STŘEDOŠKOLSKÁ ODBORNÁ ČINNOST

Obor č. 18: Informatika

Model strojového učení zaměřený na klasifikaci politického zabarvení textu

Matouš Volf
Pardubický kraj

Pardubice, 2025

STŘEDOŠKOLSKÁ ODBORNÁ ČINNOST

Obor č. 18: Informatika

Model strojového učení zaměřený na klasifikaci politického zabarvení textu

Machine learning model for classifying political leaning of text

Jméno: Matouš Volf
Škola: DELTA – Střední škola informatiky a ekonomie, s.r.o.
Kraj: Pardubický kraj
Konzultant: doc. Ing. Jakub Šimko, Ph.D.

Pardubice, 2025

Prohlášení

Prohlašuji, že jsem svou práci SOČ vypracoval samostatně a použil jsem pouze prameny a literaturu uvedené v seznamu bibliografických záznamů.

Prohlašuji, že tištěná verze a elektronická verze soutěžní práce SOČ jsou shodné.

Nemám závažný důvod proti zpřístupňování této práce v souladu se zákonem č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon) ve znění pozdějších předpisů.

V Pardubicích dne ....

Poděkování

Děkuji doc. Ing. Jakubu Šimkovi, Ph.D. za vedení, konzultace a sdílení cenných zkušeností při tvorbě projektu.

Abstrakt

Tématem této práce je úloha automatické klasifikace textů podle politického zabarvení a političnosti pomocí modelů strojového učení. Sestavili jsme obsáhlý přehled existujících datových sad a modelů pro tyto úlohy, přičemž zjišťujeme, že většina stávajících prací vytváří neuniverzální řešení s nízkou přenositelností na typy textu, které nebyly přítomny v trénovací sadě. Pro překonání tohoto omezení kombinujeme 12 existujících datových sad pro klasifikaci politického zabarvení. Pro političnost vytváříme novou datovou sadu tím, že značkujeme příslušným štítkem 18 existujících sad s různými tématy a povahami textu. Prostřednictvím rozsáhlé experimentace vyhodnocujeme úspěšnost stávajících modelů a trénujeme nové s vylepšenými generalizačními schopnostmi.

Klíčová slova

Politické zabarvení, političnost, klasifikace textu, strojové učení, transformer

Abstract

This thesis addresses the challenge of automatically classifying text according to political leaning and politicalness using machine learning models. We have composed a comprehensive overview of existing datasets and models for these tasks, finding that most current approaches create siloed solutions that perform poorly on out-of-distribution texts. To address this limitation, we have compiled a diverse dataset by combining 12 datasets for political leaning classification and creating a new dataset for politicalness by extending 18 existing vivid datasets with the appropriate label. Through extensive experiments, we have evaluated the performance of existing models and train new ones with enhanced generalization capabilities.

Keywords

Political leaning, politicalness, text classification, machine learning, transformer

Obsah

1 Úvod

Automatická klasifikace politického zabarvení textu má mnoho praktických využití a byla již předmětem výzkumu. Ten se zaměřoval především na zpravodajské články z internetových médií (Baly et al. 2020; Demidov 2023) a příspěvky ze sociálních sítí (Preoţiuc-Pietro et al. 2017; Xiao et al. 2020).

Strojová klasifikace tohoto druhu může být využívána například agregátory novinových článků, jejichž cílem je prezentovat jednotlivé zprávy ve světle různých ideologických úhlů pohledu (např. AllSides[1] nebo Ground News[2]). Jak tyto stránky uvádějí, zaujatost médií není nutně špatná – *skrytá* zaujatost je to, co klame a polarizuje společnost. Snaží se ji zmírnit tím, že čtenářům poskytují vyváženou směs názorů, což jim umožňuje vidět širší obraz a zvážit i argumenty protistrany. To zabraňuje konfirmačnímu zkreslení a vzniku tzv. sociálních bublin (ve kterých jsou uživatelé izolováni a omezeni na obsah kurátorovaný doporučovacími systémy) a komnat ozvěn (kde jsou uživatelé opakovaně vystaveni stejným názorům, které spíše posilují než zpochybňují jejich stávající přesvědčení). Tyto stránky by dokonce mohly model využít k identifikaci toho, co Ground News nazývá blindspoty[3] – novinových zpráv s politickými podtóny, které jsou nepřiměřeně pokrývány zdroji pouze z jedné strany politického spektra. Kromě médií AllSides také sleduje politické zabarvení nejvýznamnějších zdrojů fact-checků[4].

Další možné případy užití zahrnují kupříkladu nástroje pro auditování obsahu zveřejňovaného politiky za účelem ověření jejich postoje. Podobné nástroje by mohly být také použity k odhalení skrytých politických agend.

Umístění textů na politické spektrum je něco, co mohou lidští anotátoři provádět manuálně. To se však stává nákladným při práci s velkým a v poslední době stále rostoucím množstvím dostupných dat, což vede k potřebě tento proces automatizovat.

V kontextu této analýzy je *politické zabarvení* kategorizováno do tří tříd: levice, střed a pravice. Jedná se o zjednodušení skutečného problému – jak levice, tak pravice jsou samy o sobě spektra zahrnující názory od mírných po extrémnější postoje. Samotná klasifikace na jediné ose také představuje omezení – politologie uznává i vícerozměrné systémy politického zabarvení (klasifikující postoje jednotlivě např. vůči ekonomickým politikám, uznávání lidských práv či autoritářství). Navíc platí, že jeden text ve skutečnosti může podporovat obě strany (například vyjadřuje-li se rozličně k několika různým tématům),

[1] `https://www.allsides.com`

[2] `https://ground.news`

[3] `https://ground.news/blindspot/about`

[4] `https://www.allsides.com/media-bias/fact-check-bias-chart`

což jej potenciálně může zařazovat do jakési „smíšené“ třídy.

S touto úlohou vyvstává předřazený problém: pokud vstupní text vůbec není o politice, chování a výstup modelu jsou nedefinované. Řešením je filtrovat vstupní texty na základě tzv. *političnosti*. Političnost definujeme jako binární třídu: Text je politický, pokud jeho hlavním tématem je politika. Jinak je nepolitický. Ideální definice by brala v úvahu nejen samotné téma, ale spíše to, zda text vyjadřuje politický *názor*. Některé texty (např. hesla z encyklopedie) totiž mohou pojednávat o politických tématech, ve skutečnosti však neobsahují žádné známky podpory určité politické ideologie. Tento druh klasifikace je však mnohem obtížnější a vyžaduje velmi jemně rozlišená data, a proto jsme se rozhodli pokračovat se zmíněnou jednodušší definicí.

Současný stav výzkumu není uspokojivý: Jak ukážeme, existující práce zabývající se politickým zabarvením mají tendenci vytvářet řešení, která selhávají na neznámých typech textů. Na těchto textech nebylo provedeno žádné důkladné testování a **naše hypotéza je, že modely trénované na jednom typu textu nejsou dobře přenositelné na jiné typy, aniž by značně klesla jejich úspěšnost**. Navíc, ačkoli bylo publikováno několik klasifikačních modelů pro političnost, nejsou k této úloze dostupné dostatečně velké a rozmanité datové sady.

Přínosy této práce jsou následující:

1. sestavili jsme obsáhlý přehled existujících datových sad a modelů pro úlohy klasifikace političnosti a politického zabarvení,
2. vytvořili jsme novou datovou sadu pro klasifikaci političnosti spojením a označkováním několika existujících datových sad,
3. shromáždili a sjednotili jsme dostupná data se značkou politického zabarvení,
4. změřili jsme úspěšnost existujících klasifikačních modelů,
5. natrénovali a otestovali jsme sadu vlastních nových modelů.

Spolu s touto prací zveřejňujeme kompletní zdrojový kód[5] našeho výzkumu.

[5]Dostupný na `https://github.com/matous-volf/political-leaning-prediction`.

2 Související práce

2.1 Existující studie na politickou klasifikaci textu

Ačkoli existuje mnoho prací zaměřených na predikci politického zabarvení, žádná z nich se nezabývá problémem filtrování nepolitických textů. Rovněž jsme neobjevili žádné studie věnované výhradně političnosti.

Klasifikace pomocí zpracování přirozeného jazyka (NLP) není jediným způsobem, jak předpovědět politické názory jednotlivců – například lze využít vztahy mezi uživateli a příspěvky na sociálních sítích. Tento přístup zvolili Wong et al. (2016), v níž analyzovali chování uživatelů sociální sítě X (tehdy nesoucí název Twitter) při sdílení příspěvků a zkoumali překryvy publika různých populárních účtů za účelem odvození politické náklonnosti. Gu et al. (2017) využili data o sledování, zmínkách a retweetech k odhadu numerických ideologických pozic. Stefanov et al. (2020) dokonce zkombinovali tento přístup s NLP tím, že nejprve aplikovali automatické shlukování (tzv. učení modelu bez učitele) uživatelů na základě retweetů, označili jednotlivé shluky politickým zabarvením a následně trénovali model FastText na textech s těmito značkami.

Řešení této úlohy pomocí NLP je poměrně složité. Existující studie experimentovaly s použitím tradičních technik strojového učení – jako je regrese (Preoţiuc-Pietro et al. 2017) nebo SVM (Zhitomirsky-Geffet et al. 2016) – i s hlubokým učením neuronových sítí (Iyyer et al. 2014; Baly et al. 2020). Pro řešení komplexních úloh klasifikace textu představují velmi účinnou metodu modely založené na architektuře transformer. Demidov (2023) ukázal, že jsou vhodné i konkrétně pro úlohu klasifikace politického zabarvení. Všechny modely shromážděné a testované v naší studii vycházejí z některého z jazykových modelů BERT (Devlin et al. 2019), RoBERTa (Zhuang et al. 2021), DeBERTa (He et al. 2021) nebo DistilBERT (Sanh et al. 2020) postavených právě na této architektuře.

V posledních letech se schopnosti velkých generativních jazykových modelů (LLM) – ostatně také založených na tranformerech –, jako například GPT nebo Llama, výrazně posunuly kupředu, což otevírá nové možnosti v různých oblastech, včetně klasifikace textů podle politických charakteristik. Heseltine a Hohenberg (2024) rozebírají možnost využití LLM místo manuální anotace lidskými pracovníky a Törnberg (2024) dokonce naznačuje, že LLM mohou překonat amatérské anotátory i experty. Halterman (2025) navrhuje generování syntetických dat pomocí generativních LLM. Dvě datové sady použité v našem výzkumu (Jones 2024; Nayak 2024a) byly vytvořeny právě tímto způsobem. Ačkoli tyto

obrovské modely mohou velmi dobře porozumět politickým tématům a mohou být úspěšné při klasifikaci bez dalšího tréninku (tzv. zero-shot nebo few-shot), jejich použití má i své stinné stránky, kterým se podrobněji věnujeme v diskuzi (kapitola 5).

Publikovány byly také průzkumové metastudie (Doan a Gulla 2022; Németh 2022), které poskytují přehled aktuálních technik, přístupů a datových sad pro zmíněné úlohy. Velmi užitečný pro nás byl také souhrn dostupných metod, který uvádí Burnham (2024), přestože je zaměřen na detekci postojů (tzv. stance detection), která je v NLP samostatnou úlohou.

Příbuznou, ale odlišnou úlohou od těch zkoumaných v této práci je detekce *politické zaujatosti* (Gangula, Duggenpudi a Mamidi 2019) – zda a do jaké míry autor vyjadřuje preference pro nebo proti určitému politickému subjektu. Dalším blízkým problémem je predikce politické strany, jejímž je autor textu příslušníkem (Høyland et al. 2014).

2.2 Shromážděná data

Veškeré datové sady jsme získali z veřejně dostupných zdrojů na internetu. Prohledali jsme Hugging Face Hub[1], Kaggle[2] a Zenodo[3] pomocí jednoduchého univerzálního dotazu „politic“, prohlédli všechny výsledky vyhledávání a shromáždili veškeré relevantní položky. Všechny texty jsou v angličtině a týkají se kontextu Spojených států amerických. Tabulky 3.1 a 3.2 obsahují seznam všech shromážděných datových sad. Některé názvy jsme zkrátili pro snazší práci s nimi. V pozdějších sekcích analyzujeme obsah datových sad podrobněji.

2.2.1 Datové sady pro politické zabarvení

Celkem jsme vybrali 12 datových sad označkovaných podle politického zabarvení: Article bias prediction (Baly et al. 2020); BIGNEWSBLN (Liu et al. 2022); CommonCrawl news articles (Spliethöver, Keiff a Wachsmuth 2022); Dem., rep. party platform topics (Wolbrecht et al. 2023a; Wolbrecht et al. 2023b); GPT-4 political bias (Jones 2024); GPT-4 political ideologies (Nayak 2024a); Media political stance (España-Bonet 2023); Political podcasts (Narayana 2024); Political tweets (Steyn 2023); Qbias (Haak a Schaer 2023); Webis bias flipper 18 (Chen et al. 2018); Webis news bias 20 (Chen et al. 2020).

2.2.2 Datové sady pro političnost

Pro političnost jsme získali celkem 18 datových sad. Pouze dvě z nich mají požadované značky explicitně: PoliBERTweet (Kawintiranon a Singh 2022) a Political or not (Burnham et al. 2024). Zbytek jsme získali procházením výše zmíněných repozitářů, přičemž jsme vyhledávali kvalitní textová data s pestrou škálou témat. Tyto datové sady jsme museli označkovat sami (jak je popsáno v metodologické sekci 3.1.2): Amazon reviews 2023

[1] `https://huggingface.co/datasets`
[2] `https://www.kaggle.com/datasets`
[3] `https://zenodo.org`

(Hou et al. 2024); Dialogsum (Chen et al. 2021); Free news (Webhose.io n.d.); Goodreads book genres (Szemraj 2023); IMDB (Maas et al. 2011); IMDB movie genres (Heymans 2022); Medium post titles (amrrs 2022); News category (Misra a Grover 2021; Misra 2022); PENS (Ao et al. 2021); Recipes (Bian 2022); Reddit comments (Baumgartner et al. 2020); Reddit submissions (Baumgartner et al. 2020); Rotten tomatoes (Pang a Lee 2005); Textbooks (Phi 2023); Tweet topic multi (Antypas et al. 2022); Yelp review full (Zhang, Zhao a LeCun 2015).

2.3 Existující modely

Existující modely jsme vyhledávali na Hugging Face Hub[4]. Použili jsme jednoduchý a univerzální dotaz „politic“ a prohlédli všechny výsledky, abychom zajistili shromáždění co největšího počtu relevantních modelů.

Dva z nich jsou unikátní v tom, že mohou být využity pro širokou škálu úloh – nejsou natrénované pouze pro jednu konkrétní. POLITICS (Liu et al. 2022) je předtrénovaný jazykový model na zpravodajských článcích o politice, vytvořený rozšířeným trénováním (tzv. continued training) RoBERTy. Nelze jej použít přímo na konkrétní koncové úlohy, ale musí být doladěn (tzv. fine-tuning). Tento postup je totožný jako u obecných jazykových modelů vycházejících z architektury transformer (např. BERT a jeho odnože), tento by však měl dosahovat lepších výsledků na textech souvisejících s politikou, pro něž je optimalizován. Political DEBATE (Burnham et al. 2024) je klasifikátor určený pro inferenci přirozeným jazykem (NLI) trénovaný pro klasifikaci politických dokumentů bez nutnosti dalšího tréninku (tzv. zero-shot a few-shot). Tyto modely byly testovány jak na úloze političnosti, tak politického zabarvení.

2.3.1 Modely pro politické zabarvení

Celkem jsme nalezli 8 modelů trénovaných pro klasifikaci politického zabarvení: BERT political bias fine-tune (Vélez Castañeda 2024); DeBERTa political classification (Palmqvist 2024); DistilBERT-PoliticalBias (Jones 2024); DistilBERT political fine-tune (Shrimali 2024); DistilBERT political tweets (Newhauser 2022) Political bias BERT (Research 2023); Political bias prediction AllSides DeBERTa (Sahitaj 2024); Political ideologies RoBERTa fine-tuned (Nayak 2024b).

2.3.2 Modely pro političnost

Podařilo se nám najít 2 modely trénované pro klasifikaci političnosti: Classifier main subject politics (gptmurdock 2024) a Topic politics (Silcock et al. 2024).

[4] `https://huggingface.co/models`

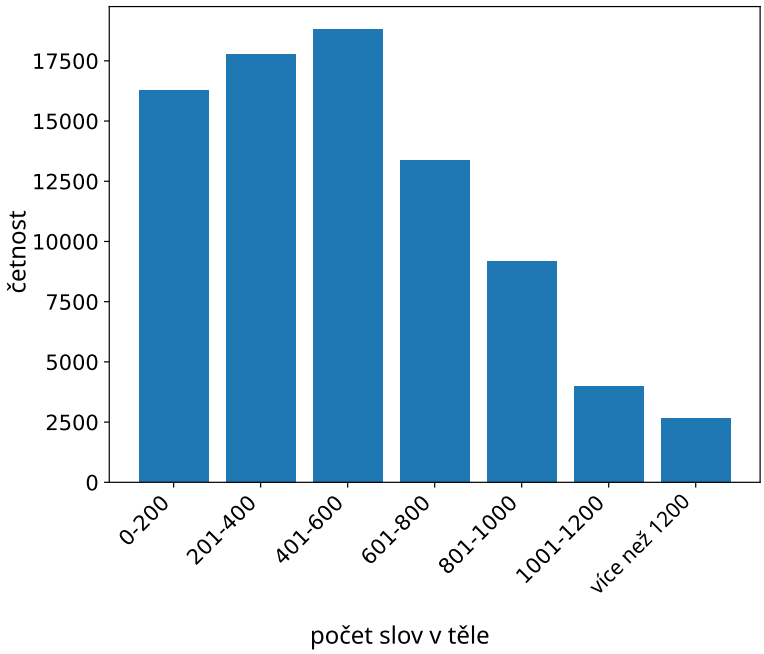
3 Metodologie

3.1 Předzpracování dat

Všechny datové sady obsahují textová těla a některé z nich i titulky. V takových případech je titulek připojen (se dvěma zalomeními řádku) k textovému tělu a až tento celek je předán jako vstup modelům.

Většina datových sad obsahuje určité množství velmi krátkých textů, které ve skutečnosti pro trénink ani měření nepřinášejí žádnou hodnotu. Pro každou z těchto sad jsme určili vhodnou spodní hranici a odstranili všechny příklady, které byly kratší.

Obrázek 3.1 ukazuje typickou distribuci délek textu, kterou sdílí téměř všechny datové sady.



Obrázek 3.1: Distribuce počtu slov textových těl v datové sadě CommonCrawl news articles. Většina datových sad má křivku distribuce velmi podobnou této – liší se pouze maximální délkou textu a celkovým počtem záznamů.

Největší z datových sad byly pro účely tohoto výzkumu zmenšeny, neboť v experimentech je důležité zachovat rovnováhu mezi počty záznamů z každé sady. To znamená, že obrovské množství dat přítomné v těchto sadách by stejně nemohlo být využito, jelikož by narušilo rovnováhu ve srovnání s menšími sadami. Datové sady BIGNEWSBLN, CommonCrawl news articles, Media political stance, Political tweets a Yelp review full byly zmenšeny na 100 000 vzorků systematickým vzorkováním (vybíráním záznamů po pevném intervalu) rovnoměrně podle délky textu, přičemž bylo zajištěno vyrovnané výsledné rozdělení tříd.

3.1.1 Politické zabarvení

Datové sady pro politické zabarvení se skládají z novinových článků, příspěvků na sociálních sítích, krátkých výroků (některé synteticky generované pomocí LLM) a popisů epizod podcastů. Tabulka 3.1 je zobrazuje spolu s dalšími vlastnostmi. Některé z nich neobsahují střední třídu. Naproti tomu datové sady CommonCrawl news articles a GPT-4 political bias obsahují pět tříd – přičemž levá a pravá strana jsou rozděleny do dvou úrovní (mírné a extrémní). Štítky byly redukovány na 3 třídy namapováním obou těchto úrovní na jedinou třídu levé nebo pravé strany. Vzhledem k tomu, že texty pocházejí ze Spojených států, je třída s levým politickým zabarvením definována především liberální ideologií a názory demokratické strany, zatímco třídě s pravým zabarvením odpovídají konzervativní a republikánské názory. Novinové články jsou označkovány agregátorem novinových článků AllSides. Tento zdroj anotací považujeme za důvěryhodný, protože – jak uvádí Baly et al. (2020) – vychází z přísného procesu zahrnujícího slepé průzkumy zaujatosti, redakční revize, analýzu třetích stran, nezávislé recenze a zpětnou vazbu komunity. Obrázek 3.2 ukazuje, jak jsou média kategorizována v době psaní tohoto článku. Datové sady Article bias prediction a Qbias jsou anotovány na úrovni jednotlivých článků, u ostatních jsou štítky automaticky odvozeny z celkového politického zabarvení mediálního zdroje. To přináší určitou míru nepřesnosti, avšak Baly et al. (2020) zjistili, že individuálně anotované příklady se od obecného politického zabarvení liší pouze v 3,11 % případů, což je přijatelné.

Po vyhodnocení stávajících modelů pro političnost (sekce 3.4.2) jsme zjistili, že ten nejlepší dosahuje velmi dobrých výsledků (sekce 4.2.2) na všech datových sadách kromě jediné – CommonCrawl news articles (která je určena pro politické zabarvení, a tudíž je zcela označena jako politická). Na této datové sadě dosáhl přesnosti pouhých 33 % – nižší, než jakou by měl model odhadující zcela náhodně. Prozkoumali jsme jednotlivé příklady, které model identifikoval jako nepolitické, a potvrdili jsme, že většina z nich skutečně byla. Zvažovali jsme úplné odstranění této datové sady z analýzy, nicméně pro její velikost a kvůli tomu, že obsahuje i texty ze střední třídy, ji považujeme za hodnotnou. Z tohoto důvodu jsme model využili jako filtr a odstranili pouze nepolitické texty. Abychom minimalizovali odstraňování falešně negativních případů (textů, které skutečně jsou politické), považovali jsme text za nepolitický pouze tehdy, když si byl model svým rozhodnutím

velmi jistý, tj. skóre na koncovém klasifikačním neuronu bylo vyšší než 0,99.

název	celkový počet	distribuce tříd (L/S/P)	průměrný počet slov ± směrodatná odchylka v těle	v titulku	druh	zdroj značek
BIGNEWSBLN	100 000	33 / 33 / 33	557 ± 411	9 ± 3	noviny	AllSides
CommonCrawl news articles	100 000	33 / 33 / 33	516 ± 351		noviny	AllSides
Media political stance	100 000	50 / 0 / 50	618 ± 360		noviny	AllSides
Political tweets	100 000	50 / 0 / 50	18 ± 5		sociální síť	author
Article bias prediction	37 550	35 / 29 / 37	1 084 ± 751	10 ± 3	noviny	AllSides
Qbias	21 715	47 / 20 / 33	66 ± 28	7 ± 3	noviny	AllSides
Dem., rep. party platform topics	11 557	53 / 0 / 47	60 ± 71		výroky	party
Political podcasts	11 047	48 / 0 / 52	101 ± 48	8 ± 4	popisy epizod	author
Webis news bias 20	7 722	47 / 16 / 37	819 ± 943	9 ± 3	noviny	AllSides
Webis bias flipper 18	6 400	37 / 24 / 39	884 ± 1 128	10 ± 3	noviny	AllSides
GPT-4 political ideologies	3 200	50 / 0 / 50	62 ± 10		výroky	GPT-4
GPT-4 political bias	612	30 / 29 / 41	11 ± 2		výroky	GPT-4

Tabulka 3.1: Datové sady pro politické zabarvení, seřazené podle velikosti. Zdroj značek: AllSides je popsán v sekci 3.1.1; příspěvky z Political tweets jsou psány členy demokratické nebo republikánské strany; každý podcast z Political podcasts je známý podporou buď liberálních nebo konzervativních postojů; výroky GPT-4 jsou generovány spolu se značkou od LLM.

Obrázek 3.2: Zaujatost amerických médií (AllSides 2024).

3.1.2 Političnost

Ze všech datových sad zaměřených na političnost mají dvě – PoliBERTweet a Political or not – explicitní štítky. Tyto sady však nejsou dostatečně rozmanité (obsahují pouze omezený rozsah témat) a jejich velikost není dostačující pro trénink modelu. Z tohoto důvodu jsme sestavili vlastní datovou sadu sloučením několika různých sad (zbylých 16), které obsahují politické a / nebo nepolitické texty. Tyto texty zahrnují novinové články, příspěvky na sociálních sítích a typicky nepolitické texty, jako jsou uživatelské recenze, shrnutí dějů knih a filmů nebo každodenní dialogy. Tabulka 3.2 je zobrazuje spolu s dalšími vlastnostmi. Pokud byly texty v datové sadě rozděleny podle témat, pečlivě jsme zkontrolovali obsah textů každého tématu, abychom buď přiřadili správný štítek, anebo všechny záznamy o daném tématu odstranili – pokud téma nebylo jednoznačně poltické či nepolitické – abychom do dat nevnášeli šum.

Datová sada Textbooks obsahuje pouze 1 795 textů. jsou však velmi dlouhé – v průměru kolem 30 000 slov. Rozhodli jsme se je rozdělit po každých 10 odstavcích.

Do analýzy jsme také zahrnuli datové sady zaměřené na politické zabarvení, přičemž všechny příklady byly označeny jako politické.

Výsledná agregovaná datová sada obsahuje texty mnoha různých typů a z rozmanitých zdrojů, a slouží tak jako univerzální standard pro měření. Model na ní natrénovaný by měl být schopen dobře generalizovat a zvládnout širokou škálu možných vstupů. Podle našeho průzkumu se jedná o první datovou sadu pro političnost o této velikosti, rozsahu a kvalitě.

název	počet nepolitických	počet politických	průměrný počet slov ± směrodatná odchylka v těle	v titulku	druh	zdroj značek
Textbooks	262 082	1 088	202 ± 100		učebnice	témata
Reddit comments	207 150	0	47 ± 79		sociální síť	témata
News category	94 566	24 986	24 ± 13	9 ± 3	noviny	témata
IMDB movie genres	104 063	0	101 ± 74		popisy filmů	témata
Yelp review full	100 000	0	96 ± 60		recenze	implicitní
IMDB	98 469	0	233 ± 173		recenze	implicitní
Reddit submissions	82 069	0	120 ± 179	10 ± 6	sociální síť	témata
PENS	70 694	45 92	599 ± 753	10 ± 3	noviny	témata
Recipes	65 390	0	21 ± 9		recepty	implicitní
Amazon reviews 2023	61 715	0	69 ± 102	4 ± 3	recenze	implicitní
Medium post titles	52 171	3 567	12 ± 6	8 ± 3	blogové příspěvky	témata
Tweet topic multi	21 416	0	28 ± 12		sociální síť	manuální
PoliBERTweet	10 000	10 000	23 ± 15		sociální síť	hashtagy
Free news	15 722	1 612	580 ± 600	12 ± 5	noviny	témata
Dialogsum	12 812	0	122 ± 68		dialogy	implicitní
Rotten tomatoes	10 647	0	21 ± 9		recenze	implicitní
Goodreads book genres	7 165	0	156 ± 89		popisy knih	témata
Political or not	2 174	1 681	245 ± 98		různé	témata

Tabulka 3.2: Datové sady pro političnost, seřazené podle velikosti. Zdroj značek: implicitní znamená, že texty jsou typicky a v drtivé většině nepolitické; témata byla označkována, jak je popsáno v sekci 3.1.2.

3.2 Průnik datových sad

Při práci s více datovými sadami je nezbytné ověřit, zda se některé z nich překrývají – zda sdílejí určité množství identických záznamů. Je důležité stanovit velikost těchto průniků a výzkum tomu přizpůsobit. Když překrytí mezi dvěma sadami není zanedbatelné (pod 10 % jejich velikosti), bude jím ovlivněn trénink i vyhodnocení modelů. Použití jedné z těchto sad pro vyhodnocení modelu, který byl natrénován na té druhé, povede k příliš optimistickým a zavádějícím výsledkům, neboť testovací sada bude pravděpodobně obsahovat příklady, se kterými se model již během tréninku setkal. Z tohoto důvodu identifikujeme datové sady s překryvy a vylučujeme je z experimentů, kde by k tomuto jevu mohlo dojít.

Průnik textů mezi každou dvojicí datových sad byl měřen porovnáním titulků a textových těl. Oba tyto prvky byly předem zbaveny všech nepísmenných znaků (dokonce i mezer) a převedeny na malá písmena, aby se neprojevily nepodstatné textové odlišnosti.

Preferovaným postupem je prosté porovnání titulků a zjištění, zda se přesně shodují. Nicméně některé záznamy nebo dokonce celé datové sady neobsahují pole s titulkem, a proto je nutné porovnávat nějakým způsobem textová těla. Primitivní porovnání řetězců znak po znaku v tomto případě není vyhovující, jelikož různé webové škrabky zahrnují do textu různé části stránky – například nadpis komentářové sekce. Sofistikované algoritmy pro měření podobnosti řetězců (jako je např. Levenshteinova vzdálenost) nejsou proveditelnou možností, protože mají vysokou časovou složitost pro tak velké množství textu. Bylo proto zvoleno výpočetně jednodušší řešení: vyříznout střední část textového těla (50 znaků, případně méně, pokud je text kratší) z textového těla v první datové sadě a ověřit, zda je tato část obsažena v nějakém těle ve druhé datové sadě. Vyříznutí střední části považujeme za optimální, neboť tato část má nejmenší pravděpodobnost nesouladu způsobeného odlišnostmi ve škrábání webu. Tato metoda ovšem poskytuje pouze odhad skutečného průniku – například se mohou objevit falešné shody, když dva texty citují ve své prostřední části stejný výrok.

Podrobné informace k měření průniku jsou uvedeny v příloze A.

3.3 Evaluace modelů

Veškeré evaluace modelů (jak existujících, tak našich nově vyvinutých) na dostupných datových sadách byla provedena s následující metodologií.

Validační a testovací sady jsou vytvořeny vzorkováním celých datových sad buď na pevný počet, nebo na určitý zlomek záznamů. Použití zlomku způsobuje, že velikost každého vzorku je různá, v závislosti na velikosti zdrojové sady. To by představovalo problém, pokud by tyto vzorky byly následně kombinovány do jediné validační nebo testovací sady, protože zastoupení jednotlivých datových sad v ní by bylo výrazně nevyvážené. Z tohoto důvodu takové vzorky nikdy nekombinujeme a místo toho testujeme modely na každé z nich samostatně, a pokud je to žádoucí a vhodné, vypočítáme průměr jednotlivých metrik. Tímto způsobem každý vzorek ovlivňuje průměr stejnou měrou, bez ohledu na svou velikost (a velikost datové sady, z níž pochází).

Vzorkování vždy zajišťuje rovnoměrné rozložení tříd přítomných ve vzorku a je systematické (vybírá řádky v pravidelných intervalech) podle délky textu v každé třídě. Pokud je počet řádků v datové sadě menší než velikost vzorku, pak je z každé třídy odebráno co nejvíce řádků, stále však se zachováním rovnoměrného rozložení tříd. Poté jsou měřeny přesnost (accuracy), preciznost (precision), recall, F_1 skóre a matice záměn (confusion matrix) modelu. Když nějaký model politického zabarvení nepodporuje střední třídu, je vyhodnocován pouze na textech s levým a pravým zabarvením.

3.4 Test existujících modelů

Existující modely byly vyhodnoceny pomocí výše popsané metodologie (sekce 3.3), přičemž každá z datových sad byla navzorkována na 1 000 záznamů.

3.4.1 Politické zabarvení

DistilBERT-PoliticalBias a Political ideologies RoBERTa fine-tuned jsou ve skutečnosti natrénovány k predikci liberálního vs. konzervativního zabarvení. Tento výstup byl namapován na levicové vs. pravicové zabarvení. Totéž platí pro DistilBERT political tweets, který je natrénován k predikci demokratického vs. republikánského zabarvení. 5 modelů podporuje třídu středového zabarvení, zbylé 4 rozlišují pouze levici a pravici. DistilBERT-PoliticalBias používá 5úrovňové spektrum, které bylo zredukováno na 3úrovňové namapováním jak silného, tak mírného levicového zabarvení do jediné třídy a u pravicového stejně tak. Political DEBATE klasifikuje pomocí inference přirozeným jazykem (NLI) a byla mu dodána jednoduchá (zero-shot) hypotéza `This text supports {left | center | right} political leaning.`

Podle jejich popisů na Hugging Face Hub byly některé z modelů trénovány na některých ze shromážděných datových sad. Tyto informace jsou zachyceny v tabulce 3.3.

název	natrénován na	podporuje střední třídu
Political bias BERT	Article bias prediction	✓
Political bias prediction AllSides DeBERTa	Article bias prediction	✓
DistilBERT-PoliticalBias	GPT-4 political bias	✓
DistilBERT political fine-tune	neznámé	✓
Political DEBATE large v1.0	Dem., rep. party platform topics; Political tweets	✓
BERT political bias fine-tune	neznámé	✗
Political ideologies RoBERTa fine-tuned	GPT-4 political ideologies	✗
DeBERTa political classification	neznámé	✗
DistilBERT political tweets	a subset of Political tweets	✗

Tabulka 3.3: Existující modely pro politické zabarvení.

3.4.2 Politično st

Každý z modelů jsme vyhodnotili zvlášť na každé datové sadě. Tyto výsledky by však mohly být zavádějící, protože většina datových sad obsahuje buď pouze politické (tak je tomu u sad k politickému zabarvení), nebo pouze nepolitické příklady – jen málo sad obsahuje obě třídy (jak ukazuje tabulka 3.2). Za těchto okolností, kdy jsou třídy v každé datové sadě drasticky nevyvážené (nebo některé dokonce zcela chybí), poskytují metriky (přesnost, preciznost, recall a F_1 skóre) klamavé výsledky a mohou být dokonce nedefinované kvůli dělení nulou. Proto jsme provedli evaluaci také na agregátu složeném ze vzorků (1 000 záznamů) všech datových sad, který má ve výsledku distribuci tříd vyváženou (49 % / 51 %).

3.5 Benchmarky datových sad

Natrénovali jsme několik skupin modelů, abychom porovnali charakter datových sad a jejich vhodnost pro trénink. Při veškerém trénování jsme využili aparát knihovnen PyTorch (Paszke et al. 2019) a Hugging Face transformers (Wolf et al. 2020) v jazyce Python. Trénovací sady pro benchmarky byly relativně malé (mohly být až 5krát větší). Když není uvedeno jinak, nebylo provedeno ani důkladné hledání optimálních hyperparametrů a byly použity výchozí hodnoty stanovené zmíněnou knihovnou transformers, jen s minimálními úpravami. Je to proto, že cílem těchto benchmarků nebylo vyvinout co nejlépe fungující model, ale spíše získat informace o samotných datových sadách a potvrdit naše hypotézy ohledně trénování.

Od výchozích hodnot Hugging Face jsme se odchýlili snížením maximální délky vstupu z 512 na 256 tokenů. Jeden token odpovídá průměrně zhruba jednomu slovu. Všechny tokeny nad tento limit jsou odříznuty, jinými slovy text je zkrácen z pravé strany. Toto snížení délky umožňuje větší velikosti dávek, zkracuje dobu trénování na polovinu a měřitelně neovlivňuje přesnost modelu. Naše vysvětlení je, že ve většině případů texty jasně vyjadřují své politické postoje v podstatě od začátku – pouze názory ukryté v jakési pozdější pointě nacházející se za maximální délkou vstupu by se mohly modelu skrýt. Pro průměrné délky textů v datových sadách vizte tabulky 3.1 a 3.2.

Navíc bylo nutné povolit hyperparametr zahřívacích kroků (warmup steps), který je ve výchozím stavu nulový. Bez něj se modely někdy naučily předpovídat pouze jedinou třídu a nikdy žádnou jinou. Důvodem je, že když se zahřívání nevyužívá, je rychlost učení nejvyšší na začátku tréninku. Tím pádem může optimalizátor drasticky přestřelit váhy sítě do nenávratných hodnot. Zvyšování rychlosti učení od nuly na maximum postupně a plynule (účel zahřívání) tomu zabraňuje. Poměr zahřívání jsme stanovili na 15 % z celkového počtu kroků.

Nutno je také poznamenat, že velikost dávky je ve výchozím nastavení zvolena automaticky a v případě našeho hardwaru (podrobnosti v příloze C) byla nastavena na 8.

Bylo natrénováno (fine-tuning), testováno a porovnáno několik jazykových modelů postavených na architektuře transformer, konkrétně BERT base (cased), RoBERTa base, DeBERTa V3 base a POLITICS.

3.5.1 Ponechání jedné

Z každé datové sady byl odebrán vzorek 2 000 záznamů (nebo méně, pokud je datová sada menší), přičemž bylo zajištěno rovnoměrné rozdělení tříd a systematické vzorkování podle délky textu v každé třídě. Na každém z těchto vzorků byl samostatně natrénován model. Předtím bylo 15 % datové sady vyhrazeno jako testovací sada a dalších 15 % jako sada validační, při dodržení výše předepsané metodologie (v sekci 3.3).

Tento přístup v podstatě odpovídá způsobu, jakým vznikla většina existujících modelů – trénováním a vyhodnocováním na jediné datové sadě, a tedy pouze na jednom typu textu.

Tento benchmark nebyl proveden na datových sadách pro političnost, protože většina z nich obsahuje čistě jednu třídu, a trénování modelu na nich je tak nemožné.

3.5.2 Vynechání jedné

Pokud se dvě datové sady významně překrývají (měřeno metodologií popsanou v sekci 3.2), menší z nich byla vyřazena z tohoto benchmarku, aby se zabránilo přítomnosti stejných záznamů v trénovací i v testovací sadě. Platí to pro dvě datové sady politického zabarvení: Webis bias flipper 18 a Webis news bias 20, které se výrazně překrývají navzájem a zároveň s Article bias prediction (zaznamenáno ve výsledcích v sekci 4.1).

Z každé datové sady byl odebrán vzorek (2 000 řádků v případě politického zabarvení, 1 000 v případě političnosti – nebo méně, když je sada menší) při zajištění rovnoměrného rozdělení tříd a systematickém vzorkování podle délky textu v každé třídě. Bylo natrénováno několik modelů, každý z nich na sjednocení všech těchto vzorků kromě jednoho (tj. kromě vynechané datové sady). Tento proces byl opakován, pokaždé s vynecháním jiné datové sady. Nejzajímavější evaluace modelu je pak právě na této vynechané sadě, protože ukazuje, jak dobře model zvládá nejen konkrétní neznámé texty, ale také různé styly textů nebo témata z jiného časového období. Proto se validační sada v tomto benchmarku skládá čistě z příkladů z vynechané datové sady, aby mohl být nejlepší checkpoint modelu vybrán na základě této úspěšnosti. Velikost testovací sady byla zvolena jako 15 % z každé datové sady. Použití zlomku místo pevné velikosti má výhodu v tom, že nevyčerpává malé datové sady, ale zároveň využívá větší množství dat ve velkých sadách, u nichž si to můžeme dovolit. Velikost validační sady byla stanovena na 1 000 (nebo méně, když je sada menší) ze zbývajících příkladů. Při vytváření validační i testovací sady jsme postupovali podle výše předepsané metodologie (v sekci 3.3).

Skutečnost, že některé datové sady pro politické zabarvení neobsahují žádné příklady ze středové třídy, má dopad na rozložení tříd při spojení vzorků do jednoho agregátu: vybrání stejně velkého vzorku z každé sady vede v tomto agregátu k nižšímu celkovému zastoupení středové třídy (neboť z některých sad záznamy z této třídy zcela chybí). Naše experimenty odhalily, že rozložení tříd při tréninku je klíčové pro to, aby model přehnaně neprioritizoval levou a pravou třídu. Nasazení různých vah optimalizátoru pro jednotlivé třídy (class weights) – technika, která by měla nevyrovnanost tříd vyvážit – tento problém

neřeší, jenom činí model méně přesným. Ve skutečnosti jsme zjistili, že je dokonce potřeba, aby třídy byly téměř naprosto rovnoměrně vyvážené. Proto jsme přidali koeficient, který zvyšuje počet vybraných záznamů středového zabarvení z datových sad, které je obsahují. Upravujeme jej vždy tak, aby se distribuce tříd vyrovnala.

Tento benchmark jsme provedli ještě jednou s druhým cílem: najít optimální nastavení tréninku, které by vedlo ke vzniku modelu pro klasifikaci politického zabarvení schopného dobře generalizovat i na jiných typech textů, než těch přítomných v jeho trénovací sadě. Naším předpokladem je, že nový model natrénovaný později přesně podle tohoto nastavení, tentokrát však na všech dostupných datových sadách, bude co nejvšestrannější, protože metodologie bude prověřena tímto přísným a náročným benchmarkem. Takový model by měl fungovat spolehlivě nejen na všech typech dat, které jsme shromáždili, nýbrž i na typech dosud neviděných a také na textech, které budou napsány teprve v budoucnosti. Tentokrát jsme vzali větší trénovací vzorek, podle dosavadních výsledků vybrali nejefektivnější jazykový model a zaměřili se na optimalizaci jeho úspěšnosti laděním hyperparametrů.

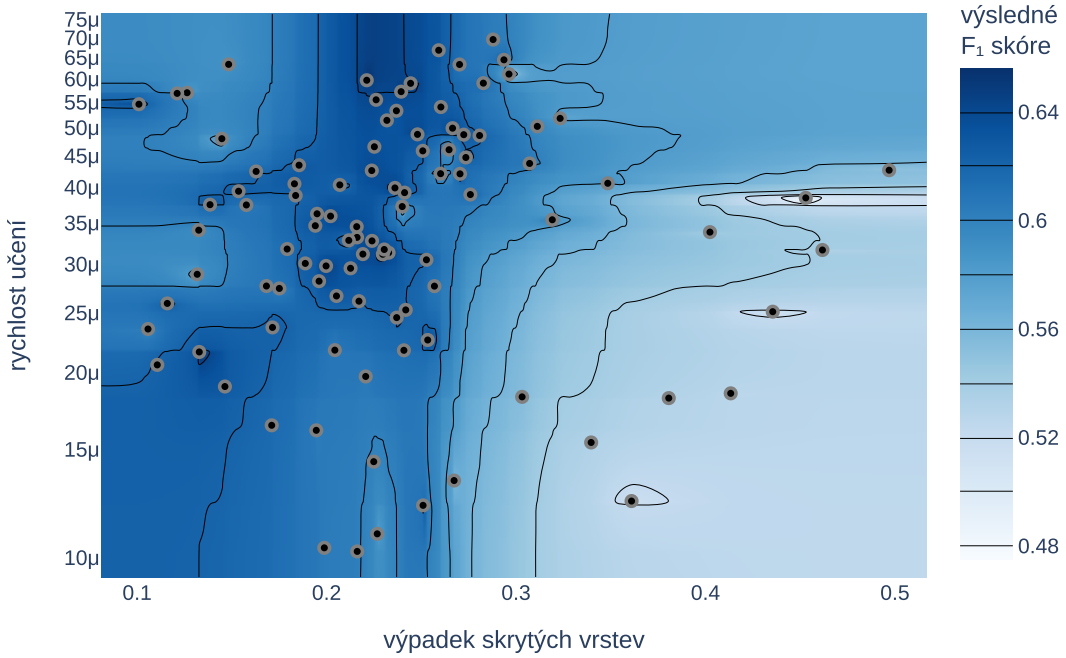
Ukázalo se, že 10 000 je dobrá velikost vzorku k odebrání z každé datové sady pro trénink, protože stále umožňuje vyvážit třídy (koeficientem zmíněným výše). Celkový počet záznamů v tréninkovém agregátu vyšel kolem 100 000.

POLITICS vyšel najevo jako nejlepší model pro pokročilé ladění, neboť konzistentně dosahoval nejvyšších přesností ve všech předchozích benchmarcích. Prozkoumali jsme konfiguraci navrženou v článku vydaném spolu s tímto modelem (Liu et al. 2022) a prozkoumali práce věnující se obecně nejvlivnějším hyperparametrům při ladění BERTa a jeho derivátů pro klasifikaci textu (Sun et al. 2019; Lorenzoni et al. 2024). Na základě těchto doporučení jsme provedli hledání hyperparametrů (hyperparameter search). Konkrétně jsme nasadili optimalizační framework Optuna (Akiba et al. 2019) v následujícím prostoru (hranaté závorky označují uzavřený interval, složené předdefinovanou množinu a hodnoty následující po šipce jsou finálně zvolené jako nejoptimálnější):

1. Výpadek pozornosti (attention dropout): [0,1; 0,3] → 0,1106.
2. Výpadek skrytých vrstev (hidden layers dropout): [0,1; 0,5] → 0,2212.
3. Výpadek klasifikační vrstvy (classifier layer dropout): [0,0, 0,5] → 0,07925.
4. Rychlost učení (learning rate): [1e-5, 7e-5], vzorkováno logaritmicky → 5,97e-5.
5. Poměr zahřívání (warmup ratio): [0,05; 0,25] → 0,20885.
6. Velikost dávky (batch size): {8; 16; 32; 48; 64} → 64.
7. Úpadek vah (weight decay): [0,0001; 0,1], vzorkováno logaritmicky → 0,00015.

Počet epoch pro každý pokus byl nastaven na 5 a pro porovnání pokusů byl vždy z tréninku vybrán checkpoint s nejlepšími výsledky na validační sadě. Obrázek 3.3 ukazuje

vrstevnice výsledných skóre na různých kombinacích hodnot dvou nejvlivnějších hyperparametrů.



Obrázek 3.3: Vrstevnice výsledných F_1 skóre v závislosti na hyperparametrech, jež měly v hledání největší vliv – výpadek skrytých vrstev (hidden layers dropout) a rychlost učení (learning rate). Tečky jsou jednotlivé pokusy. Je vidět, že úpadek skrytých vrstev nad 013 je příliš intenzivní a již znemožňuje modelu učení. Nejlepších výsledků však model dosahuje při hodnotách kolem 0,22. Optimální rychlost učení je pak přibližně 6e-5.

Pro trénovací sadu takové velikosti již může být výhodné použít větší model. POLITICS je však k dispozici pouze v základní variantě, a proto jsme se rozhodli experimentovat s laděním DeBERTy V3 large. Vybrali jsme právě ji, protože je jedním z nejmodernějších derivátů BERTa a její architektura představuje špičku mezi aktuálně dostupnými jazykovými modely. Hledání hyperparametrů již nebylo možné, protože tento model je daleko náročnější na VRAM grafické karty a trvá dlouho jej doladit na hardwaru dostupném pro tento výzkum (podrobnosti v příloze C). Hyperparametry proto musely být upraveny ručně. Po přibližně 20 pokusech jsme dospěli k těmto konečným úpravám výchozích hodnot Hugging Face:

1. Maximální délka vstupu: 512 → 256 tokenů. Důvody jsou popsané v sekci 3.5.
2. Rychlost učení (learning rate): 5e-5 → 3e-5. Zvýšení vedlo k tomu, že model přestřelil minima ztrátové funkce, a činilo trénink nestabilním. Snížení způsobovalo pomalou konvergenci a potenciální uvíznutí v lokálním minimu.

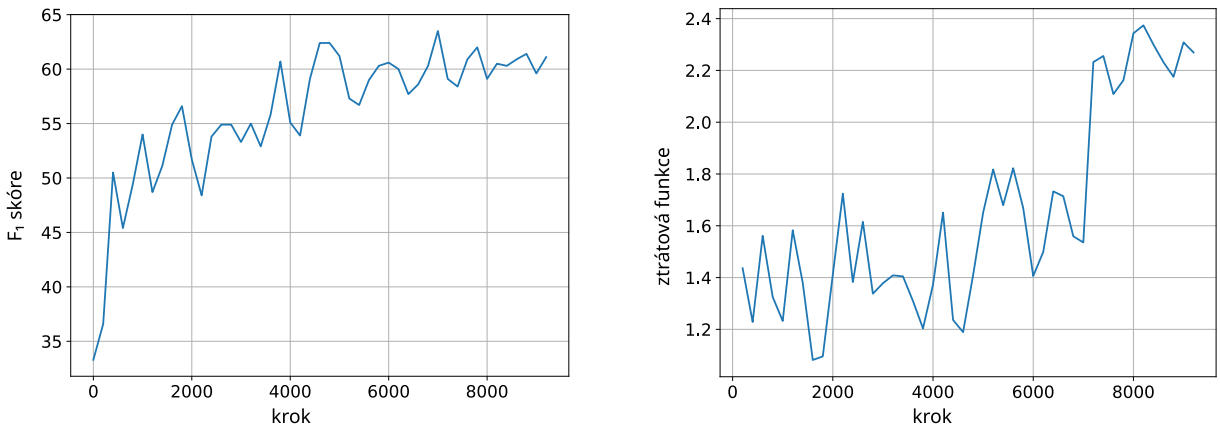
3. Poměr zahřívání (warmup ratio): 0 → 0,1. Důvody jsou popsané v sekci 3.5.

4. Velikost dávky (batch size): 8 → 40. Velikost VRAM použité grafické karty umožnila pouze skutečnou velikost dávky 10, takže musela být simulována nastavením počtu kroků akumulace gradientu (gradient accumulation steps) na 4. Malé dávky vedly k mnohem delším dobám trvání tréninku a větší dávky učinily křivku učení stabilnější, ale příliš vysoké hodnoty snižovaly maximální přesnost, které byl model schopen dosáhnout.

5. Počet epoch: 3 → 4. Zatímco přesnost modelu obvykle vrcholila před koncem třetí epochy, v některých případech bylo nutné trénovat o něco déle.

Ladění ostatních hyperparametrů nepřineslo významná zlepšení.

Jak při hledání hyperparametrů pro POLITICS, tak při ladění DeBERTy large jsme zvolili Article bias prediction jako vynechanou datovou sadu a zaměřili se na dosažení nejvyšší přesnosti při vyhodnocování modelu na ní. Pro tuto volbu existuje několik důvodů: obsahuje třídu středového zabarvení, má texty anotované individuálně (ne podle celkového zabarvení zdrojového média) a pokrývá širokou škálu témat. Interval logování byl nastaven na 200 kroků a metrikou hodnocení, na základě které jsme vybírali nejlepší checkpointy modelu, bylo F_1 skóre. Mezi checkpointy jsme také prozkoumali a zvážili matice záměn. Obrázek 3.4 ukazuje příklad průběhu tréninku.



Obrázek 3.4: Průběh F_1 skóre a ztrátové funkce na validační sadě během trénování modelu DeBERTa V3 large pro benchmark vynechávající jednu datovou sadu s 10 000 příklady z každé datové sady, přičemž byla vynechána a použita pro validaci datová sada Article bias prediction. F_1 skóre dosahuje vrcholu kolem 7 000. kroku tréninku – ve stejném okamžiku, kdy ztrátová funkce naposledy klesá, než začne výrazně růst, což indikuje přetrénování.

S vybranými optimalizovanými hyperparametry pro POLITICS a DeBERTu large jsme provedli celý benchmark znovu, přičemž jsme opět pokaždé vynechali jinou datovou sadu.

3.6 Trénování vlastního modelu

Po optimalizaci tréninku pro přesnost na neznámém typu dat v benchmarku s vynechávám jedné datové sady (metodologií popsanou v sekci 3.5.2) jsme přistoupili k trénování modelu na všech dostupných datových sadách. Použili jsme naprosto stejné nastavení, pouze tentokrát bez vynechání jakékoli sady – s výjimkou Webis bias flipper 18 a Webis news bias 20 (které nebyly zahrnuty ani do zmíněného benchmarku z důvodů popsaných v jeho metodologii). Bylo také nutné změnit utváření validační a testovací sady – z každé datové sady bylo odebráno 15 % příkladů jako testovací sada. Jako validační sada byl použit vzorek 100 příkladů odebraných z každé datové sady – v tomto případě je volba pevné velikosti vzorků validační sady lepší, protože je žádoucí, aby byly vyvážené. Jinak by totiž výběr nejlepšího checkpointu modelu byl více ovlivněn většími datovými sadami.

4 Výsledky

4.1 Průnik datových sad

U datových sad zaměřených na političnost nebyly zjištěny žádné významné průniky.

Z datových sad politického zabarvení se významně překrývají pouze Webis bias flipper 18 a Webis news bias 20 – konkrétně 5 132 instancemi (což odpovídá 80,2 % a 66,5 % celkového počtu záznamů v těchto sadách). Obě se také výrazně překrývají s Article bias prediction – nalezeno bylo 2 716 a 3 824 shod. To znamená, že modely (jak existující, tak naše nové) trénované na jedné z těchto sad budou mít uměle zvýšenou výkonnost i na zbylých dvou – nejen proto, že obsahují stejná témata, ale zejména pokud by se přímo některé články obsažené v trénovacím vzorku jedné datové sady dostaly z druhých dvou sad také do testovacího vzorku modelu. Tomuto jevu jsme se vyhnuli v benchmarku s vynecháním jedné datové sady (3.5.2).

Kompletní výsledné tabulky průniků datasetů jsou zaznamenány v příloze B.1.

4.2 Test existujících modelů

4.2.1 Politické zabarvení

Výsledná F_1 skóre jsou zaznamenána v tabulce 4.1. Nejúspěšnějším klasifikátorem je Political bias prediction AllSides DeBERTa, který dosahuje nejvyšších hodnot na typu textů jak z jeho trénování, tak mimo něj. Tento model proto považujeme za state-of-the-art mezi aktuálně dostupnými modely. DistilBERT political tweets má sice mírně vyšší průměrné skóre, avšak vzhledem k tomu, že nepodporuje třídu středového zabarvení, má statistickou výhodu při výběru 1 ze 2 tříd (šance 50 %) namísto 3 (33 %).

	Political bias BERT	Political bias prediction AllSides DeBERTa	DistilBERT-PoliticalBias	DistilBERT political fine-tune	Political DEBATE large v1.0	BERT political bias fine-tune	Political ideologies RoBERTa fine-tuned	DeBERTa political classification	DistilBERT political tweets
podporuje střední třídu	✓	✓	✓	✓	✓	✗	✗	✗	✗
Article bias prediction	57	65	17	26	51	35	43	61	57
BIGNEWSBLN	37	54	18	24	42	40	46	62	57
CommonCrawl news articles	46	68	18	28	54	40	44	65	56
Dem., rep. party platform topics	50	60	42	42	62	38	58	57	62
GPT-4 political bias	22	15	82	39	57	26	81	46	73
GPT-4 political ideologies	38	75	54	63	82	34	99	63	81
Media political stance	43	73	34	43	56	41	51	68	61
Political podcasts	26	80	34	46	69	50	53	74	71
Political tweets	41	57	42	55	57	47	58	46	74
Qbias	29	46	23	21	37	47	47	55	53
Webis bias flipper 18	38	59	16	24	42	37	47	58	53
Webis news bias 20	48	75	16	21	39	39	47	59	56
průměr na trénovaných datových sadách	48	66	82	–	60	–	99	–	74
průměr na neznámých datových sadách	37	**59**	29	–	53	–	52	–	62
celkový průměr	40	**61**	33	36	54	40	56	60	63

Tabulka 4.1: Průměrná F_1 skóre existujících modelů pro klasifikaci politického zabarvení na všech dostupných datových sadách, vzorkovaných na 1 000 záznamů z každé. Aby nedošlo k chybné interpretaci, modely s různým počtem podporovaných tříd by měly být porovnávány odděleně – zcela náhodně hádající model vybírající mezi dvěma třídami by statisticky dosáhl přesnosti kolem 50 %, ale v případě tří tříd by tato šance byla 33 %. Pozorujeme, že všechny modely dosahují konzistentně a výrazně lepších výsledků na datových sadách, na nichž byly trénovány (zaznamenáno také v tabulce 3.3).

4.2.2 Političnost

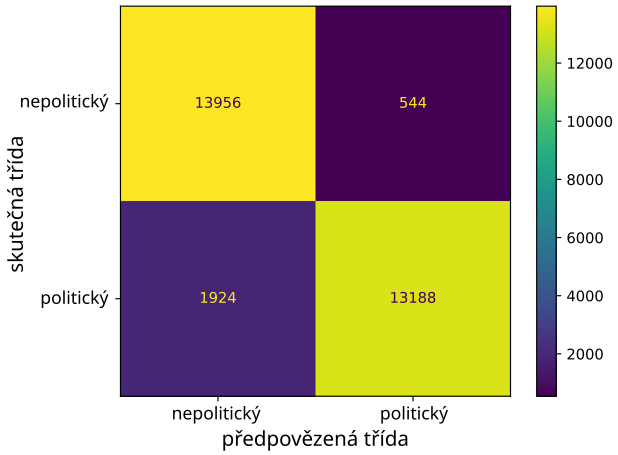
Nejdůležitější výsledky evaluace – na agregované datové sadě (popsané v sekci 3.4.2) – jsou zaznamenány v tabulce 4.2. Kompletní výslednou tabulku každého modelu vyhodnoceného na každé datové sadě zvlášť se nachází v příloze B.2.

Political DEBATE dosahuje nejlepších výsledků a je dostatečně přesný pro reálné aplikace. Jeho matice záměn na obrázku 4.1 ukazuje větší počet falešně negativních než falešně pozitivních případů, což naznačuje, že model má větší sklon klasifikovat texty, které jsou ve skutečnosti o politice, jako nepolitické než naopak. Toto je třeba zvážit při reálné implementaci, např. jako síto před klasifikátorem politického zabarvení – zda je žádoucí striktněji filtrovat nepolitické texty, nebo se snažit zpracovat co největší množství textů,

i když některé z nich mohou být potenciálně nepolitické, a tedy irelevantní. Toto ladění by mohlo být implementováno úpravou prahu pro skóre jistoty modelu – jak jsme učinili v sekci 3.1.1. Tento konkrétní model však nemusí být pro tento přístup vhodný v aplikaci mimo vědecký výzkum, jelikož distribuce jeho skóre jistoty je extrémně nevyvážená – v 90 % případů skóre přesahuje 0,99 (jinými slovy si je téměř vždy naprosto jistý).

model	přesnost	F_1 skóre
Classifier main subject politics	69,5	66,9
Political DEBATE large	**90,4**	**90,4**
Topic politics	76,3	75,1

Tabulka 4.2: Výsledky evaluace existujících modelů na politočnost na agregované datové sadě sestávající ze vzorků (1 000 záznamů) všech dostupných datových sad (jak pro politočnost, tak pro politické zabarvení).



Obrázek 4.1: Výsledná matice záměn modelu Political DEBATE large klasifikujícího podle politočnosti, změřená na agregované datové sadě sestávající ze vzorků (1 000 záznamů) všech dostupných datových sad (jak pro politočnost, tak pro politické zabarvení).

4.3 Benchmarky datových sad

4.3.1 Ponechání jedné

Tabulka 4.3 zobrazuje výsledná F_1 skóre. Podobně jako u existujících modelů – trénovaných obdobným způsobem (pouze na jedné datové sadě) – i tyto benchmarkové prototypní modely selhávají na neznámých typech dat.

PONECHANÁ DATOVÁ SADA ↓	BERT		RoBERTa		DeBERTa		POLITICS	
TYP TESTOVACÍHO VZORKU →	známý	neznámý	známý	neznámý	známý	neznámý	známý	neznámý
SE STŘEDNÍ TŘÍDOU								
Article bias prediction	64,8	39,6	58,9	33,4	70,8	37,4	81,3	51
BIGNEWSBLN	73,3	35,9	77,9	32,6	80,9	40,6	89,7	38,4
CommonCrawl news articles	74,4	33,2	78,7	37,2	78,9	35,5	82,2	49
GPT-4 political bias	90,1	35,4	91,2	37,8	92,4	36,6	91,3	37
Qbias	44,3	38,4	53,6	41,5	38,4	34,7	51,1	48,3
Webis bias flipper 18	77,5	41,6	82,1	33,6	82,3	45,4	89	50,8
Webis news bias 20	74,4	40,9	80	36,9	70,6	42,1	83,3	50,3
průměr	71,3	37,9	74,6	36,1	73,5	38,9	81,1	46,4
BEZ STŘEDNÍ TŘÍDY								
Dem., rep. party platform topics	76,6	45	75,4	53,3	57,8	38,7	76,8	60
GPT-4 political ideologies	99	53,8	99,2	53,5	99,2	56,5	98,7	57,1
Media political stance	83,7	49,2	87,9	50,7	82,1	55,3	90,3	67,5
Political podcasts	99,5	43,4	99,6	44,5	99,8	45	99,5	62,5
Political tweets	71,6	54,6	73,7	57,7	65,8	51,4	74,8	66,7
průměr	86,1	49,2	87,2	51,9	80,9	49,4	88	62,8

Tabulka 4.3: Výsledná F_1 skóre při klasifikaci politického zabarvení po evaluaci základních verzí modelů BERT (cased), RoBERTa, DeBERTa V3 a POLITICS (velikost testovacího vzorku: 15 %; velikost trénovacího vzorku: 2 000 záznamů z každé datové sady) natrénovaných vždy na jedné (ponechané) z dostupných datových sad (10) – jak je popsáno v sekci 3.5.1. Hodnoty ve sloupcích známého testovacího vzorku znamenají skóre na oné jediné ponechané datové sadě, na které model byl trénován. Sloupce neznámých testovacích vzorků ukazují průměrné skóre na všech ostatních datových sadách. Aby nedošlo k chybné interpretaci výsledků, modely s různým počtem podporovaných tříd by měly být porovnávány odděleně – zcela náhodně předpovídající model vybírající mezi dvěma třídami by statisticky dosáhl přesnosti kolem 50 %, ale v případě tří tříd by tato základní úroveň byla 33 %. Pozorujeme, že všechny klasifikátory dosahují konzistentně a výrazně lepších výsledků na datových sadách, na kterých byly trénovány.

4.3.2 Vynechání jedné

Tyto výsledky opět potvrzují fenomén poklesu úspěšnosti modelů na neznámých datech.

Skóre mimo natrénované datové sady jsou obecně vyšší, čím novější je architektura modelu. Vyjímkou je však POLITICS, který je jednoznačně nejlepší, přestože je založen na RoBERTě.

Politické zabarvení

F_1 skóre jsou zobrazena v tabulce 4.4. Při evaluaci na vynechané datové sadě, která obsahuje všechny tři třídy zabarvení (nechybí střední třída), dosahují modely obecně horších výsledků – to je očekávané jak statisticky, tak i proto, že klasifikace textů se středovým zabarvením může být obtížnější než u krajních tříd.

Ve srovnání s modelem POLITICS trénovaným na pouhých 2 000 záznamech z každé

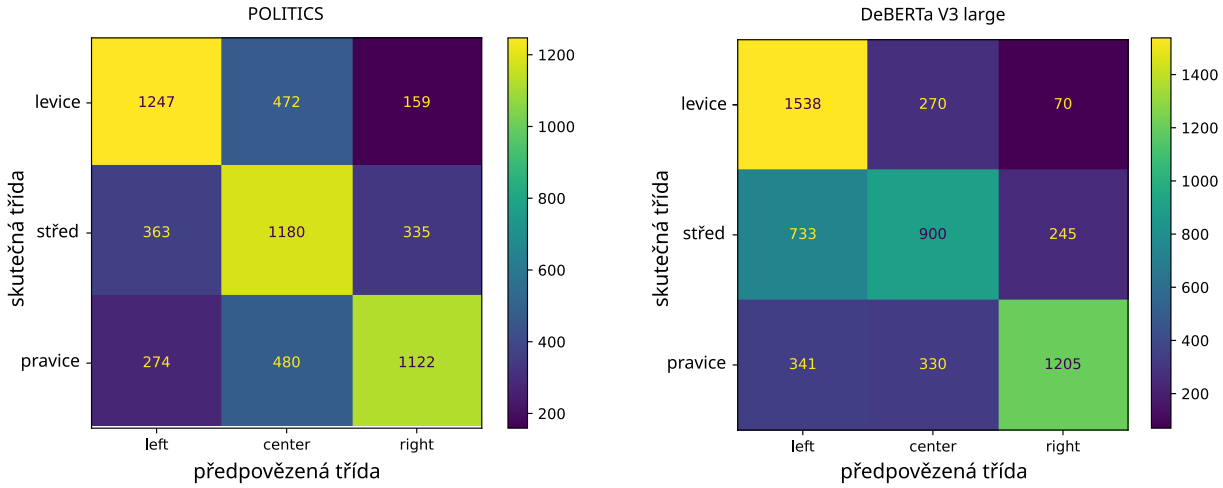
datové sady dosahuje DeBERTa large s ručně upravenými hyperparametry ve skutečnosti o 7,8 % nižšího skóre na datech z natrénovaných sad, ale dosahuje nejlepších výsledků ze všech modelů na neznámém typu dat. Tento trénink tedy dosáhl svého stanoveného cíle.

VYNECHANÁ DATOVÁ SADA ↓	BERT		RoBERTa		DeBERTa		POLITICS	
TYP TESTOVACÍHO VZORKU →	známý	neznámý	známý	neznámý	známý	neznámý	známý	neznámý
Article bias prediction	77,7	47	80,1	48,4	78,6	49,3	83,4	61,4
BIGNEWSBLN	77,5	39,3	80,7	38,1	77,1	42	83,1	56,8
CommonCrawl news articles	77	38,7	80,1	51	77,4	50	83,7	51,6
Dem., rep. party platform topics	76,6	59,6	80,3	62,4	78,2	63,1	83,8	66,4
GPT-4 political bias	74,4	42	78,7	45,8	76,3	37,2	82,7	44
GPT-4 political ideologies	74	73,5	77,9	83,2	73	79	81,1	86,3
Media political stance	73,1	54,4	78,3	57,3	76,7	64,9	82,3	66,5
Political podcasts	73,4	61,8	77	55,9	48,5	55,1	81,2	76,6
Political tweets	76,6	41	80,6	40,9	76,7	43,7	84,4	55,4
Qbias	78,9	36,9	83,4	38,3	80,5	41,6	86,2	42,6
průměr	75,9	49,4	79,7	52,1	74,3	52,6	83,2	60,8

VYNECHANÁ DATOVÁ SADA ↓	POLITICS		DeBERTa large	
TYP TESTOVACÍHO VZORKU →	známý	neznámý	známý	neznámý
Article bias prediction	84,5	63,1	78	64,3
BIGNEWSBLN	84,4	60,6	76,4	52
CommonCrawl news articles	84,2	58	78,6	45,9
Dem., rep. party platform topics	85,9	70	72,9	75
GPT-4 political bias	84,1	45,9	76	46
GPT-4 political ideologies	84	86,5	62,3	96,2
Media political stance	84,3	72,7	78,2	77,9
Political podcasts	83,1	79,1	74,6	91,2
Political tweets	86,1	62	75,8	71,3
Qbias	88,6	42,9	80,9	42,3
průměr	84,9	64,1	75,4	66,2

Tabulka 4.4: Výsledná F_1 skóre klasifikace politického zabarvení po evaluaci základních verzí modelů BERT (cased), RoBERTa, DeBERTa V3 a POLITICS (velikost testovacího vzorku: 15 %; velikost trénovacího vzorku: 2 000 z každé datové sady) a poté POLITICS a DeBERTa V3 large s optimalizovanými hyperparametry (velikost testovacího vzorku: 15 %; velikost trénovacího vzorku: 10 000 z každé datové sady) natrénovaných na všech dostupných datových sadách (10) kromě jedné (vynechané), přičemž pokaždé byla vynechána jiná datová sada (řádky tabulky) – jak je popsáno v sekci 3.5.2. Hodnoty ve sloupcích známého testovacího vzorku znamenají průměrné skóre na všech datových sadách, na kterých byl model trénován. Sloupce neznámého testovacího vzorku zobrazují skóre na oné jediné vynechané datové sadě. Pozorujeme, že všechny klasifikátory dosahují konzistentně a výrazně lepších výsledků na datových sadách, na kterých byly trénovány.

Uvádíme také příklady matic záměn obou optimalizovaných modelů na obrázku 4.2.



Obrázek 4.2: Výsledná matice záměn modelů POLITICS a DeBERTa V3 large s optimalizovanými hyperparametry (velikost testovací sady: 15 %; velikost trénovací sady: 10 000 z každé datové sady) natrénovaných na všech dostupných datových sadách (10) kromě jedné (vynechané) – v tomto případě Article bias prediction – jak je popsáno v sekci 3.5.2. Tyto matice ukazují predikce modelů na testovací sadě sestávající pouze z vynechané datové sady.

Političnost

Tabulka 4.5 zachycuje F_1 skóre. Neobsahuje výsledky na jednotlivých datových sadách (28), protože by jí to učinilo velmi rozsáhlou. Tyto výsledky uvádíme v příloze B.3. Nicméně i tato zjednodušená tabulka postačuje k demonstraci, že modely opět dosahují vyššího skóre na známých datových sadách. Ačkoli je tento efekt konzistentní, není tak drastický jako u úlohy politického zabarvení. Co jej dále činí méně znepokojivým je skutečnost, že skóre jsou celkově velmi vysoká, dokonce i na neviděných typech textů, což rovněž naznačuje, že tato úloha je výrazně jednodušší než klasifikace politického zabarvení.

TYP TESTOVACÍHO VZORKU →	známý	neznámý
BERT	97,6	93,3
RoBERTa	95,9	91
DeBERTa	97,8	92,9
POLITICS	97,8	92,5

Tabulka 4.5: Výsledná F_1 skóre klasifikace političnosti po evaluaci základních verzí modelů BERT (cased), RoBERTa, DeBERTa V3 a POLITICS (velikost testovací sady: 15 %; velikost trénovací sady: 1 000 záznamů z každé datové sady) natrénovaných na všech dostupných datových sadách (28) kromě jedné (vynechané) – jak je popsáno v sekci 3.5.2. Hodnoty ve sloupci známého testovacího vzorku znamenají průměrné skóre na všech datových sadách, na kterých byl model trénován. Sloupec neznámeho testovacího vzorku zobrazuje skóre na oné jediné vynechané datové sadě. Pozorujeme, že všechny klasifikátory dosahují konzistentně lepších výsledků na datových sadách, na kterých byly trénovány.

4.4 Trénování vlastního modelu

Naše modely trénované na všech dostupných datových sadách s využitím metodologie optimalizované pro klasifikaci neznámých typů dat (popsané v sekci 3.5.2) dosáhly F_1 skóre zaznamenaných v tabulce 4.6. Tyto modely stanovují novou state-of-the-art úspěšnost napříč všemi vyhodnocenými datovými sadami.

	POLITICS		DeBERTa V3 large	
	přesnost	F_1 skóre	přesnost	F_1 skóre
Article bias prediction	84,7	84,6	89	89
BIGNEWSBLN	89,2	89,2	88,6	88,6
CommonCrawl news articles	85,2	85,1	88,9	88,9
Dem., rep. party platform topics	77,6	77,3	85,5	85,6
GPT-4 political bias	84,8	84,2	87	86,9
GPT-4 political ideologies	97,5	97,5	99,6	99,6
Media political stance	90,8	91,1	91,6	93,1
Political podcasts	99,4	99,5	99,8	99,8
Political tweets	76,4	76,2	82,1	82,1
Qbias	51,6	51	58	57,9
průměr	83,7	83,6	87	87,2

Tabulka 4.6: Výsledky evaluace našich klasifikátorů politického zabarvení trénovaných na všech dostupných datových sadách (10) s optimalizovanými hyperparametry (jak je popsáno v sekci 3.5.2).

5 Diskuze

Naší výzkumnou hypotézou je, že modely klasifikující politické zabarvení dosahují dobrých výsledků na datových sadách, na kterých byly trénovány, ale vykazují výrazně sníženou přesnost, když jsou aplikovány na texty jim neznámých stylů nebo texty napsané v jiném časovém období. To je v souladu s existující literaturou: Cohen a Ruths (2021) ukázali, že klasifikátory dosahují vysoké přesnosti (>90 %) na datových sadách tweetů napsaných politiky a politicky aktivními uživateli, ale jejich úspěšnost prudce klesá na pouhých 65 %, jakmile jsou aplikovány na běžné uživatele. Podobně Yan, Lavoie a Das (2017) změřili, že nízká je i schopnost modelu generalizovat mezi různými doménami – napříč datovými sadami z kongresových záznamů, mediálních webů a wiki stránek, a to navzdory přesvědčivým výsledkům křížové validace v rámci jedné datové sady. Jejich práce odhalila, že obtíž nespočívá pouze v technických omezeních modelů, ale v zásadních rozdílech v povaze dat napříč doménami.

Naše evaluace existujících modelů tato tvrzení podporuje: Přes dosažení působivých výsledků na známých datových sadách jejich přesnost trpí na těch s neznámým typem textů. Ačkoli je to za do jisté míry očekávatelné, pokles přesnosti není jen konzistentní, nýbrž i významný (v nejnižším případě 7 %).

Metodologie našeho benchmarku s ponecháním jedné datové sady je velmi blízká způsobu, jakým existující práce přistupovaly k této úloze: trénování a vyhodnocování modelů na jediné datové sadě, a tedy na jediném typu textů. Naše výsledky ukázaly, že je relativně snadné dosáhnout slušné přesnosti v rámci trénovací datové sady, dokonce i s malým trénovacím vzorkem. Jakmile jsou však tyto modely vystaveny neznámým datovým sadám, drasticky selhávají.

Benchmark s vynecháním jedné datové sady přinesl velmi podobné výsledky, opět podporující naši hypotézu, navzdory tréninku na mnohem větší a rozmanitější sadě. Při zkoumání modelů trénovaných na zvýšené velikosti vzorku (10 000) a s upravenými hyperparametry pozorujeme zlepšení: Ačkoli dochází k poklesu úspěšnosti na trénovaných datových sadách, dosahují vyšších skóre mimo trénovací distribuci. Tento kompromis považujeme za úspěch – snižuje všudypřítomný pokles přesnosti na neviděných datech, kterým trpí všechny existující i naše předchozí prototypní modely. Výsledky ukazují, že tato konfigurace je použitelná a vhodná pro trénink na všech dostupných datech, což by mělo vést k modelu dobře přenosnému na budoucí nové neviděné texty.

Natrénovali jsme dva takové modely (jeden založený na POLITICS a druhý na DeBERTě V3 large) pro klasifikaci politického zabarvení. Jejich úspěšnosti představuje nový state-of-the-art na všech shromážděných datových sadách. Míra, do jaké jsou naše nové

modely trénované na všech aktuálně dostupných datových sadách schopny generalizovat napříč doménami, však může být odhalena pouze jejich vystavením datovým sadám, které teprve budou publikovány, nebo manuálně navrženým testovacím příkladům.

Obecně platí, že čím novější je architektura předtrénovaného jazykového modelu, tím lepší je výkon modelu doladěného (fine-tuning): DeBERTa V3 (2021) překonává RoBERTu (2019), která překonává BERTa (2018). Existuje však výjimka: POLITICS, založený na RoBERTě, dosahuje nejlepších výsledků ve všech provedených měřeních. Tyto výsledky prokazují, že pokročilý trénink (continued training) pro specifické textové domény je efektivní a potvrzují, že úsilí autorů (Liu et al. 2022) se vyplatilo.

Považujeme za důležité zmínit také přesnost generativních LLM se stovkami miliard parametrů, neboť se do jisté míry staly univerzálním přístupem k problémům strojového učení, a tedy i kandidáty na řešení úloh této studie. Ačkoli prokázaly působivé schopnosti v úlohách klasifikace textu, jejich adopce jak ve vědeckém výzkumu, tak v praktických aplikacích přináší významné nevýhody. Tyto masivní modely vyžadují značné výpočetní zdroje, což je činí buď drahými, anebo pomalými. Z akademického hlediska je snad nejvíce znepokojující především uzavřený charakter mnoha state-of-the-art LLM, což podkopává reprodukovatelnost dosažených výsledků a možnost zkoumat interní chování modelu. Jak Burnham et al. (2024) demonstrují se svými modely Political DEBATE (které v našich testech dosahují nejlepších výsledků pro úlohu političnosti), doménově specializované menší modely dokáží překonat proprietární LLM v úlohách klasifikace textu, přičemž jsou řádově efektivnější. Z těchto důvodů jsme se rozhodli prozkoumat doladění těchto, tisíckrát menších, jazykových modelů.

6 Omezení

Ačkoli se tato práce snaží obsáhle pokrýt svá témata, má určitá omezení.

Za prvé, klasifikace politického zabarvení je zjednodušená. Tři třídy nemusí dostatečně detailně modelovat skutečné spektrum. Zvýšení počtu tříd je jedním ze způsobů, jak zvýšit rozlišení. Případně použití regrese místo klasifikace by mohlo poskytnout jemnější reprezentaci umístěním textů na spojité číselné škále, která by odrážela různé stupně politického zabarvení. Dále je důležité vzít v úvahu, že některé texty mohou současně podporovat jak levicové, tak pravicové perspektivy, což naznačuje potřebu „smíšené“ třídy nebo metody pro predikci míry, do jaké se text shoduje s každou stranou, podobně jako u úlohy detekce postojů (stance detection) – míra pro / proti.

Za druhé, naše definice političnosti se zaměřuje pouze na to, zda je hlavním tématem textu politika, a nebere v úvahu, zda vyjadřuje politický názor. Například texty jako encyklopedická či slovníková hesla mohou pojednávat o politických tématech, aniž by projevovaly jakoukoliv zaujatost. Tyto texty by v ideálním případě měly být odfiltrovány a neměly by být povoleny jako vstupy do klasifikátoru politického zabarvení, ale s naší zjednodušenou definicí by povoleny byly. Jemnější přístup by vyžadoval data s podrobnými štítky, která v současné době nejsou k dispozici.

Za třetí, tato studie je omezena svým zaměřením na politický kontext USA, kde třída levicového zabarvení odpovídá převážně liberálním a demokratickým názorům a třída pravicového zabarvení konzervativním a republikánským názorům. Tento rámec se nepřenáší dobře do širších politologických kontextů nebo rozlišení na více osách, jako je konzervativní vs. progresivní a autoritářské vs. liberální.

Za čtvrté, naše práce nezkoumá ani data v jiných jazycích než angličtině (kterých je nedostatek), ani vícejazyčné modely.

Za páté, přesnost NLI klasifikátoru – Political DEBATE – je potenciálně omezen jednoduchostí použitých hypotéz. Neexperimentovali jsme s alternativními hypotézami pro klasifikaci bez trénovacích příkladů (zero-shot) ani jsme netestovali klasifikaci s uvedením příkladů (few-shot), což by mohlo potenciálně zlepšit výsledky – zejména u politického zabarvení. Místo toho jsme se zaměřili na učení s učitelem (supervised learning).

Za šesté, neprovedli jsme extenzivní testování našich nově natrénovaných modelů. To je nezbytné k ověření, že se nenaučily i nezamýšlené textové rysy, jako jsou specifické charakteristiky mediálních zdrojů nebo stylistické vzorce – na což poukazují (Baly et al. 2020) – kromě samotného politické zabarvení. Vyhodnocení na souboru ručně vytvořených příkladů nebo skutečných textů napsaných v budoucnosti by mohlo potenciálně odhalit nedostatky.

7 Budoucí práce

Náš výzkum odhaluje několik slibných směrů pro budoucí práci.

Za prvé, hlubší manuální analýza nesprávných predikcí našich modelů by mohla přinést cenné poznatky o potenciálních vzorcích chybného chování. Takové šetření by pomohlo identifikovat, zda určitá politická témata, styly psaní nebo konkrétní politické postoje konzistentně představují výzvu pro klasifikátory, což by naznačilo směr pro cílená vylepšování modelů.

V návaznosti na náš současný systém tří tříd politického zabarvení by budoucí práce mohla implementovat jemnější politické spektrum s pěti nebo více úrovněmi. To by přesněji zachytilo kontinuum od krajní levice po krajní pravici a lépe odráželo realitu politického diskurzu. Kromě toho by bylo vhodné zavedení „smíšené“ třídy pro texty, které současně vykazují podporu pro více politických pozic. Dalším slibným přístupem by mohl být nějaký způsob zakódování středové třídy jako sémanticky umístěné mezi levicí a pravicí, možná s využitím spojité číselné škály spíše než diskrétních kategorií.

Místo oddělených klasifikátorů pro političnost a politické zabarvení by mohl být prozkoumán jednotný přístup přidáním čtvrté „nepolitické“ třídy do modelu politického zabarvení. To by zjednodušilo klasifikační řetězec a potenciálně umožnilo modelu lépe porozumět vztahu mezi politickým a nepolitickým obsahem.

S přístupem k výkonnějším výpočetním zdrojům, než byly k dispozici pro tento výzkum (podrobnosti v příloze C), by bylo možné provést komplexní optimalizaci hyperparametrů na větších modelech, jako je DeBERTa V3 large. Náš výzkum ukazuje, že větší modely dosahují lepších výsledků na datech mimo trénovací distribuci, a optimalizace jejich hyperparametrů specificky pro klasifikaci politických textů by mohla přinést významná zlepšení.

Pro řešení datových omezení by tvorba dalších syntetických politických textů pomocí velkých generativních jazykových modelů mohla pomoci rozšířit a vyvážit naše datové sady. Podobně by využití LLM k validaci nebo upřesnění existujících štítků mohlo zlepšit kvalitu anotace, zejména u hraničních případů.

Dalším důležitým rozšířením by byl vývoj vícejazyčných modelů schopných klasifikovat politický obsah napříč jazyky. To by vyžadovalo také sběr odpovídajících dat.

V neposlední řadě představuje zásadní směr pro budoucí práci na textových klasifikačních modelech zlepšení jejich interpretovatelnosti a vysvětlitelnosti (explainability), jak apelují metastudie (Doan a Gulla 2022; Németh 2022). Vývoj metod, které by zvýrazňovaly, jaké části textu nejvíce ovlivnily jejich predikce, by učinil tyto nástroje transparentnějšími a užitečnějšími jak pro výzkum, tak pro aplikace.

8 Závěr

V této práci jsme se zabývali problémem automatické klasifikace textu podle politického zabarvení a političnosti. Shromáždili a vybrali jsme 12 datových sad s anotacemi pro politické zabarvení. Pro političnost jsme zkombinovali 18 datových sad a sjednotili jejich různorodé štítky pro binární klasifikaci. Náš komplexní přístup zahrnoval vyhodnocení existujících modelů a trénování nových s vylepšenými schopnostmi generalizace. Rozsáhlé testování prostřednictvím metodologií s vynecháním a ponecháním jednotlivých datových sad poskytlo cenné poznatky o přenositelnosti těchto modelů napříč různými textovými doménami. Ukázali jsme, že existující NLI klasifikátor Political DEBATE dosahuje vynikajících výsledků v úloze klasifikace političnosti, zatímco naše nově natrénované modely založené na POLITICS a DeBERTě large stanovují novou state-of-the-art úspěšnost na klasifikaci politického zabarvení.

Nepříjemným, ale důležitým zjištěním je, že všechny testované modely (jak existující, tak naše nově natrénované) dosahují konzistentně horších výsledků na datových sadách a typech textů, na kterých nebyly trénovány. Tato zjištění jsou v souladu s naší navrhovanou hypotézou i s existující literaturou. Podnikli jsme kroky k řešení a překonání tohoto jevu, což vedlo ke zlepšení úspěšnosti na těchto textech. Pokles přesnosti je však stále přítomen. Docházíme k závěru, že naučit model univerzálně klasifikovat texty podle politického zabarvení – na širším spektru témat a na jiném charakteru textu, než jaké jsou přítomny v trénovací sadě – je skutečně obtížná úloha.

Implikace tohoto výzkumu přesahují rámec akademického zájmu a mají praktické využití při analýze médií, automatickém kurátorství obsahu a podpoře vyvážené konzumace informací. Zjištění zdůrazňují důležitost rozmanitosti trénovacích dat a potřebu pečlivého vyhodnocení na příkladech z neviděných distribucí. Naše práce poskytuje metodologické poznatky pro výzkumníky a odborníky analyzující politicky zabarvený obsah v polarizované informační krajině.

Bibliografie

Akiba, Takuya et al. (2019). "Optuna: A Next-generation Hyperparameter Optimization Framework". In: *CoRR* abs/1907.10902. arXiv: 1907.10902. URL: http://arxiv.org/abs/1907.10902.

AllSides (2024). *AllSides Media Bias Chart*. URL: https://www.allsides.com/media-bias/media-bias-chart.

amrrs (2022). *Medium post titles*. URL: https://www.kaggle.com/datasets/nulldata/medium-post-titles.

Antypas, Dimosthenis et al. (říj. 2022). "Twitter Topic Classification". In: *Proceedings of the 29th International Conference on Computational Linguistics*. Ed. Nicoletta Calzolari et al. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, s. 3386–3400. URL: https://aclanthology.org/2022.coling-1.299/.

Ao, Xiang et al. (srp. 2021). "PENS: A Dataset and Generic Framework for Personalized News Headline Generation". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. Chengqing Zong et al. Online: Association for Computational Linguistics, s. 82–92. DOI: 10.18653/v1/2021.acl-long.7. URL: https://aclanthology.org/2021.acl-long.7/.

Baly, Ramy et al. (lis. 2020). "We Can Detect Your Bias: Predicting the Political Ideology of News Articles". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. Bonnie Webber et al. Online: Association for Computational Linguistics, s. 4982–4991. DOI: 10.18653/v1/2020.emnlp-main.404. URL: https://aclanthology.org/2020.emnlp-main.404/.

Baumgartner, Jason et al. (2020). "The pushshift reddit dataset". In: *Proceedings of the international AAAI conference on web and social media*. Sv. 14, s. 830–839.

Bian, Shengtao (2022). *Recipe*. URL: https://huggingface.co/datasets/Shengtao/recipe.

Burnham, Michael (2024). "Stance detection: a practical guide to classifying political beliefs in text". In: *Political Science Research and Methods*, s. 1–18. DOI: 10.1017/psrm.2024.35.

Burnham, Michael et al. (2024). "Political DEBATE: Efficient Zero-shot and Few-shot Classifiers for Political Text". In: *arXiv preprint arXiv:2409.02078*.

Chen, Wei-Fan et al. (lis. 2018). "Learning to Flip the Bias of News Headlines". In: *Proceedings of the 11th International Conference on Natural Language Generation*. Ed. Emiel Krahmer, Albert Gatt a Martijn Goudbeek. Tilburg University, The Netherlands: As-

sociation for Computational Linguistics, s. 79–88. DOI: 10.18653/v1/W18-6509. URL: https://aclanthology.org/W18-6509/.

Chen, Wei-Fan et al. (lis. 2020). "Analyzing Political Bias and Unfairness in News Articles at Different Levels of Granularity". In: *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*. Ed. David Bamman et al. Online: Association for Computational Linguistics, s. 149–154. DOI: 10.18653/v1/2020.nlpcss-1.16. URL: https://aclanthology.org/2020.nlpcss-1.16/.

Chen, Yulong et al. (srp. 2021). "DialogSum: A Real-Life Scenario Dialogue Summarization Dataset". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Ed. Chengqing Zong et al. Online: Association for Computational Linguistics, s. 5062–5074. DOI: 10.18653/v1/2021.findings-acl.449. URL: https://aclanthology.org/2021.findings-acl.449/.

Cohen, Raviv a Derek Ruths (2021). "Classifying Political Orientation on Twitter: It's Not Easy!" In: *Proceedings of the International AAAI Conference on Web and Social Media* 7.1, s. 91–99. DOI: 10.1609/icwsm.v7i1.14434. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14434.

Demidov, Dmitry (2023). *Political Bias of News Content: Classification based on Individual Articles and Media*. URL: https://www.researchgate.net/publication/377074277_Political_Bias_of_News_Content_Classification_based_on_Individual_Articles_and_Media.

Devlin, Jacob et al. (čvn. 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. Jill Burstein, Christy Doran a Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, s. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423/.

Doan, Tu My a Jon Atle Gulla (2022). "A Survey on Political Viewpoints Identification". In: *Online Social Networks and Media* 30. ISSN: 2468-6964. DOI: https://doi.org/10.1016/j.osnem.2022.100208. URL: https://www.sciencedirect.com/science/article/pii/S246869642200012X.

España-Bonet, Cristina (pros. 2023). "Multilingual Coarse Political Stance Classification of Media. The Editorial Line of a ChatGPT and Bard Newspaper". In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. Houda Bouamor, Juan Pino a Kalika Bali. Singapore: Association for Computational Linguistics, s. 11757–11777. DOI: 10.18653/v1/2023.findings-emnlp.787. URL: https://aclanthology.org/2023.findings-emnlp.787/.

Gangula, Rama Rohit Reddy, Suma Reddy Duggenpudi a Radhika Mamidi (srp. 2019). "Detecting Political Bias in News Articles Using Headline Attention". In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Ed. Tal Linzen et al. Florence, Italy: Association for Computational Linguis-

tics, s. 77–84. DOI: 10.18653/v1/W19-4809. URL: https://aclanthology.org/W19-4809/.

gptmurdock (2024). *Classifier main subject politics*. URL: https://huggingface.co/gptmurdock/classifier-main_subjects_politics.

Gu, Yupeng et al. (2017). "Ideology Detection for Twitter Users via Link Analysis". In: *Social, Cultural, and Behavioral Modeling*. Ed. Dongwon Lee et al. Cham: Springer International Publishing, s. 262–268. ISBN: 978-3-319-60240-0.

Haak, Fabian a Philipp Schaer (2023). "Qbias - A Dataset on Media Bias in Search Queries and Query Suggestions". In: *Proceedings of the 15th ACM Web Science Conference 2023*. WebSci '23. Austin, TX, USA: Association for Computing Machinery, s. 239–244. ISBN: 9798400700897. DOI: 10.1145/3578503.3583628. URL: https://doi.org/10.1145/3578503.3583628.

Halterman, Andrew (2025). "Synthetically generated text for supervised text analysis". In: *Political Analysis*, s. 1–14. DOI: 10.1017/pan.2024.31.

He, Pengcheng et al. (2021). *DeBERTa: Decoding-enhanced BERT with Disentangled Attention*. arXiv: 2006.03654 [cs.CL]. URL: https://arxiv.org/abs/2006.03654.

Heseltine, Michael a Bernhard Clemm von Hohenberg (2024). "Large language models as a substitute for human experts in annotating political text". In: *Research & Politics* 11.1. DOI: 10.1177/20531680241236239. eprint: https://doi.org/10.1177/20531680241236239. URL: https://doi.org/10.1177/20531680241236239.

Heymans, Adrien (2022). *IMDB movie genres*. URL: https://huggingface.co/datasets/adrienheymans/imdb-movie-genres.

Hou, Yupeng et al. (2024). "Bridging Language and Items for Retrieval and Recommendation". In: *arXiv preprint arXiv:2403.03952*.

Høyland, Bjørn et al. (čvn. 2014). "Predicting Party Affiliations from European Parliament Debates". In: *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. Ed. Cristian Danescu-Niculescu-Mizil et al. Baltimore, MD, USA: Association for Computational Linguistics, s. 56–60. DOI: 10.3115/v1/W14-2516. URL: https://aclanthology.org/W14-2516/.

Iyyer, Mohit et al. (čvn. 2014). "Political Ideology Detection Using Recursive Neural Networks". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. Kristina Toutanova a Hua Wu. Baltimore, Maryland: Association for Computational Linguistics, s. 1113–1122. DOI: 10.3115/v1/P14-1105. URL: https://aclanthology.org/P14-1105/.

Jones, Christopher (2024). *Political Bias Dataset: A Synthetic Dataset for Bias Detection and Reduction*. URL: https://huggingface.co/datasets/cajcodes/political-bias.

Kawintiranon, Kornraphop a Lisa Singh (čvn. 2022). "PoliBERTweet: A Pre-trained Language Model for Analyzing Political Content on Twitter". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Ed. Nicoletta Calzolari et

al. Marseille, France: European Language Resources Association, s. 7360–7367. URL: https://aclanthology.org/2022.lrec-1.801/.

Liu, Yujian et al. (čvc. 2022). "POLITICS: Pretraining with Same-story Article Comparison for Ideology Prediction and Stance Detection". In: *Findings of the Association for Computational Linguistics: NAACL 2022*. Ed. Marine Carpuat, Marie-Catherine de Marneffe a Ivan Vladimir Meza Ruiz. Seattle, United States: Association for Computational Linguistics, s. 1354–1374. DOI: 10.18653/v1/2022.findings-naacl.101. URL: https://aclanthology.org/2022.findings-naacl.101/.

Lorenzoni, Giuliano et al. (2024). *Exploring Variability in Fine-Tuned Models for Text Classification with DistilBERT*. arXiv: 2501.00241 [cs.CL]. URL: https://arxiv.org/abs/2501.00241.

Maas, Andrew L. et al. (čvn. 2011). "Learning Word Vectors for Sentiment Analysis". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Ed. Dekang Lin, Yuji Matsumoto a Rada Mihalcea. Portland, Oregon, USA: Association for Computational Linguistics, s. 142–150. URL: https://aclanthology.org/P11-1015/.

Misra, Rishabh (2022). "News Category Dataset". In: *arXiv preprint arXiv:2209.11429*.

Misra, Rishabh a Jigyasa Grover (led. 2021). *Sculpting Data for ML: The first act of Machine Learning*. Misra, Rishabh a Grover, Jigyasa. ISBN: 9798585463570.

Narayana (2024). *Political Podcasts listing with audio links*. URL: https://www.kaggle.com/datasets/nbandhi/political-podcasts-listing-with-audio-links.

Nayak, Jyoti Shankar (2024a). *GPT-4 political ideologies*. URL: https://huggingface.co/datasets/JyotiNayak/political_ideologies.

— (2024b). *Political ideologies RoBERTa fine-tuned*. URL: https://huggingface.co/JyotiNayak/political_ideologies_detection_roberta_finetuned.

Newhauser, Mary (2022). *DistilBERT political tweets*. URL: https://huggingface.co/m-newhauser/distilbert-political-tweets.

Németh, Renáta (pros. 2022). "A scoping review on the use of natural language processing in research on political polarization: trends and research prospects". In: *Journal of Computational Social Science* 6. DOI: 10.1007/s42001-022-00196-2.

Palmqvist, Oscar (2024). *DeBERTa political classification*. URL: https://huggingface.co/oscpalML/DeBERTa-political-classification.

Pang, Bo a Lillian Lee (čvn. 2005). "Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales". In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. Ed. Kevin Knight, Hwee Tou Ng a Kemal Oflazer. Ann Arbor, Michigan: Association for Computational Linguistics, s. 115–124. DOI: 10.3115/1219840.1219855. URL: https://aclanthology.org/P05-1015/.

Paszke, Adam et al. (2019). "PyTorch: an imperative style, high-performance deep learning library". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc.

Phi, Open (2023). *Textbooks*. URL: https://huggingface.co/datasets/open-phi/textbooks.

Preoţiuc-Pietro, Daniel et al. (čvc. 2017). "Beyond Binary Labels: Political Ideology Prediction of Twitter Users". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. Regina Barzilay a Min-Yen Kan. Vancouver, Canada: Association for Computational Linguistics, s. 729–740. DOI: 10.18653/v1/P17-1068. URL: https://aclanthology.org/P17-1068/.

Research, Bucket (2023). *PoliticalBiasBERT*. URL: https://huggingface.co/bucketresearch/politicalBiasBERT.

Sahitaj, Premtim (2024). *Political bias prediction AllSides DeBERTa*. URL: https://huggingface.co/premsa/political-bias-prediction-allsides-DeBERTa.

Sanh, Victor et al. (2020). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. arXiv: 1910.01108 [cs.CL]. URL: https://arxiv.org/abs/1910.01108.

Shrimali, Harshal (2024). *DistilBERT political fine-tune*. URL: https://huggingface.co/harshal-11/DistillBERT-Political-Finetune.

Silcock, Emily et al. (2024). *Newswire: A Large-Scale Structured Database of a Century of Historical News*. arXiv: 2406.09490 [cs.CL]. URL: https://arxiv.org/abs/2406.09490.

Spliethöver, Maximilian, Maximilian Keiff a Henning Wachsmuth (pros. 2022). "No Word Embedding Model Is Perfect: Evaluating the Representation Accuracy for Social Bias in the Media". In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Ed. Yoav Goldberg, Zornitsa Kozareva a Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, s. 2081–2093. DOI: 10.18653/v1/2022.findings-emnlp.152. URL: https://aclanthology.org/2022.findings-emnlp.152/.

Stefanov, Peter et al. (čvc. 2020). "Predicting the Topical Stance and Political Leaning of Media using Tweets". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. Dan Jurafsky et al. Online: Association for Computational Linguistics, s. 527–537. DOI: 10.18653/v1/2020.acl-main.50. URL: https://aclanthology.org/2020.acl-main.50/.

Steyn, Jacob van (2023). *Political tweets*. URL: https://huggingface.co/datasets/Jacobvs/PoliticalTweets.

Sun, Chi et al. (2019). "How to Fine-Tune BERT for Text Classification?" In: *CoRR* abs/1905.05583. arXiv: 1905.05583. URL: http://arxiv.org/abs/1905.05583.

Szemraj, Peter (2023). *Goodreads book genres*. URL: https://huggingface.co/datasets/pszemraj/goodreads-bookgenres.

Törnberg, Petter (2024). "Large Language Models Outperform Expert Coders and Supervised Classifiers at Annotating Political Social Media Messages". In: *Social Science Computer Review*. DOI: 10.1177/08944393241286471. eprint: https://doi.org/10.1177/08944393241286471. URL: https://doi.org/10.1177/08944393241286471.

Vélez Castañeda, Alejandro (2024). *BERT political bias finetune*. URL: https://huggingface.co/jhonalevc1995/BERT-political_bias-finetune.

Webhose.io (n.d.). *Free News Datasets*. URL: https://github.com/Webhose/free-news-datasets.

Wolbrecht, Christina et al. (2023a). *Policy Agendas Project: Democratic Party Platform*.

— (2023b). *Policy Agendas Project: Republican Party Platform*.

Wolf, Thomas et al. (2020). *HuggingFace's Transformers: State-of-the-art Natural Language Processing*. arXiv: 1910.03771 [cs.CL]. URL: https://arxiv.org/abs/1910.03771.

Wong, Felix Ming Fai et al. (2016). "Quantifying Political Leaning from Tweets, Retweets, and Retweeters". In: *IEEE Transactions on Knowledge & Data Engineering* 28.08, s. 2158–2172. ISSN: 1558-2191. DOI: 10.1109/TKDE.2016.2553667. URL: https://doi.ieeecomputersociety.org/10.1109/TKDE.2016.2553667.

Xiao, Zhiping et al. (2020). "TIMME: Twitter Ideology-detection via Multi-task Multi-relational Embedding". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '20. Virtual Event, CA, USA: Association for Computing Machinery, 2258–2268. ISBN: 9781450379984. DOI: 10.1145/3394486.3403275. URL: https://doi.org/10.1145/3394486.3403275.

Yan, Hao, Allen Lavoie a Sanmay Das (2017). "The Perils of Classifying Political Orientation From Text". In: *LINKDEM@IJCAI*. URL: https://api.semanticscholar.org/CorpusID:31418886.

Zhang, Xiang, Junbo Jake Zhao a Yann LeCun (2015). "Character-level Convolutional Networks for Text Classification". In: *CoRR* abs/1509.01626. arXiv: 1509.01626. URL: http://arxiv.org/abs/1509.01626.

Zhitomirsky-Geffet, Maayan et al. (čvn. 2016). "Utilizing overtly political texts for fully automatic evaluation of political leaning of online news websites". In: *Online Information Review* 40, s. 362–379. DOI: 10.1108/OIR-06-2015-0211.

Zhuang, Liu et al. (srp. 2021). "A Robustly Optimized BERT Pre-training Approach with Post-training". eng. In: *Proceedings of the 20th Chinese National Conference on Computational Linguistics*. Ed. Sheng Li et al. Huhhot, China: Chinese Information Processing Society of China, s. 1218–1227. URL: https://aclanthology.org/2021.ccl-1.108/.

Seznam obrázků

Seznam tabulek

Příloha A: Podrobnosti metodologie průniku datových sad

Celkem existuje devět možných kombinací přítomnosti nebo absence titulků a textových těl záznamů v datových sadách. Každá z nich vyžaduje vlastní přístup k porovnání, jak je uvedeno v tabulce A.1.

		řádek 1. datové sady obsahuje		
		pouze titulek	pouze tělo	obojí
řádek 2. datové sady obsahuje	pouze titulek	titulky jsou shodné	tělo obsahuje titulek	titulky jsou shodné
	pouze tělo	tělo obsahuje titulek	2. tělo obsahuje výřez z 1.	2. tělo obsahuje výřez z 1.
	obojí	titulky jsou shodné	2. tělo obsahuje výřez z 1.	titulky jsou shodné

Tabulka A.1: Zvolená porovnání na základě přítomnosti nebo absence těl a titulků při hledání průniku datových sad.

Příloha B: Kompletní výsledné tabulky

B.1 Průnik datasetů

Tabulka B.1.

B.2 Existující modely na političnost

Tabulka B.2.

B.3 Benchmark datových sad na političnost s vynecháním jedné

Tabulka B.3.

	Article bias prediction	BIGNEWSBLN	CommonCrawl news articles	Dem., rep. party platform topics	GPT-4 political bias	GPT-4 political ideologies	Media political stance	Political podcasts	Political tweets	Qbias	Webis bias flipper 18	Webis news bias 20
Article bias prediction	–	1	2	0	0	0	2	0	0	0	7	10
BIGNEWSBLN	0	–	1	0	0	0	1	0	0	0	0	0
CommonCrawl news articles	1	1	–	0	0	0	2	0	0	0	0	0
Dem., rep. party platform topics	0	0	0	–	0	0	0	0	0	0	0	0
GPT-4 political bias	0	0	0	0	–	0	0	0	0	0	0	0
GPT-4 political ideologies	0	0	0	0	0	–	0	0	0	0	0	0
Media political stance	1	1	2	0	0	0	–	0	0	0	0	0
Political podcasts	0	0	0	0	0	0	0	–	0	0	0	0
Political tweets	0	0	0	0	0	0	0	0	–	0	0	0
Qbias	0	0	0	0	0	0	0	0	0	–	0	1
Webis bias flipper 18	42	1	4	0	0	0	3	0	1	1	–	80
Webis news bias 20	50	1	3	0	0	0	3	0	1	2	66	–

	Amazon reviews 2023	Dialogsum	Free news	Goodreads book genres	IMDB	IMDB movie genres	Medium post titles	News category	PENS	PoliBERTweet	Political or not	Recipes	Reddit comments	Reddit submissions	Rotten tomatoes	Textbooks	Tweet topic multi	Yelp review full
Amazon reviews	–	0	0	0	1	0	1	0	0	0	0	1	1	1	0	0	0	1
Dialogsum	2	–	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
Free news	0	0	–	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Goodreads book genres	1	0	0	–	0	0	5	0	0	0	0	0	0	0	0	0	0	0
IMDB	1	0	0	0	–	0	4	0	0	0	0	0	0	0	0	0	0	0
IMDB movie genres	0	0	0	0	0	–	3	0	0	0	0	0	0	0	0	0	0	0
Medium post titles	1	0	0	1	8	7	–	0	0	0	0	0	1	0	0	0	0	0
News category	0	0	0	0	0	0	0	–	0	0	0	0	0	0	0	0	0	0
PENS	0	0	0	0	0	0	0	0	–	0	0	0	0	0	0	0	0	0
PoliBERTweet	0	0	0	0	0	0	0	1	0	–	0	0	1	0	0	0	0	0
Political or not	3	0	0	0	0	0	2	3	0	1	–	0	0	0	0	0	0	0
Recipes	0	0	0	0	0	0	0	0	0	0	0	–	0	0	0	0	0	0
Reddit comments	0	0	0	0	0	0	0	0	0	0	0	0	–	1	0	0	0	1
Reddit submissions	1	0	0	0	0	0	0	0	0	0	0	0	3	–	0	0	0	0
Rotten tomatoes	1	0	0	0	0	0	0	1	0	0	0	0	3	0	–	0	0	0
Textbooks	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	–	0	0
Tweet topic multi	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	–	0
Yelp review full	1	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	–

Tabulka B.1: Výsledné průniky mezi každou dvojicí datových sad politického zabarvení (první tabulka) a političnosti (druhá tabulka). Čísla jsou procenta a vyjadřují podíl velikosti průniku vůči celkovému počtu záznamů v datové sadě v odpovídajícím řádku. Například u datových sad politického zabarvení má Article bias prediction průnik 740 instancí s CommonCrawl news articles, což jsou 2 % sady Article bias prediction. V řádku CommonCrawl news articles stejná velikost průniku (740) činí pouze 1 % této datové sady.

	Classifier main subject politics	Political DEBATE large	Topic politics
Amazon reviews	99,8	99,4	100
Dialogsum	99,6	98,8	99,9
Free news	73,8	82,8	67,4
Goodreads book genres	99,3	99,2	100
IMDB	99,9	99,7	100
IMDB movie genres	98,6	97,8	99,7
Medium post titles	62,9	93,7	81,9
News category	68,8	95,3	80,3
PENS	75,4	97,8	87,9
PoliBERTweet	56	85,9	81,7
Political or not	73,5	99,8	83,4
Recipes	100	100	100
Reddit comments	98,2	95,2	99,3
Reddit submissions	99,3	98,5	99,7
Rotten tomatoes	99,8	96,2	99,6
Textbooks	58	95,6	54,5
Tweet topic multi	99,5	98,6	99,8
Yelp review full	99,8	99,8	100
Article bias prediction	69,6	93,4	76,7
BIGNEWSBLN	69,6	84,3	65
CommonCrawl news articles	71,1	98,8	70,5
Dem., rep. party platform topics	70,1	98,2	81,4
GPT-4 political bias	66,5	99,8	48,9
GPT-4 political ideologies	18,3	98,6	86,9
Media political stance	60,3	81,4	57,4
Political podcasts	10,2	89,7	41,5
Political tweets	54,3	85,2	56,4
Qbias	58,5	90,7	68
Webis bias flipper 18	75,2	92,2	81,8
Webis news bias 20	71,9	94,6	78,2

Tabulka B.2: Výsledná F_1 skóre po evaluaci existujících modelů političnosti na všech dostupných datových sadách (jak političnosti, tak politického zabarvení), každé navzorkované na 1 000 záznamů. Při interpretaci těchto metrik je třeba vzít v úvahu distribuce tříd v datových sadách, jak je vysvětleno v sekci 3.4.2. Pro zjednodušené, vyvážené a přehledné srovnání vizte sekci 4.2.2.

VYNECHANÁ DATOVÁ SADA ↓	BERT		RoBERTa		DeBERTa		POLITICS	
TYP TESTOVACÍHO VZORKU →	známý	neznámý	známý	neznámý	známý	neznámý	známý	neznámý
Amazon reviews	97,6	99,9	97,9	100	97,9	99,9	97,8	100
Dialogsum	97,6	99	97,9	98,7	97,9	98,1	97,6	98,5
Free news	97,5	96,9	97,9	97,2	97,9	97,8	98	95,7
Goodreads book genres	97,3	98,6	97,6	98,7	97	99,3	97,8	99,2
IMDB	97,4	98,3	97,3	94,6	97,8	96,6	97,9	98,9
IMDB movie genres	97,5	99,6	97,2	98,9	97,9	99,5	97,8	99,8
Medium post titles	98	69,1	98,4	65	98,4	67,9	98,4	71,7
News category	97,7	99,2	97,7	96,8	97,8	98,9	97,4	92,3
PENS	97,5	100	98	100	98	100	97,7	100
PoliBERTweet	97,7	98,8	97,2	99,8	97,7	94,6	97,7	96,3
Political or not	97,6	99,8	97,4	99,8	97,8	99,8	97,8	99,7
Recipes	97,5	99	97,1	96,4	96,8	94,3	97,9	99
Reddit comments	97,6	92,7	97,9	85,5	98,1	91,4	98,1	90,8
Reddit submissions	98,1	88,8	97,6	87,1	98,3	90,8	98,4	88,6
Rotten tomatoes	97,6	86,8	97,8	92,3	98,1	92,1	97,7	86,6
Textbooks	97,7	91,3	97,8	90,8	97,9	90,4	98,3	78
Tweet topic multi	97,7	89,9	50,7	33,3	97,5	92,9	97,9	84,7
Yelp review full	97,6	79,4	97,8	79,4	98	77,3	97,8	75,8
Article bias prediction	97,6	89,8	97,7	87,2	98	94,1	97,6	87,5
BIGNEWSBLN	97,6	51,3	97,9	62	97,3	76,6	97,7	73,9
CommonCrawl news articles	97,6	97,7	97,6	100	98	97,7	97,9	96,1
Dem., rep. party platform topics	97,6	100	95	98,6	95,8	100	97,8	100
GPT-4 political bias	97,6	98	97,9	96,7	97,8	96,8	98	96,4
GPT-4 political ideologies	97,5	99	97,7	99	97,9	99,6	97,9	99,6
Media political stance	97,6	99,2	97,1	98,9	97,4	99,7	97,9	99,3
Political podcasts	97,7	93,6	97,9	93,3	98,1	95,3	97,9	90,8
Political tweets	97,5	96,7	97,1	98,1	97,1	95,1	97,6	91,9
Qbias	97,6	99,9	97,8	99,9	97,8	100	97,5	99,9

Tabulka B.3: Výsledná F_1 skóre při klasifikaci političnosti po evaluaci základních verzí modelů BERT (cased), RoBERTa, DeBERTa V3 a POLITICS (velikost testovacího vzorku: 15 %; velikost trénovacího vzorku: 1 000 záznamů z každé sady) trénovaných na všech dostupných datových sadách (28) kromě jedné (vynechané) – jak je popsáno v sekci 3.5.2. Hodnoty ve sloupcích známého testovacího vzorku znamenají průměrné skóre na všech datových sadách, na kterých byl model trénován. Sloupce neznámého testovacího vzorku zobrazují skóre na oné jediné vynechané datové sadě. Pro průměry těchto skóre vizte sekci 4.3.2.

Příloha C: Hardware

Průniky datových sad byly měřeny na Intel Core i7-3770, 16 GB RAM. Trvalo to 50 hodin.

Na stroji s NVIDIA RTX 2060, AMD Ryzen 7 4800H a 16 GB RAM byly vyhodnoceny existující modely (přibližně 5 minut na model na jedné datové sadě) a natrénovány prototypní modely pro benchmarky datových sad bez optimalizace hyperparametrů (zhruba 15 minut na model pro ponechání jedné, 35 minut na model pro vynechání jedné).

Hledání hyperparametrů pro POLITICS probíhalo na stroji s NVIDIA RTX 4000 Ada, 5 vCPU, 47 GB RAM a jeden pokus trval 70 minut.

Ladění hyperparametrů pro DeBERTa V3 large bylo provedeno na NVIDIA RTX 4070 Ti, 8 vCPU, 30 GB RAM a jedna epocha trvala přibližně 1 hodinu.

Zadání maturitního projektu z informatických předmětů

Jméno a příjmení:	*Matouš Volf*
Pro školní rok:	*2024/2025*
Třída:	*4. A*
Obor:	*Informační technologie 18-20-M/01*
Téma práce:	*Analýza textů internetových článků a příspěvků pomocí LLM*
Vedoucí práce:	*doc. Ing. Jakub Šimko, PhD.*

Způsob zpracování, cíle práce, pokyny k obsahu a rozsahu práce:

Cílem projektu je analyzovat a klasifikovat textový obsah internetových novinových článků, příspěvků na sociálních sítích a jejich diskuzí. K tomuto účelu budou využity modely strojového učení a velké jazykové modely. Analýza bude provedena na existujících datových sadách.

Na textech budou pozorovány a vyhodnocovány znaky:

- téma,
- sentiment,
- vzájemný vztah,
- vzájemná shoda či neshoda.

Výstupem práce bude vědecký článek obsahující výsledky výzkumu a jejich interpretaci.

Stručný časový harmonogram (s daty a konkretizovanými úkoly):

září, říjen:

- nabytí teoretických znalostí potřebných k realizaci
- rešerše existujících prací zabývajících se stejným tématem
- průzkum dostupných nástrojů a modelů

listopad:

- čištění a preprocessing dat
- experimentace s využitím různých modelů a postupů

prosinec, leden:

- analýza samotná

únor, březen:

- interpretace a porovnání výsledků
- vyvození závěrů