

Technická univerzita v Košiciach
Fakulta elektrotechniky a informatiky

Využitie transformerov v detekcii toxicity na sociálnych sieťach

Bakalárska práca

2025

Artem Mykhailichenko

Technická univerzita v Košiciach
Fakulta elektrotechniky a informatiky

Využitie transformerov v detekcii toxicity na sociálnych sieťach

Bakalárska práca

Študijný program: Inteligentné systémy
Študijný odbor: Informatika
Školiace pracovisko: Katedra kybernetiky a umelej inteligencie (KKUI)
Školiteľ: prof. Ing. Kristína Machová, PhD.
Konzultant:

Košice 2025

Artem Mykhailichenko

Abstrakt v SJ

Práca sa zaoberá problematikou detekcie toxicity v online komunikácii so zameraním na sociálne siete. Úvodná časť je venovaná teoretickej analýze problému, v rámci ktorej boli analyzované súčasné vedomosti v oblasti spracovania prirodzeného jazyka (natural language processing, NLP) a obširna vedecká literatúra týkajúca sa automatického klasifikovania toxického obsahu. Na základe týchto vedomostí boli systematicky opísané rôzne prístupy – od klasických algoritmov strojového učenia až po hlboké neurónové siete a transformátorové modely s mechanizmom sebavnímania (self-attention). Praktická časť práce obsahovala sériu experimentov, počas ktorých bolo otestovaných niekoľko modelov z rôznych kategórií s cieľom zistiť najefektívnejšie prístupy. Na základe získaných výsledkov boli vybrané najefektívnejšie modely BERT (Bidirectional Encoder Representations from Transformers), RoBERTa (Robustly Optimized BERT Pretraining Approach) a Tiny-Toxic-Detector (kompaktný transformerový model optimalizovaný na detekciu toxicity s nízkymi výpočtovými požiadavkami). Tieto boli následne implementované do webovej aplikácie, ktorá umožňuje porovnávať ich výstupy na rôznych typoch vstupných textov. Pri testovaní sa zaznamenávali kľúčové metriky ako presnosť (accuracy), F1-skóre, citlivosť (recall) a latencia. Výsledkom aplikácie týchto nástrojov bolo vytvorenie funkčného webového rozhrania, ktoré umožňuje koncovému používateľovi zadávať vlastné texty a získať hodnotenie miery toxicity na základe výpočtov viacerých modelov naraz. Celý systém predstavuje prepojenie teoretických poznatkov s praktickou implementáciou, pričom využíva overené NLP prístupy na spracovanie reálnych textových dát s cieľom pozrieť sa na výsledky modelov v rôznych situáciách.

Kľúčové slová

Detekcia toxicity, transformer, NLP, strojové učenie, bakalárska práca

Abstrakt v AJ

This thesis deals with the issue of toxicity detection in online communication with a focus on social networks. The introductory part is devoted to a theoretical analysis of the problem, which includes an analysis of current knowledge in the field of natural language processing (NLP) and extensive scientific literature on automatic classification of toxic content. Based on this knowledge, various approaches were systematically described, ranging from classic machine learning algorithms to deep neural networks and transformer models with self-attention mechanisms. The practical part of the thesis included a series of experiments in which several models from different categories were tested to identify the most effective approaches. Based on the results obtained, the most effective models were selected: BERT (Bidirectional Encoder Representations from Transformers), RoBERTa (Robustly Optimized BERT Pretraining Approach), and Tiny-Toxic-Detector (a compact transformer model optimized for toxicity detection with low computational requirements). These were then implemented in a web application that allows their outputs to be compared on different types of input texts. Key metrics such as accuracy, F1 score, recall, and latency were recorded during testing. The result of applying these tools was the creation of a functional web interface that allows end users to enter their own texts and obtain a toxicity rating based on calculations from multiple models simultaneously. The entire system represents a combination of theoretical knowledge and practical implementation, using proven NLP approaches to process real text data in order to examine the results of models in different situations.

Klíčové slová v AJ

Toxicity detection, transformer, NLP, machine learning, bachelor thesis

78055

TECHNICKÁ UNIVERZITA V KOŠICIACH
FAKULTA ELEKTROTECHNIKY A INFORMATIKY
Katedra kybernetiky a umelej inteligencie

ZADANIE BAKALÁRSKEJ PRÁCE

Študijný odbor: **Informatika**
Študijný program: **Inteligentné systémy**

Názov práce:

Využitie transformerov v detekcii toxicity na sociálnych sieťach
The use of transformers in toxicity detection on social networks

Študent: **Artem Mykhailichenko**
Školiteľ: **prof. Ing. Kristína Machová, PhD.**
Školiace pracovisko: **Katedra kybernetiky a umelej inteligencie**
Konzultant práce:
Pracovisko konzultanta:

Pokyny na vypracovanie bakalárskej práce:

1. Podat' stručný prehľad o probléme detekcie online toxicity.
2. Analyzovať súčasný stav v oblasti transformerov a zvoliť vhodné metódy pre danú úlohu.
3. Vyhľadať, analyzovať a pred-spracovať dáta pre danú úlohu.
4. Použiť zvolené metódy na naučenie modelov na detekciu toxicity.
5. Generované modely testovať a vyhodnotiť.
6. Vypracovať dokumentáciu podľa pokynov vedúcej práce.

Jazyk, v ktorom sa práca vypracuje: slovenský
Termín pre odovzdanie práce: 23.05.2025
Dátum zadania bakalárskej práce: 31.10.2024

prof. Ing. Liberios Vokorokos, PhD.
dekan fakulty

Čestné vyhlásenie

Vyhlasujem, že som diplomovú prácu vypracoval(a) samostatne s použitím uvedenej odbornej literatúry.

Košice 30. 5. 2025

.....

Vlastnoručný podpis

Podakovanie

Ďakujem vedúcemu bakalárskej práce za poskytnutie študijných materiálov a odborné konzultácie, ktoré mi pomohli lepšie pochopiť tému a nasmerovali ma pri ďalšom postupe.

Predhovor

Téma využitie transformerov v detekcii toxicity na sociálnych sieťach bola zvolená vzhľadom na jej aktuálnosť. Cieľom práce je výskum a porovnanie modelov spracovania prirodzeného jazyka (NLP), ktoré sú vhodné na klasifikáciu toxických výrokov v textoch. V rámci riešenia bolo analyzovaných viacero vedeckých prác, boli otestované rôzne modely, ktoré boli prispôbované danej úlohe, a bola vytvorená webová aplikácia, ktorá umožňuje porovnávať najvhodnejšie transformátorové modely v reálnom čase.

Obsah

| | |
|--------------------------------------------------------------------------------------------------------------------------------------|----------|
| Úvod | 1 |
| 1 Formulácia úlohy | 3 |
| 2 Teoretický základ riešenej problematiky | 4 |
| 2.1 Online toxicita ako spoločenský problém | 4 |
| 2.1.1 Definícia a formy toxického správania | 5 |
| 2.1.2 Vplyv anonymity a algoritmov sociálnych sietí | 6 |
| 2.1.3 Potreba automatizovanej detekcie | 7 |
| 2.2 NLP prístupy k detekcii toxicity | 8 |
| 2.2.1 Klasické modely (Logistic Regression, SVM, Naive Bayes) . . . | 9 |
| 2.2.2 Neuronové siete (CNN, RNN, LSTM) | 10 |
| 2.2.3 Transformery a self-attention architektúra | 11 |
| 2.3 Porovnanie najznámejších transformerov | 12 |
| 2.3.1 BERT | 13 |
| 2.3.2 RoBERTa | 14 |
| 2.3.3 DistilBERT | 15 |
| 2.3.4 GPT a iné veľké jazykové modely | 15 |
| 2.4 Súvisiace práce | 16 |
| 2.4.1 Civil Rephrases of Toxic Texts With Self-Supervised Trans- formers (Laugier et al., 2021) (7) | 16 |
| 2.4.2 Social Media Toxicity Classification Using Deep Learning: Real- World Application UK Brexit (Fan et al., 2021)(5) | 17 |
| 2.4.3 GTH-UPM at DETOXIS-IberLEF 2021 (Romero et al., 2021)(14) | 18 |
| 2.4.4 DeTox at GermEval 2021: Toxic Comment Classification – Mina Schütz et al., november 2021(1) | 19 |
| 2.4.5 Toxicity detection in online Georgian discussions (Lashkarash- vili & Tsintsadze, 2022)(9) | 19 |

| | | |
|----------|---------------------------------------------------------------------------------------------------------------------------------------|-----------|
| 2.4.6 | Hate Speech and Toxic Comment Detection using Transformers – Pierre Guillaume et al., apríl 2022(8) | 20 |
| 2.4.7 | A Comparative Study of Attention-Based Transformer Networks and Traditional ML Methods – Sihao Wang, Bingjie Chen, september 2023 (6) | 21 |
| 2.4.8 | Comparison between Machine Learning and Deep Learning Approaches – Andrea Bonetti et al., máj 2023(13) | 21 |
| 2.4.9 | Comparing Different Transformer Models' Performance for Identifying Toxic Language – Carl Sundelin, jar 2023(3) | 22 |
| 2.4.10 | Lightweight Toxicity Detection in Spoken Language – Ahlam Husni Abu Nada et al., apríl 2023(10) | 23 |
| 2.4.11 | COGNITIVE METHOD TO DETECT TOXIC COMMENTS IN SOCIAL MEDIA – Jomy Joseph et al., marec 2024(15) | 24 |
| 2.4.12 | TINY-TOXIC-DETECTOR: A Compact Transformer-Based Model – Michiel Kamphuis, august 2024(16) | 24 |
| 2.4.13 | The Impact of Transformer Models on Detecting Hate Speech and Toxicity – Jiawei Li, Yuqing Xie, 2024(17) | 25 |
| 2.4.14 | A Systematic Review of Toxicity in Large Language Models – Guillermo Villate-Castillo et al., júl 2024(4) | 26 |
| 2.5 | Dátové sady používané na detekciu toxicity | 27 |
| 2.5.1 | Jigsaw Toxic Comment Classification Dataset | 28 |
| 2.5.2 | Civil Comments | 29 |
| 2.5.3 | ToxiGen | 30 |
| 2.6 | Výzvy a etické aspekty | 32 |
| 3 | Praktická realizácia riešenia | 33 |
| 3.1 | Cieľ praktickej časti | 33 |
| 3.2 | Práca s dátami | 34 |
| 3.3 | Práca s modelmi | 34 |

| | | |
|----------|---------------------------------------------------------------|-----------|
| 3.4 | Metódy detekcie toxicity | 36 |
| 3.5 | Mechanizmus self-attention | 37 |
| 3.6 | Hodnotiace metriky | 38 |
| 3.7 | Popis webového riešenia | 40 |
| 3.7.1 | Architektúra systému | 40 |
| 3.7.2 | Použité technológie | 41 |
| 3.7.3 | Všetky funkcie webovej aplikácie | 42 |
| 3.7.4 | Celkový výhľad aktívne používanej webovej aplikácie | 46 |
| 3.7.5 | Záver a možnosti rozšírenia | 46 |
| 3.7.6 | Testovanie a vyhodnotenie výstupov modelov | 47 |
| 4 | Záver | 49 |

Zoznam obrázkov

| | | |
|-----|----------------------------------------------------------------------|----|
| 3–1 | Schéma architektúry webovej aplikácie | 41 |
| 3–2 | Pole pre zadanie textu na kontrolu | 42 |
| 3–3 | Aktívne tlačidlo | 42 |
| 3–4 | Neaktívne tlačidlo | 43 |
| 3–5 | Pole výsledok | 43 |
| 3–6 | Príklad toho, ako môžu vyzerat farby výsledkov | 44 |
| 3–7 | Príklad toho, ako môže vyzerat história výsledkov | 45 |
| 3–8 | Príklad správania tlačidiel za predpokladu, že história je prázdna . | 45 |
| 3–9 | Webová aplikácia | 46 |

Zoznam tabuliek

| | |
|-----------------------------------------------------------------------|----|
| 3–1 Ukážka výsledkov detekcie toxicity pre rôzne modely | 46 |
| 3–2 Výsledky testovania toxicity modelmi Tiny, RoBERTa a BERT | 47 |

Úvod

Každým rokom sociálne siete začínajú nahrádzať osobnú komunikáciu, kde si vymieňame všetko, od bežných každodenných informácií až po politickú sféru našich krajín, ktorá po zverejnení na internete môže formovať verejný názor. Vzhľadom na anonymitu mnohých účtov však ľudia môžu otvorene vyjadrovať svoje skutočné emócie a názory, čo môže viesť k urážkam, zastrašovaniu, hrozbám a šíreniu dezinformácií. To všetko má vážne sociálne a psychologické následky, vrátane vytvárania nepravdivých predstáv o realite (2).

Na riešenie týchto problémov sa už niekoľko rokov používajú automatizované modely filtrovania a moderovania komentárov alebo iného obsahu na internete založené na spracovaní prirodzeného jazyka (NLP). Jedným z najvýznamnejších úspechov v tejto oblasti je vznik transformerov – modelov hlbokého učenia, ktoré využívajú mechanizmy sebavnímania (Self-Attention). Architektúry ako BERT, RoBERTa a ich špecializované verzie preukázali vysokú účinnosť pri analýze textov a odhaľovaní toxického obsahu (6).

Význam týchto metód sa prejavuje aj v globálnych štúdiách. V mnohých krajinách sa technológie na detekciu toxicity stávajú súčasťou nástrojov na zvýšenie bezpečnosti používateľov internetu, napríklad iniciatíva Perspective API od Google, ako aj odporúčania Európskej únie na automatické odstraňovanie nenávisťného obsahu. Zároveň však viacero prieskumov ukazuje, že modely typu transformery sú schopné zachytiť jazykové nuansy, ako je irónia, sarkazmus alebo nepriame urážky, s ktorými majú tradičné prístupy často problémy (6).

Zvýšená úroveň toxického správania na internete má priamy vplyv na psychické zdravie, verejnú diskusiu a kvalitu demokratického dialógu. Preto je vývoj spoľahlivých a účinných nástrojov na odhaľovanie a preventívne opatrenia proti toxicite nielen technickou, ale aj spoločenskou úlohou (2).

V porovnaní s tradičnými metódami strojového učenia (ako SVM, Naïve Bayes) a staršími neurónovými sieťami (LSTM, GRU) transformery sú schopné lepšie zachytiť

zložité jazykové vzťahy, kontext a sarkazmus, čo z nich robí vysoko efektívny nástroj na analýzu toxicity v sociálnych sieťach. Napriek vysokej presnosti však vyžadujú značné výpočtové zdroje a ich interpretácia je zložitá, čo otvára možnosti pre ďalší výskum a vylepšenia.

Cieľom práce je v prvej fáze oboznámiť sa s aktuálnymi poznatkami v oblasti spracovania prirodzeného jazyka (NLP) so zameraním na transformery a ich využitie pri detekcii toxicity. Následne sa práca venuje praktickej časti, v rámci ktorej boli analyzované a testované vybrané modely na reálnych textových vstupoch. Výsledkom riešenia je vytvorenie webovej aplikácie, ktorá umožňuje používateľom overiť mieru toxicity vo vlastnom texte pomocou viacerých predtrénovaných modelov.

1 Formulácia úlohy

V rámci riešenia bakalárskej práce sa majú postupne splniť nasledovné úlohy:

- Vypracovať prehľad problematiky detekcie toxicity v online prostredí, so zameraním na jej spoločenský a technický kontext.
- Študovať a analyzovať vývoj v oblasti spracovania prirodzeného jazyka (NLP), so zameraním na architektúru transformerov a ich uplatnenie v úlohách klasifikácie textu.
- Zvoliť vhodné transformerové modely a zdokumentovať princípy ich fungovania vrátane použitia mechanizmu sebavnímania (self-attention).
- Preskúmať, aké typy textových dát sa používajú v najznámejších datasetoch určených na detekciu toxicity, a vybrať z nich vhodné príklady na testovanie modelov.
- Otestovať väčšie množstvo dostupných NLP modelov na vybraných dátach a na základe porovnania výkonnosti zvoliť tie najvhodnejšie na ďalšie použitie.
- Implementovať webovú aplikáciu, ktorá umožní používateľovi v reálnom čase overiť mieru toxicity vlastného textu pomocou modelov, ktoré dosiahli najlepšie výsledky.
- Vypracovať dokumentáciu v súlade s pokynmi vedúceho práce a pravidlami pre záverečné práce.

2 Teoretický základ riešenej problematiky

2.1 Online toxicita ako spoločenský problém

Digitálne technológie radikálne zmenili spôsob komunikácie a výmeny názorov medzi ľuďmi. Okrem výhod však sociálne siete so sebou prinášajú aj riziká, medzi ktoré patrí šírenie toxického správania.

Za toxické správanie na internete možno považovať širokú škálu negatívnych prejavov – od urážok a vysmievania sa až po šírenie nenávisti a dezinformácií. Takéto správanie je často anonymné a cielené, a jeho prejavy sa môžu rozlišovať v závislosti od platformy, jazyka alebo kultúrneho kontextu. Štatistiky ukazujú, že obzvlášť zraniteľnými skupinami sú ženy, menšiny, netradičné spoločenstvá, imigranti a príslušníci rôznych náboženstiev (9). Tieto skupiny sú vystavené vyššiemu riziku online prenasledovania a útokov, čo môže viesť k psychickým traumám alebo ešte vážnejším následkom.

Jedným z faktorov, ktorý vo veľkej miere prispieva k šíreniu toxicity, je práve anonymita používateľov. Často vedie k zníženiu prekážok a strate zodpovednosti za vlastné správanie. Algoritmy sociálnych sietí formujú stream publikácií nielen podľa záujmov konkrétného používateľa, ale propagujú aj takzvané «odporúčania», ktoré získavajú najväčšiu pozornosť. V dôsledku toho sa zintenzívňujú emocionálne a konfliktné materiály, čo môže tlačiť na zraniteľné skupiny a zosilňovať atmosféru nenávisti v online priestore.

Online toxicita nie je len technický problém, ale má aj vážne sociálne dôsledky. Znižuje kvalitu verejnej diskusie, narúša dôveru medzi skupinami a môže šíriť nepravdivé informácie širokej verejnosti. V krajných prípadoch môže dokonca prispieť k fyzickému násiliu v reálnom živote. Príklady z rôznych krajín, ako sú online výzvy k násiliu voči menšinám alebo koordinované dezinformačné kampane, ukazujú, že toxické správanie na internete nemožno posudzovať izolovane od širších sociálno-politických procesov.

Z toho vyplýva, že online toxicita predstavuje komplexný spoločenský problém,

ktorý si vyžaduje nielen technologické riešenia, ale aj etické, právne a edukačné prístupy. Prevencia a moderovanie toxického obsahu sú kľúčovými výzvami súčasnej digitálnej spoločnosti.

2.1.1 Definícia a formy toxického správania

Toxické správanie na internete sa prejavuje v jazykových aj vizuálnych formách, ktoré úmyselne narúšajú sociálnu komunikáciu, vyvolávajú negatívne emócie, strašia alebo znevažujú jednotlivých ľudí alebo skupiny. Takéto formy správania sa často vyskytujú v podmienkach zníženej sociálnej zodpovednosti a nedostatočnej regulácie online priestoru.

V vedeckej literatúre sa opisujú rôzne formy a druhy toxicity v závislosti od obsahu a účelu vyjadrení. Medzi najčastejšie uvádzané kategórie patria:

- **Nenávistné prejavy (hate speech)** – vyjadrenia namierené proti ľuďom na základe rasy, rodu, sexuálnej orientácie, náboženstva alebo inej sociálnej alebo kultúrnej príslušnosti.
- **Kyberšikana** – opakované cielené útoky vo forme posmeškov, ohovárania alebo hrozieb.
- **Vulgárna a neslušná lexika** – výrazy s otvorene sexuálnym alebo agresívnym obsahom.
- **Osobné urážky a útoky** – priame útoky na vzhľad, schopnosti alebo súkromný život človeka.
- **Dezinformácia s agresívnym podtextom** – úmyselné šírenie nepravdivého alebo manipulatívneho obsahu s cieľom spôsobiť škodu.
- **Provokatívne alebo pasívne agresívne výroky** – sarkazmus, irónia a skryté frázy, ktoré formálne neporušujú pravidlá, ale sú vnímané ako urážlivé.

Rôzne výskumy zdôrazňujú, že nie všetka toxicita sa prejavuje v jasnej forme – automatické systémy majú často ťažkosti s rozpoznávaním skrytých prejavov, ako je ironická agresia alebo jazykové maskovanie. Napríklad v práci Rajeshwari a Manish (2024) sa skúmajú špecifiká zdržanlivého a ironického jazyka, ktorý predstavuje hrozbu aj napriek formálnej zdvorilosti (2). Podobne Wang a Chen (2023) poukazujú na to, že aj moderné transformerové modely, vrátane BERT a RoBERTa, vykazujú zníženú presnosť pri analýze pasívnej agresie a sarkastických výrokov (6). Podobné prípady predstavujú výzvu pre automatickú moderáciu, pretože vyžadujú nielen sémantickú interpretáciu, ale aj zohľadnenie kontextu a pragmatiky.

Formálne kritériá na určenie toxického obsahu sa môžu líšiť v závislosti od platformy, krajiny alebo kultúrnych noriem. V podstate však toxicita vždy predstavuje porušenie zásad dôstojnosti, rešpektu a bezpečnej komunikácie v digitálnom priestore.

2.1.2 Vplyv anonymity a algoritmov sociálnych sietí

Jedným z hlavných faktorov, ktoré prispievajú k šíreniu toxického správania na internete, je anonymita. V online prostredí sa ľudia cítia voľnejšie, keď nie je viditeľná ich identita, čo znižuje ich pocit zodpovednosti za svoje vyhlásenia, najmä ak im nehrozia žiadne sankcie ani sociálne dôsledky. Tento pocit beztretnosti často vedie k agresívnym, urážlivým alebo provokatívnym vyjadreniam, ktoré by boli v bežnej osobnej komunikácii neprijateľné. V psychológii sa tento jav nazýva «efekt deindividualizácie» – stieranie osobných hraníc a oslabovanie sociálnych noriem v anonymnom digitálnom prostredí. V kombinácii s algoritmami, ktoré určujú, aký obsah sa zobrazuje používateľom, vzniká prostredie, v ktorom je toxicita nielen tolerovaná, ale často aj nepriamo podporovaná samotným systémom (2).

Tieto algoritmy sú navrhnuté tak, aby maximalizovali zapojenie používateľa, to znamená, že v prvom rade zobrazujú príspevky, ktoré vyvolávajú silné emócie, ako je šok, hnev alebo pobúrenie. Takýto obsah zvyčajne generuje vyššiu úroveň interakcie, a preto je častejšie odporúčaný iným používateľom (2).

Problém spočíva v tom, že práve toxické alebo polarizujúce príspevky často vyvolávajú najintenzívnejšie reakcie a dostávajú sa tak do algoritmických odporúčaní alebo trendov. V dôsledku toho sú používatelia opakovane vystavení nepriateľskému alebo konfliktnému obsahu, aj keď on sami neprejavili záujem. Okrem toho môžu algoritmy vytvárať takzvané «informačné bubliny», v ktorých sa používateľ stretáva výlučne s obsahom, ktorý potvrdzuje jeho pozíciu, čo vedie k ďalšej polarizácii a zvýšeniu napätia v online priestore.

Treba tiež poznačiť, že aj moderné automatizované systémy majú obmedzenú schopnosť odhaľovať toxický obsah. Zvlášť zložité sú skryté formy toxicity – sarkazmus, irónia, pasívna agresia –, ktoré často unikajú odhaleniu. Podľa Wang a Chena (2023) modely ako BERT alebo RoBERTa vykazujú nižšiu presnosť pri analýze takýchto jemných prejavov (6). Tieto prípady predstavujú výzvu pre moderovanie, pretože vyžadujú hlbšie kontextuálne a pragmatické interpretácie.

Na prekonanie tohto problému je potrebný nielen technologický pokrok, ale aj vytvorenie pevného etického a právneho rámca, ktorý zabezpečí spravodlivé a bezpečné prostredie pre digitálnu komunikáciu.

2.1.3 Potreba automatizovanej detekcie

Rastúci objem obsahu a rýchlosť jeho objavovania sa na internete výrazne prevyšujú možnosti manuálnej moderácie. Manuálna kontrola komentárov, príspevkov alebo správ vyžaduje čas, finančné zdroje a kvalifikovaný personál, čo nie je dlhodobým a efektívnym riešením.

Okrem toho je ručná moderácia často nekonzistentná, subjektívna a náchylná na chyby či osobné predsudky. Automatická detekcia toxického obsahu založená na metódach strojového učenia a spracovania prirodzeného jazyka (NLP) ponúka možnosť škálovateľného a konzistentného prístupu k filtrovaniu obsahu.

Pomocou pokrokových modelov, najmä transformerových architektúr, ako sú BERT alebo RoBERTa, je možné analyzovať jazykové vzory, sémantický význam a kontext výrokov, čo zvyšuje presnosť odhaľovania toxického správania.

Okrem praktických výhod má automatická detekcia aj preventívnu úlohu. Používatelia si viac premýšľajú nad formuláciami, ak vedia, že ich obsah bude hodnotený automatizovaným systémom. Automatizácia tak môže prispieť k celkovému zlepšeniu online prostredia a zníženiu výskytu extrémnych alebo škodlivých vyjadrení.

Zavádzanie takýchto systémov by však malo byť sprevádzané etickým rámcom, aby sa zabránilo cenzúre, diskriminácii alebo nesprávnej interpretácii kultúrne špecifických vyjadrení. Preto je potrebné kombinovať technológiu s priehľadnosťou, možnosťou odvolania a náročným testovaním na rôznych jazykových a sociálnych dátach.

2.2 NLP prístupy k detekcii toxicity

Automatická detekcia toxického obsahu v digitálnom prostredí je typickou úlohou spracovania prirodzeného jazyka (NLP), vyžadujúcou analýzu významu, kontextu a štruktúry textu. Vzhľadom na rastúci objem komunikácie v sociálnych sieťach už manuálna moderácia nie je udržateľná, čo vedie k potrebe vytvorenia inteligentných systémov schopných automaticky odhaľovať toxické prejavy. Historický vývoj metód NLP odhalil niekoľko generácií prístupov – od systémov založených na pravidlách až po pokrokové modely hlbokého učenia.

Z hľadiska efektívnosti a možností použitia v oblasti toxicity sa v praxi najčastejšie používajú tri skupiny prístupov:

- klasické modely strojového učenia,
- neurónové siete,
- transformery a architektúry so self-attention mechanizmom.

Každá z týchto skupín predstavuje inú úroveň zložitosti, schopnosti pracovať s kontextom a presnosti predikcie. Pri porovnávaní prístupov je dôležité brať do úvahy aj jazykovú a kultúrnu variabilitu, špecifiká dátových súborov a požiadavky na interpretáciu výsledkov. V nasledujúcich častiach sa podrobne pozrieme na tieto tri kategórie modelov a ich použitie pri detekcii toxicity.

2.2.1 Klasické modely (Logistic Regression, SVM, Naive Bayes)

Klasické modely strojového učenia predstavujú prvú generáciu algoritmov využívaných na automatickú detekciu toxického obsahu v online diskusiách. Tieto modely vychádzajú z číselných reprezentácií textu, ako sú frekvencie slov (Bag-of-Words), n-gramy alebo vážené vektory **TF-IDF (term frequency–inverse document frequency)**, a aplikujú na ne štatistické klasifikačné metódy. Ich hlavnými výhodami sú výpočtová nenáročnosť, rýchla implementácia a transparentnosť rozhodovacích procesov.

Medzi najčastejšie využívané klasické algoritmy v tejto oblasti patria:

- **Logistická regresia** – model, ktorý odhaduje pravdepodobnosť, že daný text patrí do kategórie toxický. Je vhodná pre binárnu aj viacnásobnú klasifikáciu a poskytuje dobre interpretovateľné výsledky. V praxi sa často kombinuje s TF-IDF reprezentáciou vstupu a slúži ako základná porovnávacia metóda v mnohých štúdiách (6; 5).
- **SVM (Support Vector Machine)** – klasifikátor, ktorý hľadá optimálnu hranicu medzi kategóriami a snaží sa ich čo najpresnejšie oddeliť. V úlohách textovej klasifikácie sa osvedčil najmä preto, že dobre funguje aj pri veľkom počte znakov (napríklad tisíce slov v komentároch). Je známy svojou spoľahlivosťou a často sa používa ako referenčný model pri porovnávaní s novšími neurónovými sieťami (9).
- **Naivný Bayesov klasifikátor** – jednoduchý pravdepodobnostný model založený na Bayesovej vete, ktorý predpokladá nezávislosť medzi vlastnosťami. Napriek tejto zjednodušenej premise dosahuje dobré výsledky, najmä pri priamo formulovaných prejavoch toxicity, ako sú vulgárnosti, nadávky alebo priame útoky (8). Jeho hlavnými výhodami sú extrémna rýchlosť a nízke nároky na tréningové dáta.

Tieto modely boli široko používané v počiatočných pokusoch o automatickú mo-

deráciu, ako napríklad v raných verziách detektorov pre fórum Reddit alebo YouTube komentáre (14). Hoci moderné úlohy klasifikácie toxicity si čoraz častejšie vyžadujú hlbšie modely schopné chápať kontext a jemné jazykové nuansy, klasické algoritmy stále plnia dôležitú úlohu ako baseline metódy pri hodnotení výkonu pokročilých prístupov. Vďaka svojej jednoduchosti sú zároveň vhodné na nasadenie v prostrediach s obmedzenými výpočtovými zdrojmi.

2.2.2 Neuronové siete (CNN, RNN, LSTM)

S rastúcim objemom a zložitostou textového obsahu na internete sa klasické modely ukázali ako nedostatočné na zachytávanie kontextu, významových súvislostí a jazykových nuáns. Významný pokrok priniesli hlboké neurónové siete, ktoré umožňujú analyzovať text ako sekvenčné dáta. Medzi najčastejšie využívané architektúry v oblasti detekcie toxicity patria konvolučné neurónové siete (CNN), rekurentné neurónové siete (RNN) a modely s dlhodobou krátkodobou pamäťou (LSTM).

Konvolučné neurónové siete (CNN) – pôvodne určené na spracovanie vizuálnych dát – sa v oblasti spracovania textu osvedčili najmä pri detekcii lokálnych jazykových vzorov. CNN dokážu efektívne extrahovať významové n-gramy, ktoré môžu signalizovať prítomnosť toxického obsahu, napríklad urážlivé frázy alebo expresívne slovné spojenia. Vďaka nízkej výpočtovej náročnosti sú vhodné na rýchle spracovanie veľkého množstva dát.

Rekurentné neurónové siete (RNN) – spracúvajú text ako postupnosť slov a uchovávajú informácie o predchádzajúcich prvkoch. Tým dokážu reflektovať vývoj významu v rámci viet alebo komentárov. Ich kľúčovou výhodou je schopnosť zachytávať jazykový kontext, no trpia problémom «miznúceho gradientu», čo obmedzuje ich výkon pri dlhších vstupoch.

Modely s dlhodobou krátkodobou pamäťou (LSTM) – vznikli ako rozšírenie RNN a sú navrhnuté tak, aby si pamätali relevantné informácie aj na dlhšiu vzdialenosť v texte. Ich pamäťové bunky umožňujú uchovať významové vzťahy medzi slovami a vetami, ktoré sú kľúčové pri identifikácii nepriamych foriem toxicity. Podľa

Wang a Chen (2023) dokážu LSTM siete presnejšie klasifikovať toxické prejavy, ktoré sa objavujú až v neskorších častiach komentára alebo sú jazykovo maskované (6).

Neuronové siete tvoria dôležitý prechod medzi tradičnými prístupmi a transformerovými architektúrami. V porovnaní s klasickými modelmi dosahujú lepšie výsledky pri analýze kontextu a jazykovej štruktúry, no ich výkonnosť závisí od dostupných dát a technických podmienok. CNN a LSTM sú preto v praxi často využívané ako kompromisné riešenia medzi presnosťou a efektivitou.

2.2.3 Transformery a self-attention architektúra

Transformery predstavujú prelomový bod vo vývoji modelov spracovania prirodzeného jazyka. Na rozdiel od klasických algoritmov a neurónových sietí založených na sekvenčnom spracovaní vstupu, transformerová architektúra využíva mechanizmus takzvaného self-attention (sebapozornosti), ktorý umožňuje modelu súčasne analyzovať všetky časti textu bez ohľadu na ich poradie. Tento prístup výrazne zlepšil schopnosť modelov zachytávať dlhodobé závislosti, jazykové nuansy, ako aj implicitné formy toxicity.

Základ transformerovej architektúry bol predstavený v článku *Attention Is All You Need* (Vaswani et al., 2017), kde autori zaviedli nový spôsob spracovania textu založený výlučne na pozornosti, bez použitia rekurentných alebo konvolučných vrstiev. Vďaka paralelnému spracovaniu vstupov a schopnosti priradovať váhy jednotlivým slovám podľa ich významu v kontexte sa transformery stali mimoriadne efektívnymi pri riešení rôznych NLP úloh vrátane detekcie toxicity.

Mechanizmus self-attention umožňuje modelu priamo určiť, ktoré časti textu sú relevantné pri klasifikácii konkrétneho prejavu ako toxického. Napríklad model dokáže identifikovať, že urážlivé vyjadrenie v poslednej vete nadväzuje na osobný útok v úvode komentára, aj keď medzi nimi existuje jazykový odstup. Takéto prepojenie významu by bolo pre sekvenčné modely (ako RNN či LSTM) veľmi ťažké zachytiť.

Okrem základnej architektúry sa v praxi uplatnilo viacero modifikácií transformerov, ako napríklad BERT (Bidirectional Encoder Representations from Transfor-

mers), RoBERTa (Robustly Optimized BERT Pretraining Approach) alebo GPT (Generative Pretrained Transformer), ktoré sa líšia spôsobom trénovania, rozsahom údajov a aplikovanými optimalizáciami. Tieto modely dosiahli výborné výsledky v benchmarkoch ako GLUE, SuperGLUE alebo Jigsaw Toxic Comment Classification Challenge (6).

Výhodou transformerov je schopnosť porozumieť významu slov na základe ich kontextu v celej vete alebo odstavci, čo je kľúčové pri detekcii nejednoznačných alebo maskovaných prejavov toxicity (ako irónia, sarkazmus, pasívna agresia). Nevýhodou však zostáva ich výpočtová náročnosť, potreba veľkých objemov trénovacích dát a zložitosť interpretácie rozhodovania modelu.

Transformery dnes tvoria jadro mnohých moderných systémov automatickej moderácie obsahu na sociálnych sieťach a sú považované za najefektívnejšiu technológiu v detekcii online toxicity.

2.3 Porovnanie najznámejších transformerov

Od predstavenia architektúry transformera v roku 2017 sa spracovanie prirodzeného jazyka výrazne zmenilo. Modely, ktoré na začiatku pôsobili ako výskumné experimenty, sa dnes bežne používajú v rôznych oblastiach – od prekladu textu, cez sumarizáciu, až po detekciu toxicity na internete.

Transformery sú výnimočné v tom, ako spracúvajú jazyk. Na rozdiel od starších modelov nečítajú text po jednom slove za sebou, ale dokážu naraz zohľadniť celý kontext a určiť, ktoré časti spolu súvisia. Vďaka tomu vedia lepšie rozpoznať zložitejšie formy toxického správania, ako je irónia, sarkazmus alebo skryté narážky.

Od pôvodného modelu BERT vzniklo viacero ďalších verzií, ktoré sa líšia tým, ako boli trénované, aké majú rozmery, alebo ako rýchlo pracujú. Niektoré sú silnejšie a presnejšie, iné zas ľahšie a rýchlejšie na použitie v praxi.

V tejto časti sa pozriem na štyri najznámejšie a najčastejšie používané transformerové modely na detekciu toxicity:

- BERT – základný obojsmerný model,
- RoBERTa – vylepšená verzia BERT-u s lepším trénovaním,
- DistilBERT – zjednodušená verzia vhodná na rýchle nasadenie,
- GPT – generatívny model s možnosťou pracovať s rôznymi vstupmi.

Každý z nich opíšem zvlášť a porovnáam ich výhody, nevýhody a vhodnosť na detekciu toxického obsahu na sociálnych sieťach.

2.3.1 BERT

Model BERT (Bidirectional Encoder Representations from Transformers) bol predstavený v roku 2018 výskumným tímom Google AI ako prvý transformer, ktorý spracúva text obojsmerne – teda súčasne zľava doprava aj sprava doľava. Práve táto schopnosť pozeráť sa na slovo v širšom kontexte mu umožňuje lepšie rozumieť významu jednotlivých viet a súvislostiam medzi slovami.

BERT bol predtrénovaný na obrovskom množstve textov z Wikipédie a kníh, kde sa učil dopĺňať chýbajúce slová (*masked language modeling*) a porovnávať, či dve vety spolu logicky súvisia (*next sentence prediction*). Vďaka tomu sa naučil rozlišovať rôzne jazykové vzory, gramatické konštrukcie a sémantické nuansy.

V oblasti detekcie toxicity sa BERT osvedčil ako veľmi dobrý základ, najmä pri klasifikácii výrokov, ktoré obsahujú otvorenú alebo jasne vyjadrenú agresiu. Mnohé open-source projekty, ako aj súťaže (napríklad Jigsaw Toxic Comment Classification Challenge), využívajú BERT ako etalón, ku ktorému sa porovnávajú ďalšie architektúry (6).

Výhodou BERT-u je jeho schopnosť presne pracovať s významom v kontexte celého výroku. Na druhej strane, jeho výpočtová náročnosť je vyššia a pri práci s dlhšími textami môže byť menej efektívny, keďže pôvodná architektúra BERT-u má obmedzenie na maximálne 512 tokenov.

Celkovo je BERT vnímaný ako silný a overený model, ktorý sa hodí najmä na úlohy, kde je dôležité porozumieť presnému významu textu v oboch smeroch. V kombinácii s ďalším trénovaním na špecifickej doméne (ako komentáre na sociálnych sieťach) dokáže dosahovať veľmi dobré výsledky pri detekcii toxického obsahu.

2.3.2 RoBERTa

Model RoBERTa (Robustly Optimized BERT Pretraining Approach) vznikol ako vylepšenie pôvodného BERT-u. Predstavili ho výskumníci z Facebook AI v roku 2019 s cieľom zlepšiť výkon modelu pomocou rozsiahlejšieho trénovania a niekoľkých architektonických úprav. RoBERTa zachováva základnú štruktúru BERT-u, no výrazne zvyšuje kvalitu vďaka zmenám v predtrénovaní.

Najväčší rozdiel oproti BERT-u spočíva v tom, že RoBERTa bol trénovaný na oveľa väčšom množstve dát a dlhší čas. Zároveň nepoužíva úlohu *next sentence prediction*, ktorá sa v prípade BERT-u ukázala ako menej užitočná. Navyše boli zmenené niektoré technické parametre, ako veľkosť dávok (batch size), rýchlosť učenia a typ maskovania slov. Tieto úpravy viedli k tomu, že RoBERTa vo viacerých benchmarkoch (vrátane úloh toxicity) prekonal pôvodný BERT (6).

V kontexte detekcie toxického obsahu je RoBERTa často považovaný za jeden z najpresnejších dostupných modelov. Vďaka svojej stabilite dokáže spoľahlivo rozpoznať nielen priame, ale aj jemnejšie formy agresie, irónie alebo skrytého zosmiešňovania. V mnohých prípadoch sa využíva ako finálna verzia klasifikačného systému, najmä ak sú dostupné dostatočné výpočtové zdroje.

Na druhej strane, výpočtová náročnosť RoBERTy je ešte o niečo vyššia ako v prípade BERT-u, čo môže byť problémom pri nasadzovaní modelu v reálnom čase alebo v mobilných prostrediach. Pre takéto prípady sú vhodnejšie odlahčené modely ako DistilBERT.

Celkovo je RoBERTa veľmi silný a presný model, ktorý si v oblasti NLP našiel široké uplatnenie – a v detekcii toxicity patrí medzi najspoľahlivejšie možnosti.

2.3.3 DistilBERT

Model DistilBERT vznikol ako ľahšia verzia BERT-u so zámerom ponúknuť rýchlejší a menej náročný model, ktorý si pritom zachová čo najviac z pôvodnej presnosti. Predstavili ho výskumníci z Hugging Face v roku 2020 a od začiatku bol navrhnutý tak, aby bol vhodný aj pre použitie na zariadeniach s obmedzeným výkonom.

DistilBERT využíva techniku nazývanú *knowledge distillation*, pri ktorej sa menší model učí napodobňovať správanie väčšieho – v tomto prípade pôvodného BERT-u. V praxi to znamená, že hoci má DistilBERT iba približne polovicu parametrov a je o 40% menší, dokáže dosiahnuť približne 95% jeho presnosti (6).

V oblasti detekcie toxicity sa DistilBERT ukázal ako veľmi užitočný, najmä tam, kde je dôležitá rýchlosť odpovede alebo obmedzené výpočtové možnosti – napríklad v mobilných aplikáciách, chatbotoch alebo pri analýze komentárov v reálnom čase. Dokáže rozpoznať väčšinu bežných foriem toxického správania, hoci v porovnaní s plnohodnotnými modelmi môže mať nižšiu citlivosť na jemné jazykové odtiene.

Jeho hlavnou výhodou je výborný kompromis medzi rýchlosťou, presnosťou a náročnosťou. Vďaka tomu sa stal veľmi populárnym aj medzi vývojármi, ktorí potrebujú spoľahlivý, ale zároveň efektívny model pre praktické použitie.

2.3.4 GPT a iné veľké jazykové modely

Modely typu GPT (Generative Pretrained Transformer) predstavujú osobitnú vetvu vývoja transformerov. Na rozdiel od BERT-u a jeho potomkov, ktoré sú primárne určené na klasifikáciu alebo pochopenie textu, GPT je model trénovaný na generovanie textu – teda na predikciu ďalšieho slova v sekvencii. Prvý model GPT bol predstavený tímom OpenAI v roku 2018 a odvtedy prešiel niekoľkými zásadnými vylepšeniami, ktoré vyvrcholili vo veľkých jazykových modeloch ako GPT-2, GPT-3 a novšie verzie.

GPT je jednosmerný model – číta text zľava doprava – čo pri klasifikácii môže znamenať slabšiu orientáciu na úplný kontext. Napriek tomu však jeho silné stránky

spočívajú v schopnosti pracovať s veľmi dlhými vstupmi a chápať kontext na vysokej úrovni. Vďaka veľkému množstvu tréningových dát a miliardám parametrov dokáže GPT reagovať na otázky, generovať zrozumiteľné odpovede a dokonca si «pamätať» kontext viacerých vyjadrení.

V oblasti detekcie toxicity sa GPT dá použiť dvoma spôsobmi. Po prvé ako klasifikátor s takzvaným *few-shot* alebo *zero-shot* učením, kde model dostane len niekoľko príkladov alebo len krátku inštrukciu na identifikáciu toxických prejavov. Po druhé ako generatívny nástroj, ktorý dokáže navrhnúť miernejšie formulácie alebo vytvoriť spätnú väzbu na nevhodné komentáre.

Výhodou GPT modelov je ich flexibilita – nie sú obmedzené na jednu úlohu a dajú sa ľahko adaptovať na rôzne typy vstupov. Zároveň však ide o extrémne výpočtovo náročné systémy, ktoré si vyžadujú špecializovaný hardvér a starostlivé riadenie pri použití, najmä pokiaľ ide o etiku a spoľahlivosť výstupov.

Okrem GPT existujú aj iné veľké jazykové modely, ako napríklad T5, XLNet, ERNIE či LLaMA, ktoré majú rôzne technické vylepšenia a sú zamerané na konkrétne aplikácie. V praxi sa však GPT vďaka svojej univerzálnosti a schopnosti generovať prirodzený jazyk stal najznámejším predstaviteľom tejto skupiny modelov.

2.4 Súvisiace práce

2.4.1 Civil Rephrases of Toxic Texts With Self-Supervised Transformers (Laugier et al., 2021) (7)

- **Obsah a cieľ** – Článok sa zameriava na preformulovanie toxických komentárov do civilnejšej podoby pomocou modelu CAE-T5, ktorý je založený na text-to-text transformerovej architektúre a trénovaný samo-supervidovaným spôsobom. Na rozdiel od klasickej detekcie toxicity, autori navrhujú riešenie, ktoré generuje menej urážlivé verzie vstupných výrokov bez nutnosti manuálne vytvoreného paralelného korpusu. Model bol trénovaný na dátach z datasetu Civil Comments a využíva denoising autoencoder a cyklickú konzistenciu na

zachovanie pôvodného významu výroku.

- **Aktuálnosť a súvislosť s témou práce: Čiastočne** – Práca priamo využíva transformerový model CAE-T5 na úlohu súvisiacu s toxicitou textov. Aj keď cieľom nie je klasifikácia, ale generovanie preformulovaných viet, článok demonštruje silu transformerov pri jemnej úprave štýlu komunikácie. Prístup bez potreby párových dát je obzvlášť zaujímavý pre jazyky a platformy, kde anotovaný materiál chýba.
- **Záver a výhody článku** – Článok ponúka dôkaz koncepcie pre využitie veľkých predtrénovaných transformerov na transformáciu toxických výrokov do prijateľnejšej formy. Ukazuje, že self-supervised prístup môže viesť k zmysluplným výstupom bez ručného prepisovania. Tento prístup môže byť relevantný nielen na moderovanie obsahu, ale aj ako doplnok ku klasifikačným modelom detekcie toxicity.

2.4.2 Social Media Toxicity Classification Using Deep Learning: Real-World Application UK Brexit (Fan et al., 2021)(5)

- **Obsah a cieľ** – Cieľom článku bolo vytvoriť a otestovať model detekcie toxicity na reálnych dátach zo sociálnych sietí, konkrétne z diskusií súvisiacich s Brexitom na Twitteri. Autori použili transformerový model BERT a jeho varianty RoBERTa, DistilBERT, Multilingual BERT, ktoré boli doladené na dátach z Kaggle Toxic Comment Classification Challenge. Modely následne testovali na dvoch samostatných datasetoch z Twitteru zozbieraných cez API v rokoch 2019–2020. Výsledky ukazujú, že BERT dosiahol najvyššiu presnosť a výkonnosť pri klasifikácii toxických príspevkov.
- **Aktuálnosť a súvislosť s témou práce: áno** – Práca sa priamo zameriava na detekciu toxicity pomocou transformerov, a to na praktickej a spoločensky dôležitej téme – online diskusiách o Brexite. Použitie BERT-u aj jeho porovnanie s inými variantmi poskytuje cenný prehľad o výkonnosti týchto modelov

v reálnych podmienkach. Dátová aj metodická časť sú dobre prenositeľné na iné jazyky a sociálne platformy.

- **Záver a výhody článku** – Článok demonštruje silu transformerových modelov pri analýze toxicity vo veľkom množstve používateľského obsahu. Práca je prakticky orientovaná a využíva reálne dáta z Twitteru, čo zvyšuje jej aplikčný potenciál. Získané výsledky podporujú použitie predtrénovaných modelov a ukazujú možnosti ich doladenia na konkrétne prípady.

2.4.3 GTH-UPM at DETOXIS-IberLEF 2021 (Romero et al., 2021)(14)

- **Obsah a cieľ** – Cieľom článku bolo zúčastniť sa výzvy DETOXIS-IberLEF 2021 v oblasti detekcie toxicity v komentároch na sociálnych sieťach. Autori testovali viacero hlbokých modelov vrátane rekurentných sietí (biLSTM) a transformerových architektúr. Najlepšie výsledky dosiahol špeciálne doladený model BETO – španielska verzia BERT-u – ktorý bol rozšírený o nové tokeny relevantné pre danú úlohu. Článok opisuje aj metódu rozšírenia slovnej zásoby predtrénovaných modelov pomocou doménovo špecifických tokenov.
- **Aktuálnosť a súvislosť s témou práce: áno** – Táto práca sa priamo venuje použitiu transformerov pri klasifikácii toxicity, pričom využíva reálne údaje a zapája sa do súťažného benchmarku. Doladenie modelu BETO a práca s rozšírenou slovnou zásobou ponúka relevantné poznatky pre adaptáciu transformerov na špecifické jazykové alebo tematické oblasti.
- **Záver a výhody článku** – Článok potvrdzuje, že transformerové modely, najmä tie tréňované v konkrétnom jazyku, ako napríklad BETO pre španielčinu, môžu dosahovať najvyššiu presnosť pri detekcii toxicity. Výhodou je aj prezentácia jednoduchej a efektívnej stratégie rozšírenia slovníka, ktorá zlepšuje výkonnosť modelu. Tento prístup môže byť užitočný pri práci s málo podporovanými jazykmi alebo špecifickými témami.

2.4.4 DeTox at GermEval 2021: Toxic Comment Classification – Mina Schütz et al., november 2021(1)

- **Obsah a cieľ** – Článok opisuje účasť tímu DeTox na súťaži GermEval 2021, ktorá sa zameriavala na klasifikáciu toxických komentárov v nemeckom jazyku. Autori použili kombináciu predtrénovaných jazykových modelov vrátane mBERT, XLM-R a ensemble metód s cieľom dosiahnuť čo najvyššiu presnosť pri detekcii toxicity. Okrem toho implementovali techniky ako oversampling a analýzu chýb, čo umožnilo zvýšiť spoľahlivosť ich systému.
- **Relevancia pre moju tému: áno** – Článok sa priamo zameriava na detekciu toxicity pomocou transformerových modelov a poskytuje praktickú aplikáciu týchto architektúr v nemeckom kontexte. Ukazuje tiež, ako kombinovať rôzne modely a prístupy pre zlepšenie výsledkov.
- **Záver a výhody článku** – Práca poskytuje jasný prehľad o tom, ako efektívne využiť transformerové modely na klasifikáciu toxicity a aké doplnkové techniky môžu zlepšiť výkon systému. Výhodou je aj metodologická otvorenosť (replikovateľnosť), čo umožňuje využiť ich postupy v ďalších výskumoch alebo aplikáciách detekcie toxicity.

2.4.5 Toxicity detection in online Georgian discussions (Lashkarashvili & Tsintsadze, 2022)(9)

- **Obsah a cieľ** – Článok sa venuje detekcii toxicity v online diskusiiach na gruzínskom fóre (Tbilisi Forum) a predstavuje prvý výskum tohto typu pre jazyk s minimálnou podporou NLP. Autori vytvorili anotovaný dataset s 10 000 komentármi (rozdelenými na toxické a netoxické) a otestovali klasické algoritmy (Naive Bayes, SVM), hlboké neurónové siete (CNN, biLSTM, biGRU) a transformerový model. Najlepšie výsledky dosiahla CNN (presnosť 88,8 %, AUC 94,2 %).

- **Aktuálnosť a súvislosť s témou práce:** Nie – Hoci transformer nebol hlavným predmetom výskumu, bol súčasťou porovnávaných modelov, čo z článku robí cenný doplnkový zdroj pre analýzu metód klasifikácie toxicity.
- **Záver a výhody článku** – Hlavná hodnota článku spočíva v demonštrácii použitia rôznych modelov pre málo podporovaný jazyk a v jasnom popise procesu budovania klasifikačného systému. Práca ukazuje, že aj relatívne jednoduché modely ako CNN môžu byť vysoko efektívne pri kvalitnej anotácii a presne stanovenej úlohe.

2.4.6 Hate Speech and Toxic Comment Detection using Transformers – Pierre Guillaume et al., apríl 2022(8)

- **Obsah a cieľ** – Článok prezentuje porovnanie viacerých transformerových modelov (HateBERT, RoBERTa, BERTweet) pri detekcii nenávistných a toxických komentárov. Autori testovali modely na datasete Jibes&Delights (komentáre zo subredditu RoastMe a FreeCompliments) a študovali vplyv rôznych techník augmentácie dát ako spätný preklad, vkladanie a nahradzovanie tokenov. Najlepšie výsledky dosiahol model RoBERTa st4-aug, ktorý využíva výstupy zo štyroch posledných vrstiev transformera.
- **Relevancia pre moju tému: áno** – Práca sa zameriava priamo na použitie transformerov pre úlohu binárnej klasifikácie toxických výrokov a porovnáva ich účinnosť pri rôznych konfiguráciách trénovania. Dôležité sú aj praktické poznatky o tom, ako zvyšovať výkon modelov bez nutnosti zvyšovania ich zložitosti.
- **Záver a výhody článku** – Autori ukazujú, že nie zložitosť modelu, ale výber výstupných vrstiev a techniky augmentácie môžu mať zásadný vplyv na presnosť. RoBERTa st4-aug dosiahla najlepšie výsledky bez potreby rozsiahlejších modelov. Článok poskytuje praktický návod na efektívne doladenie transformerových modelov pre detekciu toxicity v sociálnych sieťach.

2.4.7 A Comparative Study of Attention-Based Transformer Networks and Traditional ML Methods – Sihao Wang, Bingjie Chen, september 2023 (6)

- **Obsah a cieľ** – Článok sa zameriava na porovnanie výkonnosti transformero- vých sietí (ako BERT, GPT) a tradičných metód strojového učenia (Logistic Regression, SVM, Random Forest) pri klasifikácii toxických komentárov. Autori využili benchmarkový dataset Toxic Comment Classification Dataset a analyzovali presnosť, recall, F1-score a AUC-ROC. Výsledky ukázali, že transformerové modely (najmä BERT) prekonávajú klasické prístupy takmer vo všetkých metrikách, pričom najvyššia presnosť dosiahla 92,14 %.
- **Relevancia pre moju tému: áno** – Práca sa priamo venuje porovnaniu transformerov a klasických algoritmov pri detekcii toxicity, pričom využíva dôsledné experimentálne nastavenie a detailnú analýzu. Obsahuje aj úvahy o interpretovateľnosti, výpočtovej náročnosti a vhodnosti pre nasadenie v praxi.
- **Záver a výhody článku** – Článok demonštruje jasné výhody transformero- vých modelov pri analýze toxických komentárov vrátane schopnosti pracovať s kontextom a dlhodobými závislosťami. Výhodou je aj systematické porovnanie s tradičnými metódami a hodnotenie ich praktickej využiteľnosti pri detekcii toxicity.

2.4.8 Comparison between Machine Learning and Deep Learning Approaches – Andrea Bonetti et al., máj 2023(13)

- **Obsah a cieľ** – Cieľom článku bolo porovnať tradičné metódy strojového učenia (Logistic Regression, Random Forest, SVM) s hlbokými modelmi, reprezentovanými transformerom BERTweet, v úlohe detekcie toxických komentárov na sociálnych sieťach. Klasické modely boli kombinované s technikami modelovania tém (LSI, LDA). Hoci BERTweet mierne prekonal ostatné prí-

stupy v presnosti (ako F1 91,4%), rozdiely neboli výrazné. Autori zdôraznili aj otázku výpočtovej náročnosti a interpretovateľnosti modelov.

- **Relevancia pre moju tému: áno** – Článok sa priamo venuje porovnaniu transformerov s tradičnými prístupmi pri detekcii toxicity, pričom ponúka kvantitatívne výsledky a analýzu ich výhod a obmedzení. Relevantné je aj použitie rôznych NLP techník a ich kombinácií v reálnom dátovom prostredí (Twitter).
- **Záver a výhody článku** – Práca ukazuje, že hoci transformerové modely poskytujú najvyššiu presnosť, tradičné modely môžu byť porovnateľne účinné v prostrediach s obmedzenými výpočtovými zdrojmi. Výhodou článku je detailné metodologické porovnanie vrátane časovej náročnosti, výstupných metrík a diskusie o možnej optimalizácii oboch prístupov.

2.4.9 Comparing Different Transformer Models' Performance for Identifying Toxic Language – Carl Sundelin, jar 2023(3)

- **Obsah a cieľ** – Diplomová práca sa zameriava na porovnanie troch transformerových modelov (RoBERTa, ALBERT a DistilBERT) pri detekcii toxického jazyka. Modely boli trénované a testované na kombinovanom datasete s anotáciami ako «abusive», «hateful», «harassing» a iné formy toxicity. Okrem klasických metrík bola hodnotená aj výkonnosť na reálnych dátach z Redditu. Cieľom bolo zistiť, ktorý z modelov dosahuje najlepšiu rovnováhu medzi presnosťou a výpočtovou efektivitou.
- **Relevancia pre moju tému: áno** – Práca priamo testuje transformerové architektúry v kontexte detekcie toxicity a ponúka praktické porovnanie medzi plnohodnotnými a odľahčenými modelmi. Obsahuje tiež úvahy o vhodnosti nasadenia v reálnych podmienkach, vrátane hodnotenia na dátach zo sociálnych sietí.

- **Záver a výhody článku** – Výsledky ukazujú, že aj keď RoBERTa dosahuje najvyššiu presnosť, model DistilBERT predstavuje najlepšiu rovnováhu medzi rýchlosťou trénovania a kvalitou výstupu. Výhodou práce je dôkladná experimentálna metodológia a dôraz na reálne scenáre použitia pri klasifikácii toxického obsahu.

2.4.10 Lightweight Toxicity Detection in Spoken Language – Ahlam Husni Abu Nada et al., apríl 2023(10)

- **Obsah a cieľ** – Článok sa zameriava na detekciu toxicity v hovorenom jazyku pomocou ľahkého transformerového modelu založeného na wav2vec2.0. Cieľom bolo vytvoriť riešenie použiteľné v reálnych fyzických prostrediach, napríklad na periférnych zariadeniach (edge devices). Autori aplikovali techniky ako kvantizácia a knowledge distillation, čím sa podarilo znížiť veľkosť modelu viac než 14-násobne pri zachovaní vysokej presnosti (F1-skóre až 90,3 %).
- **Relevancia pre moju tému: Áno** – Práca predstavuje unikátny prístup k detekcii toxicity mimo klasického textového prostredia, pričom demonštruje možnosti využitia transformerov aj pri zvukových dátach. Téma je rozšírením tradičných NLP prístupov a zároveň veľmi aktuálna v súvislosti s nasadzovaním AI v reálnych zariadeniach.
- **Záver a výhody článku** – Autori ukazujú, že modely založené na transformerovej architektúre môžu byť efektívne optimalizované pre nízkonákladové prostredia bez výraznej straty výkonu. Výhody práce spočívajú v praktickej použiteľnosti, dôraze na nasadenie v reálnych podmienkach a inovatívnom rozšírení oblasti detekcie toxicity do hovoreného jazyka.

2.4.11 COGNITIVE METHOD TO DETECT TOXIC COMMENTS IN SOCIAL MEDIA – Jomy Joseph et al., marec 2024(15)

- **Obsah a cieľ** – Článok predstavuje model na automatickú detekciu toxických komentárov na sociálnych sieťach pomocou hybridnej neurónovej siete (Hybrid Neural Network – HNN), ktorá kombinuje CNN a RNN. Cieľom bolo vytvoriť systém, ktorý dokáže efektívne klasifikovať komentáre ako toxické alebo netoxické a zároveň identifikovať ich mieru toxicity. Model je navrhnutý tak, aby bol schopný nasadenia v rámci webových aplikácií a podporoval aj viacjazyčné vstupy s prevodom do angličtiny.
- **Relevancia pre moju tému: Nie** – Práca ponúka praktický príklad vývoja systému detekcie toxicity vrátane návrhu architektúry, zberu a predspracovania dát, trénovania modelu a jeho integrácie do webového rozhrania. Hoci nepracuje priamo s transformerovými modelmi, kombinuje hlboké architektúry a reflektuje problémy a výzvy typické pre túto doménu.
- **Záver a výhody článku** – Článok ukazuje, že aj kombinované modely CNN a RNN môžu byť efektívne pri detekcii toxicity a poskytuje úplný prehľad o celom vývojovom cykle riešenia. Výhodou je dôraz na efektívnosť, reálne nasadenie a schopnosť pracovať s rôznymi jazykmi, čo je dôležité pre škálovateľnosť detekčných systémov v praxi.

2.4.12 TINY-TOXIC-DETECTOR: A Compact Transformer-Based Model – Michiel Kamphuis, august 2024(16)

- **Obsah a cieľ** – Článok predstavuje model Tiny-toxic-detector, kompaktný transformer navrhnutý špeciálne na detekciu toxického obsahu s dôrazom na nízke nároky na výpočtové zdroje. Model má len 2,1 milióna parametrov, ale dosahuje porovnateľné výsledky s omnoho väčšími architektúrami: 90,97 % presnosť na datasete ToxiGen a 86,98 % na Jigsaw. Je postavený na 4 vrstvách

encoderov, s 2 attention headmi a bez predtrénovania – trénovaný len na označených dátach.

- **Relevancia pre moju tému: áno** – Práca sa výslovne zameriava na využitie transformerov pri detekcii toxicity a ponúka zaujímavý prístup k optimalizácii modelu pre nasadenie v prostredí s obmedzenými zdrojmi. Zahŕňa aj porovnanie s inými modelmi a analýzu výkonnosti z hľadiska presnosti, rýchlosti a pamäťovej náročnosti. Tento model som využil aj v praktickej časti mojej bakalárskej práce ako jeden z porovnávaných systémov.
- **Záver a výhody článku** – Model Tiny-toxic-detector ukazuje, že aj veľmi malé modely môžu byť konkurencieschopné v detekcii toxicity. Výhody práce spočívajú v prehľadnej dokumentácii architektúry, dôraze na ekologickú udržateľnosť a praktickú použiteľnosť v reálnych podmienkach ako edge zariadenia alebo mobilné aplikácie.

2.4.13 The Impact of Transformer Models on Detecting Hate Speech and Toxicity – Jiawei Li, Yuqing Xie, 2024(17)

- **Obsah a cieľ** – Článok skúma vplyv transformerových modelov na detekciu nenávisťného a toxického jazyka v sociálnych médiách. Autori porovnali viacero architektúr vrátane BERT, RoBERTa a DistilBERT na viacerých verejných datasetoch, ako aj na dátach s rôznou mierou toxicity. Cieľom bolo vyhodnotiť nielen presnosť modelov, ale aj ich stabilitu voči variabilite jazyka, sarkazmu a implicitným formám nenávisti.
- **Relevancia pre moju tému: áno** – Práca priamo analyzuje výkonnosť transformerových modelov pri úlohe detekcie toxicity, pričom kladie dôraz na praktickú použiteľnosť, výpočtovú efektívnosť a prácu s komplexnými jazykovými štruktúrami. Prináša aj pohľad na to, ako rôzne formy jazyka ovplyvňujú výstupy modelov.

- **Záver a výhody článku** – Článok potvrdzuje, že transformerové modely sú mimoriadne efektívne pri klasifikácii toxického obsahu, a zároveň ukazuje, že stabilita modelu závisí nielen od architektúry, ale aj od typu trénovacích dát. Výhody práce spočívajú v širokom experimente, dôraze na kvalitu hodnotenia a záveroch, ktoré možno uplatniť pri výbere modelu pre konkrétne nasadenie.

2.4.14 A Systematic Review of Toxicity in Large Language Models – Guillermo Villate-Castillo et al., júl 2024(4)

- **Obsah a cieľ** – Článok predstavuje systematický prehľad literatúry týkajúcej sa toxicity v rámci veľkých jazykových modelov (LLM), pričom sa zameriava na definície toxicity, dostupné datasety, metodiky detekcie a techniky zmierňovania. Autori analyzovali 206 výskumov a identifikovali rad problémov, ako napríklad chýbajúce jednotné definície toxicity, problémy s generalizáciou modelov, absenciu štandardizovaných benchmarkov a dôležitosť transparentnosti v anotácii dát.
- **Relevancia pre moju tému: áno** – Práca má vysokú výpovednú hodnotu pre výskum detekcie toxicity, najmä v kontexte transformerových architektúr. Sústreďí sa na aktuálne výzvy v oblasti detekcie a zmierňovania toxicity v generatívnych modeloch a sumarizuje stav poznania do roku 2023, čo poskytuje široké teoretické pozadie k mojej praktickej analýze modelov.
- **Záver a výhody článku** – Článok pomáha lepšie pochopiť komplexnosť problematiky toxicity, upozorňuje na nedostatky súčasných metodík a potrebu štandardizácie v oblasti anotácie, hodnotenia a riešení zaujatosti modelov. Výhodou je systematické spracovanie, rozsiahly zoznam referencií a formulácia otvorených výskumných otázok.

Zhrnutie súvisiacich prác

Naštudované vedecké články a výskumné správy zohrali kľúčovú úlohu pri budovaní teoretického základu a formovaní praktickej časti mojej práce. Vďaka ich analýze som získal prehľad o aktuálnych trendoch vo využívaní transformerových modelov na detekciu toxicity, ako aj o výhodách a obmedzeniach rôznych prístupov. Získané poznatky mi umožnili:

- pochopiť rôzne stratégie doladenia transformerov (ako BERT, RoBERTa, DistilBERT) a ich efektivitu v porovnaní s klasickými modelmi.
- identifikovať bežne používané datasety (ako Jigsaw, ToxiGen) a metriky na hodnotenie presnosti detekcie toxicity.
- oboznámiť sa s praktickými problémami, ako sú irónia, pasívna agresia alebo jazyková nevyváženosť, a ich vplyvom na výkonnosť modelov.
- inšpirovať sa konkrétnymi implementáciami a rozhodnúť sa pre vhodné modely a technológie použité v praktickej časti.
- rozšíriť pohľad na etické a metodologické otázky, ktoré súvisia s automatizovanou moderáciou online obsahu.

2.5 Dátové sady používané na detekciu toxicity

Úspešná detekcia toxicity v online komunikácii si vyžaduje prístup ku kvalitným a reprezentatívnym anotovaným dátam. Práve od štruktúry, rôznorodosti a spoľahlivosti použitých datasetov závisí, do akej miery je model schopný rozpoznať agresívne, urážlivé alebo manipulatívne výroky v rôznych kontextoch. Pri tom je dôležité brať do úvahy nielen jazykové zvláštnosti, ale aj kultúrne, platformové a žánrové rozdiely jednotlivých zdrojov.

V posledných rokoch vzniklo viacero datasetov určených na tréning a testovanie modelov pre klasifikáciu toxicity. Niektoré sa zameriavajú na anglicky písané

komentáre v školských, novinových alebo sociálnych prostrediach, iné sú orientované na konkrétne jazyky, ako napríklad nemčina, španielčina či gruzínčina. Mnohé z nich obsahujú nielen javnú, ale aj skrytú formu toxicity – vrátane sarkazmu, pasívnej agresie alebo kultúrne špecifických výrazov.

V tejto práci som sa rozhodol sústrediť na tie datasetové zdroje, ktoré sú najčastejšie používané v odbornej literatúre a praktických aplikáciách, a zároveň poskytujú kvalitnú anotáciu a rozmanité formy toxického obsahu. Tieto datasety umožňujú nielen efektívne trénovanie modelov, ale aj objektívne porovnávanie ich výkonnosti v rôznych podmienkach.

2.5.1 Jigsaw Toxic Comment Classification Dataset

Jedným z najznámejších a najpoužívanějších datasetov pre detekciu toxicity je **Jigsaw Toxic Comment Classification Dataset**, ktorý bol publikovaný v rámci súťaže na platforme Kaggle v spolupráci s organizáciou Jigsaw (patriacou pod Google). Dataset obsahuje viac než 160 000 komentárov extrahovaných z anglickej Wikipédie, pričom každý komentár je anotovaný nielen binárne (toxický / netoxický), ale aj podľa viacerých konkrétnych kategórií toxického správania.

Hlavné triedy toxicity použité pri anotácii:

- **toxic** – všeobecne urážlivý alebo nepriateľský jazyk.
- **severe toxic** – extrémne agresívne alebo výhražné vyjadrenia.
- **obscene** – vulgarizmy a neslušné výrazy.
- **threat** – vyhrážky fyzickým alebo psychickým násilím.
- **insult** – osobné urážky a znevažovanie.
- **identity hate** – nenávisťné prejavy voči konkrétnym skupinám.

Komentáre sú často viacnásobne anotované (multi-label), keďže toxický výrok môže súčasne spĺňať viacero kritérií. Výhodou datasetu je jeho veľký rozsah, kvalita anotácií a dostupnosť pre komunitu. Práve preto sa stal základom pre výučbu a hodnotenie množstva klasifikačných modelov, vrátane transformerov ako BERT, RoBERTa, DistilBERT a ďalších.

Jigsaw dataset sa často používa ako benchmark v akademických aj komerčných projektoch zameraných na moderovanie obsahu. Umožňuje testovať modely na rôznych úrovniach toxicity a poskytuje dobré pokrytie bežných foriem online agresie. Vzhľadom na pôvod dát (anglická Wikipédia) však môže byť štýl jazykovo špecifický a menej reprezentatívny pre sociálne siete alebo neformálne platformy.

2.5.2 Civil Comments

Je to dataset vytvorený v rámci projektu moderácie online diskusií, ktorý vznikol v spolupráci s mediálnymi portálmi a platformou Jigsaw. Dataset pochádza z reálnych komentárov používateľov, ktorí reagovali na spravodajské články a diskusie na rôznych webových stránkach. Obsahuje približne 2 milióny komentárov, z ktorých bola vybraná reprezentatívna podmnožina (cca 450 000 komentárov) a podrobená viacnásobnému hodnoteniu.

Unikátnosť Civil Comments spočíva v tom, že každý komentár bol hodnotený viacerými anotátormi – bežnými používateľmi internetu, ktorí sa zúčastnili na hodnotení v rámci crowdsourcingovej platformy. Výsledkom je takzvaná «distribučná anotácia», kde toxickosť komentára nie je binárna, ale vyjadrená ako podiel používateľov, ktorí ho označili ako problematický.

Anotované atribúty zahŕňajú:

- **toxicity** – celková miera urážlivosti alebo nevhodnosti výrazu.
- **severe toxicity** – vážne alebo extrémne prejavy nenávisť.

- **insult** – priame alebo nepriamo formulované osobné urážky, ktoré sa týkajú intelektu, vzhľadu, charakteru alebo iných vlastností osoby.
- **obscene** – používanie vulgárnych alebo hrubých výrazov.
- **threat** – vyhrožovanie sa násilím, ublížením alebo inou formou zastrašovania, ktoré je adresované konkrétnej osobe alebo skupine.
- **attack on commenter** – osobný útok na autora iného komentára.
- **inflammatory language** – provokatívne, zbytočne konfrontačné formulácie.

Tento dataset je často využívaný na výskum «fairness» a «biasu» v NLP modeloch, pretože obsahuje aj metadáta o tom, či sa komentár týka určitých skupín (ako ženy, náboženstvá, etnické skupiny), čo umožňuje testovať rovnosť výkonnosti modelu naprieč rôznymi kategóriami.

Civil Comments je ideálny pre úlohy, kde je dôležitá presná a citlivá detekcia jemne formulovanej toxicity, ktorá sa môže pohybovať na hranici prípustnosti. Modely trénované na tomto datasete dokážu lepšie zohľadniť kontext a rôzne perspektívy používateľov. Vďaka charakteru dát je dataset využiteľný nielen na klasifikáciu, ale aj na vývoj modelov pre spravodlivé a transparentné rozhodovanie v systémoch automatizovanej moderácie.

2.5.3 ToxiGen

Ide o veľký dataset, ktorý navrhli vedci z MIT a Allen Institute for AI s cieľom identifikovať citlivé, skryté a modelmi generované formy toxického obsahu. Obsahuje viac než 270 000 anglických výrokov, pričom výnimočnosť datasetu spočíva v tom, že značná časť týchto výrokov bola vygenerovaná pomocou veľkých jazykových modelov (ako GPT-3), a následne manuálne anotovaná ľuďmi.

Dataset pokrýva široké spektrum toxických prejavov, pričom anotácie rozlišujú výroky ako:

- **toxické voči identitám** – stereotypné alebo znevažujúce výroky namierené proti etnickým, náboženským, rodovým a iným menšinám,
- **toxické vo forme latentnej agresie** – vyjadrenia, ktoré vyzerajú slušne, no ich obsah má urážlivý alebo dehumanizujúci charakter,
- **toxické v sémantickom kontexte** – výroky, ktorých urážlivý význam vyplýva z kontextu a nie je prítomný priamo na úrovni lexiky.

Na rozdiel od klasických datasetov, ToxiGen umožňuje modelom učiť sa rozpoznávať aj nepriame formy predsudkov a diskriminácie, ktoré môžu byť štylisticky neutrálne, ale hodnotovo problematické. Je to dôležité najmä v súvislosti s tým, ako sa jazykové modely používajú v reálnych aplikáciách – generovaný text môže vyzeráť korektne, ale zároveň obsahovať škodlivé stereotypy.

Tento dataset slúži aj ako benchmark pre testovanie takzvaného *toxicity classifiers*, ktoré sú určené na detekciu problematického výstupu generatívnych AI systémov. Modely trénované na ToxiGen, ako napríklad **ToxiGen Classifier**, dosahujú vysokú citlivosť na jemné formy verbálnej toxicity a sú preto využívané v prostredí automatizovanej moderácie obsahu a AI bezpečnosti.

Vďaka svojej koncepcii je ToxiGen vhodným doplnkom k datasetom ako Jigsaw či Civil Comments – neobsahuje len reálny, používateľmi napísaný text, ale aj syntetické výroky, čo umožňuje komplexnejšie trénovanie a testovanie stability klasifikačných modelov.

Spôsoby anotácie, jazyková variabilita a bias

Aj keď jednotlivé datasety používané pri detekcii toxicity sa líšia svojím pôvodom, všetky riešia podobné výzvy spojené s anotáciou a jazykovou diverzitou. Rozdiely v typoch anotácie (binárna vs. distribučná), úrovni subjektivity hodnotenia či prítomnosti kultúrne špecifických výrazov môžu výrazne ovplyvniť výkonnosť modelov.

Zároveň väčšina dostupných datasetov pochádza z anglicky hovoriaceho prostredia, čo sťažuje prenosnosť modelov na iné jazyky. Problémom môže byť aj tzv. mo-

delový bias – teda tendencia modelov nesprávne klasifikovať výroky určitých skupín ako toxické. Tieto faktory je preto dôležité zohľadniť pri výbere trénovacích dát a interpretácii výsledkov.

2.6 Výzvy a etické aspekty

Aj napriek technickému pokroku v oblasti NLP a transformerových modelov, detekcia toxicity so sebou prináša viacero otvorených výziev a etických otázok. Automatizované systémy môžu vykazovať sklony k falošne pozitívnym klasifikáciám – teda označiť ako toxický aj komentár, ktorý je v skutočnosti neškodný. K tomu často dochádza pri používaní irónie, sarkazmu alebo pri citáciách, kde význam závisí od širšieho kontextu.

Ďalším problémom je subjektivita hodnotenia. Vnímanie urážlivosti sa líši medzi jednotlivcami aj kultúrami, čo sťažuje tvorbu univerzálnych pravidiel. Modely trénované na konkrétnych datasetoch môžu byť náchylné na bias voči určitým skupinám alebo štýlom vyjadrovania, čím môže dôjsť k nespravodlivému penalizovaniu používateľov.

Z hľadiska spoločenskej zodpovednosti je dôležité zvážiť, kto nesie zodpovednosť za moderáciu – či už ide o prevádzkovateľov platforiem, tvorcov modelov alebo ich používateľov. Automatizácia nesmie nahrádzať kritické hodnotenie kontextu a musí byť doplnená o možnosť transparentného odvolania alebo revízie. Eticky navrhnuté systémy by mali rešpektovať slobodu prejavu, no zároveň aktívne chrániť používateľov pred škodlivým obsahom.

3 Praktická realizácia riešenia

3.1 Cieľ praktickej časti

Praktická časť práce bola zameraná na dôkladné oboznámenie sa s modernými metódami spracovania prirodzeného jazyka (NLP), ktoré sa používajú pri detekcii toxicity v online textoch. Mojmím zámerom bolo pochopiť princípy fungovania týchto prístupov — od klasických algoritmov až po súčasné transformerové modely (podrobnejšie) — a získať praktickú skúsenosť s ich aplikáciou.

Rovnako dôležitým cieľom bolo prejsť si celý praktický proces implementácie systému na detekciu toxicity. Tento proces zahŕňal:

- výber a analýzu reálnych datasetov obsahujúcich toxické komentáre.
- predspracovanie textových dát vrátane čistenia, normalizácie a tokenizácie.
- implementáciu inferencie pomocou predtrénovaných transformerových modelov.
- testovanie modelov na vybranej vzorke textov a zber výstupov.
- vyhodnotenie výkonnosti modelov pomocou metrík ako accuracy, F1-score, recall a inference latency.
- porovnanie výstupov viacerých modelov na rovnakých vstupoch a identifikácia ich silných a slabých stránok.

Záverečným krokom bolo vytvorenie webovej aplikácie, ktorá umožňuje koncovému používateľovi prostredníctvom prehľadného rozhrania overiť mieru toxicity vo vlastnom texte. Aplikácia je napojená na modely, ktoré počas experimentov dosiahli najlepšie výsledky, čím zabezpečuje kvalitné a spoľahlivé výstupy. Tento nástroj zároveň slúži ako praktická ukážka toho, ako možno transformerové modely efektívne využiť v reálnych podmienkach pri detekcii škodlivého obsahu.

3.2 Práca s dátami

Spracovanie dát prebieha po jednotlivých etapách nasledovne:

1. **Načítanie dát** – import súborov vo formáte CSV alebo JSON a získanie textového obsahu spolu s anotáciami toxicity.
2. **Čistenie textu** – odstránenie HTML značiek, URL odkazov, špeciálnych znakov a zbytočných medzier.
3. **Normalizácia** – prevod textu na malé písmená, štandardizácia formátu a odstránenie diakritiky (ak je to potrebné).
4. **Tokenizácia** – rozdelenie textu na jednotlivé slová alebo tokeny; spôsob tokenizácie závisí od typu použitého modelu.
5. **Vektorizácia** – konverzia tokenov do numerickej podoby:
 - pri klasických modeloch (ako Logistic Regression) pomocou TF-IDF alebo n-gramových reprezentácií,
 - pri hlbokých modeloch (ako BERT) cez `input_ids` a `attention_mask`.
6. **Formátovanie vstupov** – vytvorenie jednotného dátového formátu vhodného pre vstup do modelu (ako slovníky alebo JSON objekty).
7. **Uloženie pripravených dát** – export finálnej verzie vstupov na ďalšie spracovanie modelmi.

3.3 Práca s modelmi

Klasické modely (Logistic Regression, SVM, Naive Bayes) — Tieto modely pracujú na základe vektorových reprezentácií textu pomocou metód ako TF-IDF alebo n-gramy. Logistic Regression modeluje pravdepodobnosť triedy, SVM optimalizuje hranicu medzi triedami a Naive Bayes využíva štatistické rozdelenie výskytu slov s predpokladom ich nezávislosti.

- **Výhody:** veľmi rýchle tréovanie, nízke nároky na výpočtové zdroje, jednoduchá interpretácia výsledkov, spoľahlivé pri explicitne formulovanej toxicite.
- **Nevýhody:** slabá práca s kontextom a jazykovými vzťahmi, neschopnosť zachytiť iróniu alebo latentné formy toxicity.

Neuronové siete (CNN, RNN, LSTM) — Tieto modely pracujú so sekvenčným charakterom textu. CNN extrahujú lokálne jazykové vzory, RNN modelujú postupnosť slov v čase a LSTM zlepšujú pamäť modelu pri dlhodobých závislostiach v texte.

- **Výhody:** schopnosť pracovať s kontextom, lepšie zachytávanie významových súvislostí než klasické modely, vhodné na dlhšie texty.
- **Nevýhody:** pomalšie tréovanie, vyššia výpočtová záťaž, obmedzenia pri veľmi dlhých alebo zložito štruktúrovaných vstupoch.

Transformery (BERT, RoBERTa, DistilBERT, Tiny-toxic-detector, GPT) — Modely založené na architektúre self-attention umožňujú paralelné spracovanie vstupu a efektívne modelovanie vzťahov medzi slovami bez ohľadu na ich poradie. BERT využíva obojsmerný kontext, RoBERTa zlepšuje tréovanie, DistilBERT ponúka zjednodušenú verziu, Tiny-toxic-detector je extrémne kompaktný a GPT funguje ako generátor textu s možnosťou klasifikácie pomocou promptovania.

- **Výhody:** vysoká presnosť, výborné spracovanie kontextu, schopnosť identifikovať jemné jazykové nuansy vrátane irónie a nepriamej agresie.
- **Nevýhody:** veľmi vysoké nároky na výpočtový výkon, dlhší čas inferencie, zložitá interpretácia výstupov, potreba veľkých množstiev dát.

Na základe testovania rôznych modelov som dospel k záveru, že najpresnejšie výsledky poskytli modely **BERT, RoBERTa** a **Tiny-toxic-detector**, ktoré boli následne implementované do webovej aplikácie ako základ detekčného systému.

3.4 Metódy detekcie toxicity

Detekcia toxicity môže byť formulovaná rôznymi spôsobmi v závislosti od cieľa systému a charakteru trénovacích dát.

Binárna klasifikácia — Model rozdeľuje vstupy do dvoch tried: `toxic` a `non-toxic`. Ide o základný typ klasifikácie, ktorý je často využívaný ako referenčný systém alebo pri nedostatku dát na detailnejšie rozlíšenie.

- **Výhody:** jednoduchá implementácia, rýchly tréning, nízke nároky na dáta, ľahká interpretácia výstupu.
- **Nevýhody:** neumožňuje rozlišovať medzi rôznymi typmi toxicity, znižuje prínos pri hlbšej analýze prejavov.

Multi-label klasifikácia — Každý komentár môže byť zaradený do viacerých toxických kategórií naraz (ako `obscene`, `insult`, `identity hate`). Tento prístup sa uplatňuje pri rozsiahlejšie anotovaných datasetoch, ako je Jigsaw.

- **Výhody:** podrobnejšia analýza prejavov toxicity, vyššia výpovedná hodnota pre moderovanie, možnosť štatistického sledovania jednotlivých typov útokov.
- **Nevýhody:** vyššia zložitosť klasifikácie, náročnosť na kvalitu a rovnováhu anotácií, komplikovanejšia interpretácia výsledkov.

Citlivá klasifikácia a jazykové zvláštnosti — Niektoré prejavy toxicity sú vyjadrené nepriamo alebo s použitím irónie, sarkazmu a kultúrne podmienenej symboliky. Ich odhalenie si vyžaduje nielen prácu s významom slov, ale aj analýzu širšieho kontextu a pragmatiky.

- **Výhody:** schopnosť zachytiť skrytú agresiu a pasívne výrazy, lepšia ochrana pred nenápadnými formami útokov, relevantnejšie výstupy pre moderáciu.

- **Nevýhody:** ťažšie anotovateľné, vyžadujú výkonné modely ako transformery, riziko nesprávnej interpretácie pri nedostatočnom kontexte.

3.5 Mechanizmus self-attention

Mechanizmus *self-attention* je základným stavebným prvkom transformerových modelov. Na rozdiel od sekvenčných architektúr, ktoré spracúvajú text postupne, self-attention umožňuje každému prvku vstupu hodnotiť svoj vzťah ku všetkým ostatným slovám v rámci celej vety alebo dokumentu súčasne. Tento princíp je kľúčový pre efektívne pochopenie globálneho kontextu, čo je obzvlášť dôležité pri detekcii jazykovo komplexných prejavov toxicity.

Základný princíp — Každé slovo vo vstupe je zakódované ako vektor a následne transformované do troch reprezentácií: **query** (dotaz), **key** (kľúč) a **value** (hodnota). Pre každé slovo sa vypočíta váha pozornosti voči ostatným slovám ako skalárny súčin **query** a **key**, normalizovaný pomocou funkcie **softmax**. Tento výstup určuje, na ktoré slová má model «sústrediť pozornosť». Finálny vektor sa potom vypočíta ako vážený súčet všetkých **value** vektorov. Výsledkom je kontextualizovaná reprezentácia každého tokenu.

Tento proces možno formálne zapísať ako:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

kde Q , K , V sú matice dotazov, kľúčov a hodnôt, a d_k je rozmer kľúčových vektorov. Táto operácia umožňuje modelu priamo prepájať významovo vzdialené časti textu bez ohľadu na ich pozíciu.

Prínos pre detekciu toxicity — Self-attention je obzvlášť účinný pri rozpoznávaní implicitného, nepriamo vyjadreného alebo ironického obsahu. Umožňuje modelu zohľadniť celkový význam výroku, čo je zásadné pri moderovaní komentárov, kde sa toxicita často maskuje v sarkazme alebo dlhších stylistických konštrukciách.

- **Výhody:** paralelné spracovanie vstupov, schopnosť zachytiť dlhodobé jazykové závislosti, presná analýza kontextu, aplikovateľnosť v rôznych NLP úlohách.
- **Nevýhody:** vysoké nároky na výpočtové zdroje, rast pamäťovej náročnosti so zvyšujúcou sa dĺžkou vstupu, zložitá interpretácia rozhodovacích procesov modelu.

3.6 Hodnotiace metriky

Úspešnosť modelov na detekciu toxicity sa hodnotí pomocou štandardných metrík klasifikácie, ktoré vyjadrujú mieru správnosti predikcií v porovnaní so skutočnými anotáciami. Výber konkrétnych metrík závisí od typu úlohy (binárna alebo multi-label klasifikácia) a od dôležitosti rôznych aspektov výstupu — napríklad presnosti pri odhaľovaní agresívneho obsahu alebo schopnosti minimalizovať falošné aktivácie.

Presnosť (Precision) — Vyjadruje podiel správne klasifikovaných toxických komentárov z celkového počtu tých, ktoré model označil ako toxické.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Napríklad ak model označí 100 komentárov ako toxické, ale len 70 z nich skutočne toxické, presnosť je 70 %.

- **Výhody:** dôležitá metrika pri obmedzení falošnej cenzúry — teda pri minimalizácii zásahov do neškodného obsahu.
- **Nevýhody:** ignoruje počet toxických komentárov, ktoré model nezachytil (falošné negatíva).

Citlivosť (Recall) — Miera zachytenia všetkých skutočne toxických prípadov modelom.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Ak je medzi 100 toxickými komentármi zachytených 85, recall je 85 %.

- **Výhody:** dôležitá metrika pri ochrane pred škodlivým obsahom, najmä v moderátorských systémoch.
- **Nevýhody:** vysoký recall môže byť dosiahnutý na úkor presnosti (model označuje veľa komentárov ako toxické, aj keď nie sú).

F1-skóre — Harmonický priemer presnosti a citlivosti. Používa sa v prípadoch, kde je dôležitá rovnováha medzi týmito dvoma faktormi.

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **Výhody:** poskytuje vyvážený pohľad na kvalitu modelu pri nevyvážených dátach, kde je počet toxických a netoxických komentárov nerovnomerný.
- **Nevýhody:** môže maskovať extrémne hodnoty jednej zo zložiek (ako veľmi nízku presnosť pri vysokom recalle).

Accuracy (presnosť klasifikácie) — Vyjadruje celkový podiel správnych predikcií zo všetkých vstupov. Napriek popularite však pri nevyvážených dátach môže zavádzať — ak je len 5 % komentárov toxických, model, ktorý vždy predikuje `non-toxic`, bude mať 95 % accuracy, hoci nezachytí žiadnu toxicitu.

- **Výhody:** jednoduchá interpretácia, vhodná pri vyvážených datasetoch.
- **Nevýhody:** nevhodná pre unbalanced dáta — neodráža reálnu výkonnosť pri detekcii menšinovej triedy.

Pri multi-label klasifikácii sa jednotlivé metriky môžu počítať osobitne pre každú triedu a následne sa agregujú buď **makro-priemerom** (rovnaká váha pre každú triedu) alebo **mikro-priemerom** (vážený priemer podľa počtu prípadov). Výber závisí od cieľa analýzy — makro priemer zdôrazňuje výkon na menej častých triedach, zatiaľ čo mikro priemer preferuje dominantné kategórie.

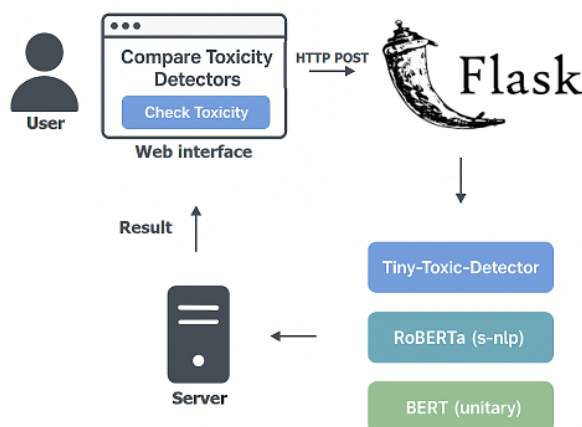
3.7 Popis webového riešenia

V rámci praktickej časti bol vytvorený webový systém, ktorý umožňuje používateľom otestovať mieru toxicity textu pomocou niekoľkých transformerových modelov. Aplikácia beží na serveri, ktorý poskytuje rozhranie pre analýzu vstupného textu a vráti porovnanie výstupov troch modelov: **BERT**, **RoBERTa** a **Tiny-Toxic-Detector**.

3.7.1 Architektúra systému

Celý systém je navrhnutý ako klient–server aplikácia. Používateľská časť (frontend) beží v prehliadači a komunikuje s backend serverom, ktorý zabezpečuje predspracovanie vstupu, volanie modelov a spätnú väzbu.

- **Frontend** – zobrazuje formulár na zadanie textu, vizualizuje výsledky a históriu dotazov. Interakcia prebieha cez HTML a JavaScript.
- **Backend** – zabezpečuje logiku aplikácie v prostredí Flask (Python), volá jednotlivé NLP modely a generuje odpoveď vo forme štruktúrovaného JSON objektu.
- **Modelová vrstva** – integruje tri transformerové modely (Tiny-Toxic-Detector, RoBERTa a BERT), ktoré vracajú skóre toxicity a latenciu.



Obrázok 3 – 1 Schéma architektúry webovej aplikácie

3.7.2 Použité technológie

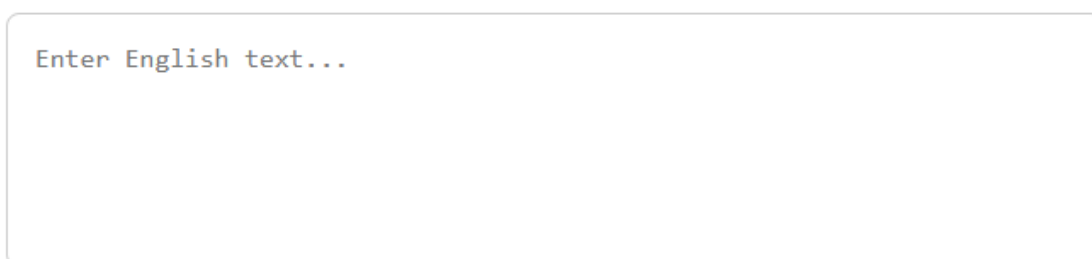
- **Flask** – webový framework v jazyku Python, ktorý slúži ako backend server, obsluhuje HTTP požiadavky a prepája používateľské rozhranie s NLP modelmi.
- **Hugging Face Transformers** – knižnica, cez ktorú boli načítané modely BERT a RoBERTa (pipeline a tokenizácia).
- **Tiny-Toxic-Detector** – Kompaktný transformerový model založený na práci (10), implementovaný ako samostatný modul v jazyku Python pomocou knižnice PyTorch. Na rozdiel od modelov BERT a RoBERTa, tento model nebol načítaný cez API, ale bol manuálne integrovaný. Model je optimalizovaný pre rýchlu inferenciu a nízku pamäťovú záťaž.
- **HTML/CSS + JavaScript** – používateľské rozhranie s možnosťou vkladať text, odosielať dotazy, sledovať výsledky a históriu.

3.7.3 Všetky funkcie webovej aplikácie

Používateľ po načítaní stránky môže:

1. Zadať ľubovoľný anglický text (max. 512 symbolov) do textového poľa.

Compare Toxicity Detectors

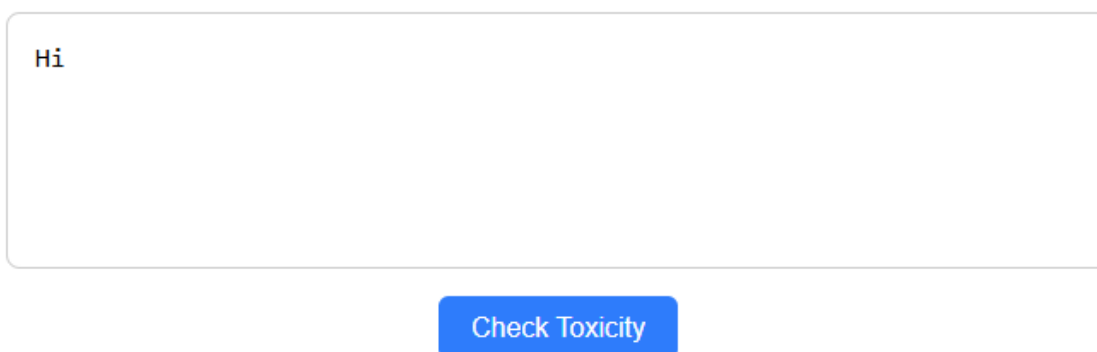


Enter English text...

Obrázok 3–2 Pole pre zadanie textu na kontrolu

2. Spustiť analýzu toxicity stlačením tlačidla «Check Toxicity», ktoré môže byť aktívne, ak je v textovom poli aspoň 1 znak, alebo neaktívne, ak je pole prázdne.

Compare Toxicity Detectors



Hi

Check Toxicity

Obrázok 3–3 Aktívne tlačidlo

Compare Toxicity Detectors

Enter English text...

Check Toxicity

Obrázok 3–4 Neaktívne tlačidlo

3. Zobrazit výsledky v prehľadnej tabuľke: názov modelu, skóre toxicity a latencia.

Results

| Model | Toxicity Score | Latency (ms) |
|---------------------|----------------|--------------|
| Tiny-Toxic-Detector | 0.0 | 8.54 |
| RoBERTa (s-nlp) | 0.0 | 51.82 |
| BERT (unitary) | 0.001 | 61.09 |

Obrázok 3–5 Pole výsledok

4. Vizualne identifikovať úroveň toxicity pomocou farebného rozlíšenia (zelená = nízka, žltá = stredná, červená = vysoká).

Results

| Model | Toxicity Score | Latency (ms) |
|---------------------|----------------|--------------|
| Tiny-Toxic-Detector | 1.0 | 6.33 |
| RoBERTa (s-nlp) | 0.064 | 51.42 |
| BERT (unitary) | 0.559 | 55.86 |

Obrázok 3 – 6 Príklad toho, ako môžu vyzeráť farby výsledkov

Vzhľadom na to, že každý model je iný, niekedy sa vyskytujú rôzne výsledky. V tomto prípade sme takýto výsledok dostali na túto správu «You're not the worst person I've ever worked with, but you're definitely in the running. Mediocrity seems to follow you like a shadow.».

5. Prehliadať históriu predchádzajúcich vstupov, všetko, čo bolo zadané predtým (od spustenia aplikácie až doteraz, ak nebolo použité tlačidlo «Clear History»).

History

| Input | Tiny | RoBERTa | BERT |
|----------------------------------------------------------------------------------------------------------------------------------------------------------|-------|---------|-------|
| You're completely useless in every meeting, and the only thing you contribute is dead air. It's like you make the whole room dumber just by being there. | 1.0 | 0.998 | 0.706 |
| I love you!! | 0.0 | 0.0 | 0.001 |
| If there was an award for incompetence, you'd not only win it — you'd redefine it. Honestly, you're a burden to the team. | 0.999 | 0.244 | 0.148 |
| Well, I didn't expect much from you, and somehow you still | 0.0 | 0.0 | 0.001 |

[Download CSV](#)
[Clear History](#)

Obrázok 3–7 Príklad toho, ako môže vyzeráť história výsledkov

6. Možnosť stiahnuť históriu vo formáte CSV alebo ju vymazať pomocou tlačidiel «Download CSV» a «Clear History», ak nie je prázdna.

History

| Input | Tiny | RoBERTa | BERT |
|-------|------|---------|------|
|-------|------|---------|------|

[Download CSV](#)
[Clear History](#)

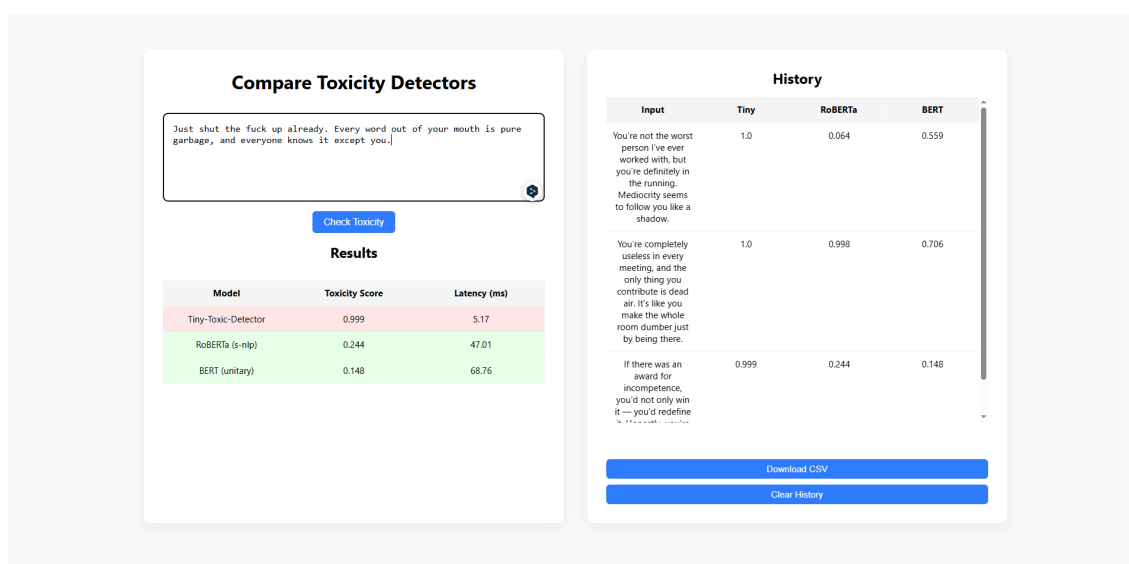
Obrázok 3–8 Príklad správania tlačidiel za predpokladu, že história je prázdna

7. Po stlačení tlačidla «Download CSV» za predpokladu, že história nie je prázdna, je možné získať tabuľku hodnotenia toxicity konkrétneho textu.

| Text | Tiny | RoBERTa | BERT |
|----------------------------------------------------|------|---------|-------|
| You're not the worst person I've ever worked with. | 1.0 | 0.064 | 0.559 |

Tabuľka 3 – 1 Ukážka výsledkov detekcie toxicity pre rôzne modely

3.7.4 Celkový výhľad aktívne používanej webovej aplikácie



Obrázok 3 – 9 Webová aplikácia

3.7.5 Záver a možnosti rozšírenia

V rámci tejto aplikácie sa podarilo úspešne realizovať všetko, čo bolo pôvodne naplánované. Používateľské rozhranie bolo navrhnuté s dôrazom na jednoduchosť, prehľadnosť a interaktivitu. Všetky výpočty prebiehajú lokálne, bez potreby vzdialeného servera.

Aj napriek tomu vždy existuje priestor na zlepšenie. Do budúcnosti by bolo možné aplikáciu rozšíriť o porovnanie s väčším množstvom modelov, nasadiť ju na verejný

hosting, pridať podporu pre ďalšie jazyky a takisto implementovať jemnejšiu klasifikáciu podľa skupín (napríklad: urážky, nenávisť, výsmech).

Takéto rozšírenia by mohli výrazne zlepšiť presnosť a použiteľnosť aplikácie v reálnych scenároch.

3.7.6 Testovanie a vyhodnotenie výstupov modelov

Na účely testovania boli v aplikácii zadané rôzne typy textov — od pozitívnych a neutrálnych po ironické, pasívne agresívne a otvorene toxické. V tabuľke nižšie sú uvedené výsledky hodnotenia toxicity pre každý text, vygenerované tromi modelmi.

Tabuľka 3–2: Výsledky testovania toxicity modelmi
Tiny, RoBERTa a BERT

| Vstupný text | Tiny | RoBERTa | BERT |
|-------------------------------------------------------------|-------|---------|-------|
| If we were grading on effort, you'd still fail. | 0.000 | 0.000 | 0.027 |
| Shut up already. You're embarrassing yourself. | 1.000 | 0.998 | 0.948 |
| That's an interesting idea — not smart, but interesting. | 0.000 | 0.000 | 0.001 |
| I love working with you, you're always so helpful and kind. | 0.000 | 0.000 | 0.001 |
| You're the reason projects fail. | 1.000 | 0.006 | 0.466 |
| Nobody asked for your opinion, just stop. | 1.000 | 0.000 | 0.001 |

| Vstupný text | Tiny | RoBERTa | BERT |
|-------------------------------------------------------------|-------|---------|-------|
| You clearly have no idea what you're doing. | 0.000 | 0.000 | 0.002 |
| Wow, it takes real skill to mess things up this badly. | 0.002 | 0.000 | 0.001 |
| Could you please try to be on time next time? | 0.559 | 0.000 | 0.001 |
| Thank you for your help today, it really made a difference. | 0.000 | 0.000 | 0.001 |
| It's truly astonishing how consistently unhelpful you are. | 0.986 | 0.013 | 0.019 |
| Let me know if you need any help, I'm happy to support you. | 0.000 | 0.000 | 0.001 |

Z výsledkov vyplýva, že:

- **Tiny-Toxic-Detector** má tendenciu výrazne reagovať na toxické vstupy — vo viacerých prípadoch vrátil skóre 1.0, čo hovorí o vysokej citlivosti. V niektorých prípadoch označil aj neutrálny alebo mierne kritický text ako toxický (*Could you please try to be on time... – 0.559*).
- **RoBERTa** je oveľa konzervatívnejšia — skóre takmer vždy ostáva blízko 0, výnimkou je len silne toxický vstup (ako *Shut up already*, kde skóre dosiahlo 0.998).
- **BERT** sa správa vyvážene — nereaguje veľmi výrazne, ale pri otvorené toxických vetách vráti skóre 0.9+, čím preukazuje rozumné rozlíšenie medzi mierne a vysoko toxickým obsahom.

4 Záver

Vykonaná práca predstavuje komplexný základný výskum v oblasti detekcie toxicity v textoch s využitím transformerových modelov. Pozornosť bola venovaná nielen analýze dostupných prístupov v rámci spracovania prirodzeného jazyka (NLP), ale aj ich praktickému nasadeniu. Výstupom je funkčné riešenie, ktoré umožňuje testovanie a porovnávanie rôznych modelov v reálnom čase prostredníctvom vytvoreného webového rozhrania.

Práca zároveň poskytuje praktický návod pre tých, ktorí chcú začať alebo pokračovať v tejto oblasti. Popisuje celý cyklus vývoja NLP riešenia — od predspracovania údajov, cez výber a testovanie architektúr, až po ich aplikáciu na konkrétnu úlohu. Aj keď je dôraz kladený na transformerové modely, čitateľ získa širší prehľad o súčasných prístupoch k detekcii toxického obsahu.

Medzi výhody riešenia patrí jeho komplexnosť, praktická použiteľnosť a otvorenosť pre ďalší rozvoj. Webová aplikácia poskytuje možnosť interaktívne testovať rôzne modely a porovnávať ich výsledky, čo zvyšuje prehľadnosť celého riešenia. Samotná práca má ambíciu byť dobrým základom pre ďalšie realizácie.

Nevýhodou je, že testovanie bolo vykonané iba na anglických údajoch a že niektoré časti by si vyžadovali hlbšiu technickú analýzu.

Napriek dosiahnutému pokroku ostáva problematika detekcie toxicity otvorenou a trvalo aktuálnou výzvou. V budúcnosti je možné riešenie ďalej vylepšovať — rozšíriť o viacjazyčnú podporu, zapojiť pokročilejšie modely či doplniť webové rozhranie o analytické nástroje alebo možnosti interakcie s existujúcimi platformami

Ale potrebné si rozumieť, že v tejto oblasti pravdepodobne nikdy nebude možné nájsť úplné a univerzálne riešenie. Dôvodom je vysoká jazyková, kultúrna a hodnotová rozmanitosť globálneho komunikačného priestoru, ktorá sťažuje vytvorenie jednotného prístupu k definícii a klasifikácii toxického správania. Akékoľvek pokusy o absolútnu kontrolu by narazili na etické a sociálne obmedzenia, ktoré sú nevyhnutné na zachovanie slobody prejavu v demokratickej spoločnosti. Z tohto dôvodu bude

detekcia toxického správania vždy vyžadovať citlivú rovnováhu medzi technologickou presnosťou a rešpektovaním ľudských práv.

Literatúra

- [1] Schütz, M., Demus, C., Pitz, J., Probol, N., Siegel, M., Labudde, D. DeTox at GermEval 2021: Toxic Comment Classification. 2021.
- [2] Rajeshwari, K., Manish, B. S. A Review on Automated Approaches to Detection of Social Media Toxicity. 2024.
- [3] Sundelin, C. Comparing Different Transformer Models' Performance for Identifying Toxic Language Online. 2023.
- [4] Villate-Castillo, G., Del Ser, J., Sanz Urquijo, B. A Systematic Review of Toxicity in Large Language Models. 2024.
- [5] Fan, H., Du, W., Dahou, A., Ewees, A. A., Yousri, D., Abd Elaziz, M., Elsheikh, A. H., Abualigah, L., Al-qaness, M. A. A. Social Media Toxicity Classification Using Deep Learning: Real-World Application UK Brexit. 2021.
- [6] Wang, S., Chen, B. A Comparative Study of Attention-Based Transformer Networks and Traditional Machine Learning Methods for Toxic Comments Classification. 2023.
- [7] Laugier, L., Pavlopoulos, J., Sorensen, J., Dixon, L. Civil Rephrases Of Toxic Texts With Self-Supervised Transformers. 2021.
- [8] Guillaume, P., Duchêne, C., Dehak, R. Hate Speech and Toxic Comment Detection using Transformers. 2022.
- [9] Lashkarashvili, N., Tsintsadze, M. Toxicity Detection in Online Georgian Discussions. 2022.
- [10] Abu Nada, A. H., Latif, S., Qadir, J. Lightweight Toxicity Detection in Spoken Language: A Transformer-based Approach for Edge Devices. 2023.

- [11] Jigsaw AI Research Team. ToxiGen Dataset: Advancing Detection of Subtle Toxicity. 2021.
- [12] Google Jigsaw. Civil Comments Dataset: A Scalable Approach to Evaluating Online Toxicity. 2019.
- [13] Bonetti, A., Martínez-Sober, M., Torres, J. C., Vega, J. M., Pellerin, S., Vila-Francés, J. Comparison between Machine Learning and Deep Learning Approaches for the Detection of Toxic Comments on Social Networks. 2023.
- [14] Romero, S. E., Castellana, D., Barrón-Cedeño, A. GTH-UPM at DETOXIS-IberLEF 2021: Automatic Detection of Toxic Comments in Social Networks. 2021.
- [15] Joseph, J., Rose, S., Jayakumar, M. Cognitive Method to Detect Toxic Comments in Social Media. 2024.
- [16] Kamphuis, M. TINY-TOXIC-DETECTOR: A Compact Transformer-Based Model for Toxic Content Detection. 2024.
- [17] Li, J., Xie, Y. The Impact of Transformer Models on Detecting Hate Speech and Toxicity. 2024.

Využitie umelej inteligencie pri príprave práce

Pri písaní tejto bakalárskej práce bola použitá umelá inteligencia, konkrétne jazykový model ChatGPT od spoločnosti OpenAI, ktorý slúžil ako pomocný nástroj v niekoľkých oblastiach:

- pre lepšie pochopenie vedeckých článkov a ich zhrnutie do zrozumiteľnejšej podoby,
- ako pomoc pri riešení technických problémov v jazykoch Python a LaTeX,
- ako jazykový pomocník pri preklade do slovenského jazyka (ktorý nie je mojím materinským jazykom), na kontrolu pravopisu a interpunkcie,
- pri formulovaní textu v akademickom a profesionálnom štýle.

Prílohy

Príloha A zip súbor s projektom Python, ktorý implementuje webovú aplikáciu, je nahraný na git.

Príloha B Používateľská príručka

Príloha C Systémová príručka

Príloha B

Táto používateľská príručka popisuje spôsob používania webovej aplikácie na detekciu toxicity textov

Použitie aplikácie

1. Po otvorení stránky používateľ uvidí jednoduché rozhranie s textovým vstupným poľom. Do tohto poľa je možné vložiť anglický text na analýzu (maximálne **512 znakov**).
2. Po zadaní textu stlačte tlačidlo **Check Toxicity**.
3. Aplikácia následne zobrazí výsledok určenia toxicity a spotrebovaného času
4. Každá analýza sa automaticky uloží do **histórie**. História sa zobrazí pod výsledkom analýzy ako zoznam predchádzajúcich vstupov a ich hodnotení.
5. Používateľ má možnosť s históriou ďalej pracovať:
 - Pomocou tlačidla **Download CSV** si môže stiahnuť všetky predchádzajúce záznamy vo forme CSV súboru.
 - Pomocou tlačidla **Clear History** môže používateľ vymazať všetky predchádzajúce záznamy. Táto akcia je **nevratná**.

Obmedzenia

- Texty musia byť zadané v anglickom jazyku. Detekcia pre iné jazyky nemusí byť presná.
- Maximálna dĺžka vstupného textu je 512 znakov.

Príloha C

Táto príručka popisuje základné kroky potrebné na spustenie aplikácie na detekciu toxicity.

1. Stiahnite a nainštalujte programovací jazyk Python vo verzii **3.10**. Odporúčaný zdroj: <https://www.python.org/downloads/release/python-3100/>.
2. Zo GitHub si stiahnite projektový archív vo formáte **.zip**.
3. Rozbaľte stiahnutý archív do požadovaného priečinka na vašom počítači.
4. Otvorte celý priečinok projektu pomocou vývojového prostredia pre Python, ako napríklad **PyCharm**, **VS Code** alebo iný editor podľa výberu.
5. Vo vývojovom prostredí otvorte súbory **app.py** a **tiny_toxic_detector.py**. Systém pravdepodobne podčiarkne červenou farbou názvy chýbajúcich knižníc.
6. Postupne nainštalujte všetky potrebné knižnice.
7. Keď budú všetky knižnice úspešne nainštalované, spustite aplikáciu pomocou súboru **app.py**.
8. Po úspešnom spustení bude aplikácia dostupná cez webový prehliadač na lokálnej adrese, napríklad **http://127.0.0.1:5000**.