-3-5].

[13] *What is Deep Learning?*. Online. In: Amazon Articles. Dostupné z: https://aws.amazon.com/what-is/deep-learning/. [cit. 2024-3-5].

[14] *What is a neural network?*. Online. In: Cloudflare Resources. Dostupné z: https://www.cloudflare.com/learning/ai/what-is-neural-network/. [cit. 2024-3-7].

[15] BANOULA, Mayank. *What is Perceptron: A Beginners Guide for Perceptron*. Online. In: Simplilearn Articles. Dostupné z: https://www.simplilearn.com/tutorials/deep-learning-tutorial/perceptron. [cit. 2024-3-9].

[16] *Feed Forward Neural Network*. Online. In: DeepAI Documents. Dostupné z: https://deepai.org/machine-learning-glossary-and-terms/feed-forward-neural-network. [cit. 2024-3-10].

[17] *What are convolutional neural networks?*. Online. In: IBM Topics. Dostupné z: https://www.ibm.com/topics/convolutional-neural-networks. [cit. 2024-3-10].

[18] GeeksforGeeks. *Introduction to Convolution Neural Network*. Online. In: GeeksforGeeks Articles. Dostupné z: https://www.geeksforgeeks.org/introduction-convolution-neural-network/. [cit. 2024-3-14].

[19] Sky Engine AI. *What is a Convolutional Neural Network?.* Online. In: Sky Engine AI Developer Blog. Dostupné z: https://skyengine.ai/se/skyengine-blog/125-what-is-a-convolutional-neural-network. [cit. 2024-3-14].

[20] KALITA, Debasish. *A Brief Overview of Recurrent Neural Networks (RNN)*. Online. In: AnalyticsVidhya Blog. Feb 7 2024. Dostupné z: https://www.analyticsvidhya.com/blog/2022/03/a-brief-overview-of-recurrent-neural-networks-rnn/. [cit. 2024-3-16].

[21] *Natural Language Processing (NLP).* Online. In: SAS Insights. Dostupné z: https://www.sas.com/en_us/insights/analytics/what-is-natural-language-processing-nlp.html. [cit. 2024-3-19].

[22] SATYAM, Kumar. *Computer Vision Tutorial*. Online. In: GeeksforGeeks Articles. Dostupné z: https://www.geeksforgeeks.org/computer-vision/. [cit. 2024-3-25].

[23] *What is Large Language Model (LLM)?*. Online. In: Amazon Articles. Dostupné z: https://aws.amazon.com/what-is/large-language-model/. [cit. 2024-3-26].

[24] GHOSAL, Souma Suvra; CHAKRABORTY, Souradip; GEIPING, Jonas; HUANG, Furong; MANOCHA, Dinesh; BEDI, Amrit Singh. *Towards Possibilities & Impossibilities of AI-generated Text Detection: A Survey*. Online. Oct 23 2023. Dostupné z: https://arxiv.org/pdf/2310.15264. [cit. 2024-3-30].

[25] *What is a transformer model?*. Online. In: IBM Topics. Dostupné z: https://www.ibm.com/topics/transformer-model. [cit. 2024-4-3].

[26] DROST, Dorian. *Different ways of training LLMs*. Online. In: Medium News. Jul 21 2023. Dostupné z: https://towardsdatascience.com/different-ways-of-training-llms-c57885f388ed. [cit. 2024-3-28].

[27] HEWITT, Craig. *The Complete Guide to AI Text Generators for Creators (How They Work, Limitations, and How to Use Them).* Online. In: Castos Blog. Dostupné z: https://castos.com/ai-text-generators/. [cit. 2024-4-2].

[28] AIContentfy Team. *The Step-by-Step Text Generation Process Demystified*. Online. In: AIContentfy Blog. Nov 6 2023. Dostupné z: https://aicontentfy.com/en/blog/step-by-step-text-generation-process-demystified. [cit. 2024-4-2].

[29] KANADE, Vijay. *What Is ChatGPT? Characteristics, Uses, and Alternatives*. Online. In: Spiceworks News. May 16 2023. Dostupné z: https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-chatgpt/. [cit. 2024-4-7].

[30] EMMANUEL, Chude. *GPT-3.5 and GPT-4 Comparison*. Online. In: Medium News. Aug 4 2023. Dostupné z: https://medium.com/@chudeemmanuel3/gpt-3-5-and-gpt-4-comparison-47d837de2226. [cit. 2024-4-13].

[31] MCLEAN, Deanna. *What is Copy.ai and How to Use It (2024 Guide)*. Online. In: ElegantThemes Blog. Dostupné z: https://www.elegantthemes.com/blog/marketing/copy-ai. [cit. 2024-4-19].

[32] CAULFIELD, Jack. *How Do AI Detectors Work? | Methods & Reliability*. Online. In: Scribbr Knowledge Base. Sep 6 2023. Dostupné z: https://www.scribbr.com/ai-tools/how-do-ai-detectors-work/. [cit. 2024-4-23].

[33] CRAIG, Lev. *AI watermarking*. Online. In: TechTarget Topics. Dostupné z: https://www.techtarget.com/searchenterpriseai/definition/AI-watermarking. [cit. 2024-4-23].

[34] GILLHAM, Jonathan. *ZeroGPT AI Content Detector Review*. Online. In: Originality.ai Blog. Apr 22 2024. Dostupné z: https://originality.ai/blog/zerogpt-ai-content-detector-review. [cit. 2024-4-26].

[35] *Our detection technology*. Online. In: GPTZero Detector. Dostupné z: https://gptzero.me/technology. [cit. 2024-4-26].

[36] *Frequently asked questions.* Online. In: Scribbr Detector. Dostupné z: https://www.scribbr.com/frequently-asked-questions/. [cit. 2024-4-26].

[37] WEBER-WULFF, Debora; ANOHINA-NAUMECA, Alla; BJELOBABA, Sonja; FOLTÝNEK, Tomáš; GUERRERO-DIB, Jean; POPOOLA, Olumide; ŠIGUT, Petr; WADDINGTON, Lorna. *Testing of Detection Tools for AI-Generated Text*. Online. Jun 21 2023. Dostupné z: https://arxiv.org/pdf/2306.15666. [cit. 2024-5-3].

SEZNAM POUŽITÝCH SYMBOLŮ A ZKRATEK

AI	Umělá inteligence
ML	Strojové učení
DL	Hluboké učení
ANI	Úzká umělá inteligence
AGI	Obecná umělá inteligence
ASI	Superinteligentní umělá inteligence
NN	Neuronová síť
FNN	Dopředná neuronová síť
CNN	Konvoluční neuronová síť
RNN	Rekurzivní neuronová síť
NLP	Zpracování přirozeného jazyka
CV	Počítačové vidění
LLM	Velký jazykový model
L_θ	Jazykový model
V	Slovní zásoba
h	Vstupní sekvence tokenů
h_p	Dodatečný dotaz
s	Výstupní sekvence tokenů
t	Časový krok
ℓ_t	Logit vektor
p_t	Distribuce pravdepodobnosti
N	Délka výstupní sekvence
G_ϕ	Model pro parafrázování
s_k	Náhradní token

ϵ Maximální dovolená změna

D Nástroj pro detekci

I Funkce indikátoru

SEZNAM OBRÁZKŮ

SEZNAM TABULEK