FAKULTA APLIKOVANÝCH VĚD
ZÁPADOČESKÉ UNIVERZITY
V PLZNI

KATEDRA INFORMATIKY
A VÝPOČETNÍ TECHNIKY

Bakalářská práce

Metody Active learning pro úlohy automatického zpracování přirozeného jazyka

Jakub Mladý

PLZEŇ

2023

FAKULTA APLIKOVANÝCH VĚD
ZÁPADOČESKÉ UNIVERZITY
V PLZNI

KATEDRA INFORMATIKY
A VÝPOČETNÍ TECHNIKY

Bakalářská práce

Metody Active learning pro úlohy automatického zpracování přirozeného jazyka

Jakub Mladý

Vedoucí práce
Ing. Jakub Sido

PLZEŇ **2023**

Citace v seznamu literatury:
MLADÝ, Jakub. *Metody Active learning pro úlohy automatického zpracování přirozeného jazyka*. Plzeň, 2023. Bakalářská práce. Západočeská univerzita v Plzni, Fakulta aplikovaných věd, Katedra informatiky a výpočetní techniky. Vedoucí práce Ing. Jakub Sido.

ZÁPADOČESKÁ UNIVERZITA V PLZNI
Fakulta aplikovaných věd
Akademický rok: 2022/2023

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

(projektu, uměleckého díla, uměleckého výkonu)

Jméno a příjmení:	**Jakub MLADÝ**
Osobní číslo:	**A20B0190P**
Studijní program:	**B0613A140015 Informatika a výpočetní technika**
Specializace:	**Informatika**
Téma práce:	**Experimenty s metodami Active Learning na rozsáhlých datasetech přirozeného jazyka**
Zadávající katedra:	**Katedra informatiky a výpočetní techniky**

Zásady pro vypracování

1. Prostudujte problematiku Active Learning pro použití ve strojovém učení užívající neuronové sítě pro úlohy zpracování přirozeného jazyka.
2. Na základě předchozí analýzy vyberte z existujících již implementovaných technik reprezentativní vzorek a navrhněte množinu experimentů nad dodanými datovými sadami. Experimenty budou probíhat v plně automatizovaném režimu a poskytovat výsledky v podobě vhodných metrik.
3. Navrženou množinu experimentů realizujte a na základě porovnání výše uvedených metrik změřte přínosy metod Active Learning v každém experimentu.
4. Dosažené výsledky pro jednotlivé metody zpracujte do grafů závislostí počtu označených vzorků na výsledné úspěšnosti systému. Výsledky kriticky zhodnoťte.

Rozsah bakalářské práce: **doporuč. 30 s. původního textu**
Rozsah grafických prací: **dle potřeby**
Forma zpracování bakalářské práce: **tištěná/elektronická**

Seznam doporučené literatury:

Dodá vedoucí bakalářské práce

Vedoucí bakalářské práce: **Ing. Jakub Sido**
Nové technologie pro informační společnost

Datum zadání bakalářské práce: **3. října 2022**
Termín odevzdání bakalářské práce: **4. května 2023**

L.S.

Doc. Ing. Miloš Železný, Ph.D.
děkan

Doc. Ing. Přemysl Brada, MSc., Ph.D.
vedoucí katedry

V Plzni dne 25. října 2022

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného akademického titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Západočeská univerzita v Plzni má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

Ve Františkových Lázních dne 21. června 2023

. . . . . .
Jakub Mladý

V textu jsou použity názvy produktů, technologií, služeb, aplikací, společností apod., které mohou být ochrannými známkami nebo registrovanými ochrannými známkami příslušných vlastníků.

Abstrakt

Aktivní učení je přístup k trénování modelů umělé inteligence v rámci učení s učitelem. Motivací k jeho zavedení je šetření času a finančních prostředků při pořizování anotací trénovacích dat, jež jsou v mnoha úlohách strojového učení potřebné. Jeho paradigma je jednoduché: předpoklady jsou datová sada s několika málo označenými daty, model a anotátor. Následně do splnění vhodně zvolené ukončovací podmínky probíhá následující cyklus – naučit model na označené podmnožině, vybrat několik neoznačených vzorků a dotázat se anotátora na označení vybraných vzorků. Kritickou částí systému je pak výběr vzorků. Při použití vhodné strategie je možné vybrat takové prvky, jejichž naučením se model zlepší nejvíce. Právě tyto strategie jsou hlavním předmětem výzkumu aktivního učení. Tato práce nabízí průzkum již existujících strategií a poskytuje hodnocení přínosu některých strategií a aktivního učení jako celku z výsledků navržených experimentů.

Abstract

Active learning is an approach to training artificial intelligence models within supervised learning. The motivation for its introduction comes from the need to spare time and financial resources in obtaining labels for the training data, which are fundamental for many machine learning tasks. Its paradigm is simple: the preconditions are a training dataset with a few data labeled, the model and an annotator. Then, the following cycle is repeated until some suitable terminal condition is met – train the model on labeled data, query for more unlabeled data and let the annotator provide the labels for the queried instances. The critical part of such a system lies in the startegy of querying for new data samples. With a desirable strategy, such data labels could be obtained, that imporve the model the most. These query strategies are the main subject of ongoing research in active learning. This work offers a survey of existing strategies and provides evaluation of contribution of some chosen strategies and active learning as a whole from the results of proposed experiments.

Klíčová slova

aktivní učení • zpracování přirozeného jazyka • umělá inteligence • strojové učení • neuronové sítě • transformery • HuggingFace • Metacentrum • PyTorch • modAL

• Small-Text • Weights and Biases

Poděkování

Tímto bych rád poděkoval Ing. Jakubu Sidovi za nápomocné vedení bakalářské práce, odborné rady a ochotu věnovat mi svůj cenný čas a energii.

Chtěl bych také poděkovat své rodině a přátelům za podporu během svého studia.

Dále bych rád poděkoval projektu e-INFRA CZ (ID:90140) za poskytnutí výpočetních zdrojů pro realizaci experimentů této práce[1].

Nakonec děkuji Piotru Bialeckimu za jeho vytrvalé pomáhání komunitě vývojářů ve frameworku PyTorch.

[1] znění oficiálního poděkování: Computational resources were provided by the e-INFRA CZ project (ID:90140), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

Obsah

1 Úvod **3**

2 Související práce **5**

3 Aktivní učení 7
- 3.1 Třídění metod . . . 8
- 3.2 Metody založené na heterogenitě . . . 8
  - 3.2.1 Vzorkování založené na nejistotě . . . 8
  - 3.2.2 Komisní vzorkování . . . 10
  - 3.2.3 Očekávaná změna modelu . . . 10
- 3.3 Metody založené na výkonnosti . . . 11
- 3.4 Metody založené na reprezentativnosti . . . 11
- 3.5 Bayesovksé aktivní učení . . . 12
- 3.6 Dávkové vzorkování . . . 13
  - 3.6.1 Hodnocené dávkové vzorkování . . . 13
  - 3.6.2 Nejistota a diverzita . . . 14
  - 3.6.3 Diskriminativní aktivní učení . . . 16
- 3.7 Možnosti Implementace . . . 16
  - 3.7.1 Studený začátek . . . 17

4 Experimenty **19**
- 4.1 Vymezení pojmů . . . 19
- 4.2 Typy a cíle experimentů . . . 20
  - 4.2.1 Experimenty východisek . . . 20
  - 4.2.2 Experimenty aktivního učení . . . 21
- 4.3 Datové sady . . . 22
- 4.4 Modely . . . 23
  - 4.4.1 Modely závislé na jazyku . . . 24
  - 4.4.2 Modely nezávislé na jazyku . . . 24
  - 4.4.3 Trénink . . . 25
- 4.5 Návrh Experimentů . . . 26

4.6 Zaznamenané potíže 27

5 Výsledky a Diskuze **29**
5.1 Grafy 29
5.2 Diskuze 29
5.2.1 Výpočetní složitost některých strategií 29
5.2.2 Přínos aktivního učení 30
5.2.3 Zhodnocení 32

6 Závěr **35**

A Grafy experimentů **37**
A.0.1 Experimenty aktivního učení 37
A.0.2 Experimenty východisek 44

Bibliografie **47**

Seznam obrázků **51**

Úvod 1

Zpracování přirozeného jazyka (angl. natural language processing, NLP) je jednou z mnoha oblastí strojového učení, jež se snaží reprezentovat přirozený jazyk, tedy jazyk lidí, vhodným způsobem za účelem jeho porozumění jazykovými modely a následného plnění lidmi požadovaných úkonů – klasifikace textů, analýza sentimentu, rozpoznávání pojmenovaných entit, nebo i obecnější zodpovídání dotazů a mnoho dalších – pomocí těchto modelů. V současné době zažíváme obrovský rozmach tohoto oboru; můžeme se setkat s velmi výkonnými modely jako např. BERT, GPT-4, ChatGPT, Llama, jejichž jazyková znalost se mnohdy vyrovná rodilým mluvčím. Kromě jazykových dovedností však také vynikají jinými znalostmi; z obou důvodů pomoc těchto modelů vítají lidé z různých oborů. Efektivita dnešních jazykových modelů je často neodborným publikem nepochopená, k její demystifikaci je třeba rozklíčovat princip jejich úspěchu. Ten tkví zejména v architektuře modelu a datech, kterými je model učen. Znalosti modely čerpají z dat, architektura jim pomáhá v jejich lepším pochopení. Vyspělejší architektury, přístupy učení a zpracování již obdržených dat jsou předmětem výzkumu, data je však nutné v první řadě získat, což se může stát časově nebo i finančně náročným úkolem. Kromě jich samotných vyžaduje mnoho úloh strojového učení (nejen NLP) také anotace těchto dat – správné odpovědi, z nichž se modely mohou učit. Získávání anotací pak představuje další časově/finančně obtížný problém. Na komplikace s pořizováním anotací reaguje právě schéma aktivního učení (angl. active learning, AL), jehož principem je umožnit modelu výběr datových vzorků, které potřebuje označit. Sběr anotací pak nemusí být proveden pro celou datovou množinu před učením, ale může probíhat inkrementálně a nikoliv pro všechna data.

Cílem této práce je prozkoumat metodiky (strategie) aktivního učení zejména se zaměřením na modely neuronových sítí používaných v oblasti zpracování přirozeného jazyka. Současně budou tyto postupy taxonomicky kategorizovány podle stávající literatury z oblasti, zejména průzkumů aktivního učení. Studium metod zahrne jak starší (původní, ale stále používané), tak i modernější strategie, které cílí právě na neuronové sítě. Ke zjištění přínosu aktivního učení v oblasti zpracování přirozeného jazyka budou následně vybrány vhodné nastudované metody s

ohledem na jejich rozšíření, pokrok a diverzitu. S nimi budou navrženy smysluplné experimenty odpovídající možné praxi nad moderními jazykovými modely (tzv. transformery) i nad modely zcela odlišné architektury – náhodnými lesy – a nad rozsáhlými (s více než dvaceti tisíci až s přes půl milionu vzorky) datovými sadami v anglickém či českém jazyce. Experimenty proběhnou automatizovaně v gridovém výpočetním prostředí metacentra a jejich výsledky budou zpracovány do grafů závislostí metriky představující výkonnost modelu – přesnosti – na počtu označených dat. Posledním úkolem bude z grafů kriticky posoudit, zda a případně jak moc aktivní učení ovlivňuje schopnosti výsledného modelu oproti náhodnému výběru vzorků a pasivnímu učení.

Související práce 2

Experimentování s metodami aktivního učení je předmětem zejména výzkumných prací přinášejících nové strategie do oboru – v zájmu autorů je totiž zjistit, jakých jimi vyvinutá strategie dosahuje výsledků v porovnání s ostatními strategiemi [KAG19; GIG17; Ash+19; SS18; GS19; Car+17]. Tyto práce se však nezaměřují pouze na oblast zpracování přirozeného jazyka a často nezkoumají přínosy aktivního učení pro moderní modely NLP – transformery. Vybírají datové sady z různých oblastí a modely různých architektur (z neuronových sítí zejm. konvoluční sítě). Experimenty mají ale také velice podobné vlastnosti. Prezentace výsledků je provedena grafem závislosti přesnosti (nebo výjimečně jiné metriky) na počtu označených vzorků případně vyjádřeném jako procento z celkové velikosti sady. Své strategie porovnávají s náhodným vzorkováním, pak často s některou z heterogenních metod založených na nejistotě – entropie, největší rozdíl nebo nejmenší jistota – a případně dalšími, v době výzkumu špičkovými, strategiemi.

Další podstatnou částí této práce je představení některých strategií aktivního učení. Stejný cíl si kladou autoři průzkumů této oblasti, kteří se snaží zmapovat metody AL nebo jeho podoblastí v jedné práci. Základním a nejzásadnějším dílem této kategorie jistě je průzkum *Active Learning Literature Survey* od Burra Settlese [Set09]. Obdobnou prací je průzkum Charu C. Aggarwala a dalších [Agg+14]. Jeden z novějších průzkumů se pak zaměřuje na aktivní učení pro neuronové sítě – *A Survey of Deep Active Learning* od autorů Pengzhen Ren, Yun Xiao et al. Ve své práci například jmenují některé známé problémy aktivního učení s těmito modely, jako třeba nutnost dávkového přístupu nebo nemožnost uspokojivé reprezentace nejistoty modelu.

Existují také výzkumy zabývající se úspěšností aktivního učení – ty jsou této bakalářské práci nejbližší. Například Siddharth Karamcheti et al. ve své práci [Kar+21] zkoumají, proč mnohé strategie selhávají pro úlohu dotazování vizuální informace (angl. visual question answering). Přicházejí s možným vysvětlením, že strategie aktivního učení často vybírají kolektivní odlehlé vzorky, tj. vzorky zásadně se lišící od zbytku sady, které se ale (v příznakovém nebo jiném prostoru) shlukují k sobě, jež jsou pro model složité na nauku. Dále tyto kolektivní odlehlé vzorky v rámci datové

sady identifikují a pozorují průběhy aktivního učení při postupném odstraňování těchto prvků. Z experimentů docházejí k závěru, že tyto vzorky skutečně negativně ovlivňují aktivní učení, neboť při postupném ořezávání sady o 10, 25 a 50 % dochází ke stále lepším výsledkům AL oproti náhodnému vzorkování. V experimentech se zaměřují na 8 strategií AL a 5 modelů na 4 datových sadách pro dotazování vizuální informace.

Této práci je ještě blíže článek [DD22], kde autoři zkoumají limitace aktivního učení přímo pro transformer modely. Například ověřují problematiku kolektivních odlehlých vzorků, velikosti modelu nebo „předběžného naučení“ v aktivním učení. Ze své studie vyvozují další možnost, proč aktivní učení může selhávat, a to nestabilitu tréninku – strategie totiž vybírají potencionálně užitečné prvky, které ale způsobují nestabilní učení.

Aktivní učení 3

Aktivní učení je ve strojovém učení s učitelem technika pomáhající minimalizovat jeho náklady [Ren+20; Agg+14] tak, že učený model sám vybírá vzorky pro označení orákulem – abstraktním znalcem schopným podat správné anotace datových vzorků. Vybírání je realizováno tzv. dotazovacími strategiemi (angl. query strategies). Cílem strategií je nějakým způsobem získat takové neoznačené vzorky, jejichž anotace pomůže modelu co nejvíce.

Význam. Aktivní učení má svůj význam ve strojovém učení s učitelem, kde je potřeba mít k dispozici veliké sady anotovaných dat, aby výsledný model podával uspokojivé výsledky. Toto ale představuje problém, protože takové množství dat je nákladné na obdržení, vyžaduje monotónní práci mnoha lidí a učení modelu může být pomalejší. Přesně na tyto problémy reaguje aktivní učení. Při tomto přístupu není potřeba mít označena všechna data – jen malou množinu prvotních vzorků, na kterých se model neúplně učí. Hlavní myšlenkou této metody je, že nadále si model sám určitým způsobem vybírá z neoznačené množiny dat (anebo je sám generuje) ty vzorky, které jeho učení podpoří nejvíce [Ren+20; Set09]. Tato data nechá anotovat orákulem (např. lidmi) a následně se na nich přiučí (nebo učení proběhne od nuly se všemi doposud označenými vzorky – viz sekci 3.7.1). Cyklus výběru vzorků modelem a dotázání se na anotace probíhá až do nějaké ukončovací podmínky – např. počet cyklů nebo hodnota metriky pro přesnost modelu.

Celkový počet anotovaných dat je tedy velikost prvotní množiny v součtu s počtem modelem vybraných vzorků po všech cyklech. Toto číslo musí být menší nebo rovné celkovém počtu dat (ne- i označených), neboť víc jich není, a závisí na ukončovací podmínce aktivního učení. Pokud je podmínka volena vhodně a model vybírá dobré vzorky, může být počet anotací mnohonásobně menší než velikost dat. Z toho plyne, že k celkovému anotování není zapotřebí tolik hodin lidské práce, čímž se ušetří finanční zdroje.

3.1 Třídění metod

Literatura nejčastěji dělí aktivní učení do tří skupin podle přístupů (scenarios [Set09]): *syntéza dotazů* (Membership query synthesis), *sekvenční (proudové) dotazování* (Sequential, stream sampling) a *výběrové dotazování* (Pool based sampling) [Ren+20; Set09; Agg+14]. V syntéze dotazů model vytváří nové umělé vzorky, které se nemusí nacházet v původním datasetu. Úskalím tohoto přístupu je, že pro určité problémy (např. rozpoznávání obrázků, většinu úloh NLP) tento generativní přístup syntetizuje nesmyslné a obtížně označitelné vzorky [Agg+14]. Sekvenční dotazování prochází postupně celým neoznačeným datasetem a pro každý vzorek vyhodnotí, zda má pro něj být dotázáno označení, či nikoliv. Naproti tomu výběrové dotazování vybírá nejlepší vzorek z celého (neoznačeného) datasetu. Sekvenční přístup je vhodnější pro učení na menších zařízeních (mobilní, embedded), kdežto výborový se běžně používá na silnějších strojích [Set09; Ren+20]. Všechny tři přístupy však uvažují, že vybíraný vzorek je jen jeden, což je pro některé modely nevhodné. Vznikl tedy další přístup – *dávkové vzorkování.* Strategie spadající do této třídy se však od ostatních zásadně liší tím, že potřebují informace navíc – viz sekci 3.6.

Mimo přístupů lze dělit také konkrétní strategie. Jedním z příkladů je v předchozím odstavci zmíněné dávkové vzorkování – jelikož se strategie tohoto přístupu vyznačují informací navíc. Podle Aggarwala et al. [Agg+14] lze dále dělit strategie na metody založené na: *heterogenitě*, *výkonnosti* nebo *reprezentativnosti* (anebo hybridní metody spadající do více než jedné z těchto tříd). Dále uvádějí například *Bayesovské aktivní učení* a *metody založené na hustotě informace*. V literatuře se dá ovšem setkat i s jinými druhy kategorizace – např. dělení na metody založené na nejistotě nebo diverzitě [YLB20; Ash+19].

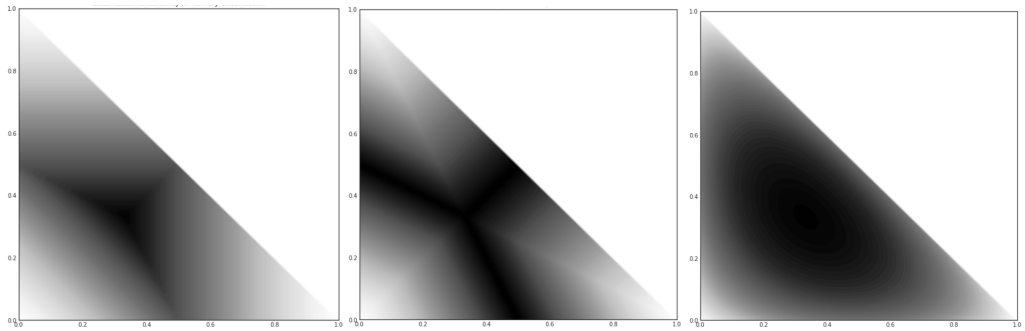
Všechny popsané třídy se zabývají učením klasifikátorů. Existují však strategie zaměřující se na regresi. Například mezi ně patří Bayesovská optimalizace [DH18, dokumentace knihovny].

3.2 Metody založené na heterogenitě

Tyto strategie nějakým způsobem měří různorodost dat od znalostí modelu a vybírají ty nejvíce odlišné. Dále se dělí na např. *vzorkování založené na nejistotě*, *komisní dotazování* nebo *očekávaná změna modelu*.

3.2.1 Vzorkování založené na nejistotě

Jedněmi z prvních a stále běžně používaných strategií jsou metody založené na nejistotě modelu. Tyto strategie vyžadují, aby byl klasifikátor schopen určit pravdě-



(a) Nejmenší jistota (b) Největší rozdíl (c) Entropie

Obrázek 3.1: Vizualizace výběru vzorků jednotlivými metodami založenými na nejistotě pro úlohu ternární klasifikace. Osy zobrazují pravděpodobnost dvou tříd, pravděpodobnost třetí třídy se dopočítá jako $1 - p_1 - p_2$. Grafy pak vyjadřují skóre jednotlivých metod, vzorky s pravděpodobnostní distribucí odpovídající tmavým bodům jsou lepší kandidáti podle dané metody. Obrázek z [DH18, dokumentace knihovny]

podobnostní distribuci nad třídami pro daný vzorek:

$$p(x) = P_\Theta(Y = y_i|x) \forall i,$$

kde malé p je distribuce nad vzorkem x a velké P_Θ je pravděpodobnost, že vzorek x patří do třídy y_i, když jsou dány parametry modelu Θ. Díky distribuci pak můžeme měřit nejistotu. Tyto strategie je možné použít v přístupech selektivního nebo výběrového vzorkování. Obrázek 3.1 znázorňuje přednostně vybírané vzorky každé z těchto metod.

Nejmenší jistota. Tato strategie vybírá prvky, pro něž i ta nejpravděpodobnější třída má stále nejnižší pravděpodobnost. U vzorků tedy vybere třídu s nejvyšší hodnotou pravděpodobnostní funkce a vybere ten vzorek, kde je tato pravděpodobnost nejnižší (neboli s největším rozdílem od maxima – 100 %). Ve výběrovém vzorkování lze použít následující vzorec k obdržení nejvhodnějšího vzorku:

$$\hat{x} = \text{argmax}_x(1 - \hat{p}(x)),$$

kde $\hat{p}(x) = \max_i\{P_\Theta(Y = y_i|x)\}$.

Největší rozdíl. Největší rozdíl měří rozdíl pravděpodobností tříd, v nichž si je model nejvíce jistý. Vybírá pak ty vzorky, pro něž je rozdíl nejmenší. Literatura často tuto metodu označuje také jako „rozdílové vzorkování“ [Set09; DH18] kvůli jinak matoucímu názvu. Vzorek pro označení je vybrán podle vztahu:

$$\hat{x} = \text{argmin}_x(\hat{p_1}(x) - \hat{p_2}(x)),$$

kde $\hat{p_1}(x)$ a $\hat{p_2}(x)$ jsou největší a druhá největší (v tomto pořadí) z hodnot $P_\Theta(Y = y_i|x)\forall i$.

Entropie. Vzorkování na základě entropie na rozdíl od předchozích bere v potaz celou distribuci nad vzorkem. Z distribucí vypočítá střední entropii neboli očekávanou míru překvapení. Čím větší je entropie, tím větší je také nejistota, neboť informační entropie nabývá maximálních hodnot při rovnoměrném rozdělení náhodné veličiny (v našem případě třída vzorku). K určení vzorku lze použít následující vzorec:

$$\hat{x} = \text{argmax}_x(-\sum_i p_i(x) \cdot \log_2 p_i(x)),$$

kde $p_i(x) = P_\Theta(Y = y_i|x)$.

Selektivní přístup. Všechny strategie založené na nejistotě byly popsány výběrovým přístupem. K převodu na proudový přístup je potřeba stanovit vhodný práh t a místo vybírání vzorků s největší (resp. nejmenší) hodnotou se dotážeme na třídu vzorku, pokud je hodnota větší (resp. menší) než daný práh. Například pro největší rozdíl vybereme k označení vzorek x, pokud $\hat{p_1}(x) - \hat{p_2}(x) < 0.1$ (s prahem $t = 0.1$).

Obdržení distribuce. Zjištění spolehlivé jistoty modelu založeném na neuronových sítích je stále zkoumaný problém v oblasti strojového učení. Nejpoužívanější technikou pro klasifikační úlohy je transformace výstupu modelu vrstvou softmax, která poskytuje pravděpodobnostní distribuci. Tato distribuce však ne příliš dobře odpovídá jistotě modelu [Gal16, s. 13–14]. Mezi některé spolehlivější techniky patří například použití bayesovských sítí nebo tzv. Monte Carlo vynechávání (angl. dropout), jehož principem je „vynechání“ (vynulování výstupu) některých náhodně vybraných neuronů při inferenci, což způsobí obdržení různých výsledků modelu pro ten samý vzorek. Tyto techniky jsou dále popsány v sekci 3.5.

3.2.2 Komisní vzorkování

Strategie komisního vzorkování vyžadují více modelů určených na stejný úkol strojového učení. Smyslem tohoto přístupu v souvislosti s AL je vybírat ty vzorky, na nichž je shoda modelů nejmenší. Jednotlivé strategie této kategorie měří právě shodu. U neuronových sítí je opět možné použít Monte Carlo vynechávání k simulaci více modelů [Agg+14].

3.2.3 Očekávaná změna modelu

Strategie této kategorie vybírají prvky, které model „změní“ nejvíce, konkrétně které způsobí největší změnu gradientu ztrátové funkce vzhledem k parametrům modelu.

Cílem je vybrat takové vzorky, které jsou modelu nejvíce cizí. Tento přístup může být však použit jen na modely, jejichž učení je založeno na gradientech [Agg+14].

3.3 Metody založené na výkonnosti

Hlavní nevýhodou heterogenních metod je vysoká pravděpodobnost výběru zašuměného nebo jinak nekvalitního vzorku. Strategie v této kategorii se snaží obejít tento problém zaměřením na prvky, které zmenší chybovost modelu (na rozdíl od prvků, v nichž si je nejistý) [Agg+14].

Hlavním představitelem je strategie nazvaná *očekávané snížení chyby* („Expected error reduction"). Jak napovídá i název, strategie se snaží snížit chybovost budoucího modelu. Zároveň se dá diskutovat, že se jedná o komplementární přístup k heterogenitě, jelikož zatímco heterogenní metody maximalizují nejistotu vybíraných prvků, metody založené na výkonnosti nejistotu minimalizují na množině zbývajících (neoznačených) vzorků [Set09; Agg+14].

Pro určení očekávaného snížení chyby je potřeba uvažovat vlastnosti modelu po přidání zkoumaného prvku, což však vyžaduje znalost jeho třídy, kterou nemáme. Ve výpočtech tedy zahrneme všechny možnosti, tedy vlastnosti všech modelů po přidání prvku za předpokladu, že prvek je v dané třídě, pro všechny možné třídy. Takto zkoumáme každý vzorek v neoznačené množině – výpočet je tedy velmi náročný. Toto je také důvodem, proč tento způsob aktivního učení není příliš rozšířený [Set09]. Jedna z možných variací této strategie pro binární klasifikaci je popsána následujícím vzorcem:

$$e(x) = \sum_{i}^{k} P_{\Theta}(Y = y_i|x) \cdot \left(\sum_{j}^{k} \sum_{u \in \mathcal{U}} \left| P_{\Theta^{+(x,y_j)}}(Y = y_j|u) - 0,5 \right| \right),$$

kde $e(x)$ je očekávané snížení chyby, k je počet tříd (a i,j tedy iterují přes třídy), $\mathcal{U}$ je neoznačená sada a $\Theta^{+(x,y_j)}$ značí parametry modelu poté, co je k aktuálnímu modelu (daný parametry Θ) přidán zkoumaný prvek x za předpokladu, že jeho třída je y_j. Strategie vybere prvek s minimálním $e(x)$.

3.4 Metody založené na reprezentativnosti

Očekávané snížení chyby je výhodnější než heterogenní metody v tom ohledu, že je odolnější vůči nekvalitním vzorkům a hledá vlastnosti celé neoznačené sady a nejen zkoumaného prvku. Hledání neokrajových prvků se jeví jako úspěšný přístup, avšak kvůli výpočetní složitosti očekávaného snížení chyby je strategie většinou nepoužitelná, byly tedy zkoumány další postupy. Metody založené na reprezentativnosti se snaží vybírat informativní vzorky, které jsou ale podobné zbytku neoznačených

dat. Hodnocení vzorku vznikne násobkem hodnoty libovolné strategie založené na heterogenitě a míry reprezentativnosti neboli hodnoty (odhadnuté) funkce hustoty v bodě zkoumaného vzorku. Odhad může být proveden např. jádrovým odhadem hustoty [Agg+14].

3.5 Bayesovksé aktivní učení

Tento přístup k aktivnímu učení se zaměřuje na „Bayesovské modely“, které nad svými parametry udržují pravděpodobnostní distribuci namísto bodových odhadů [KAG19; GG15; GIG17]. Autoři v této oblasti upozorňují na neschopnost modelů hlubokých sítí reprezentovat nejistotu, přičemž mnohé strategie aktivního učení právě na nejistotu spoléhají [GIG17; Ren+20; Gal16]. Strategie této kategorie se opírají o Bayesovské aktivní učení neshodou (angl. Bayesian active learning by disagreement, BALD), kritériem měřícím vzájemnou informaci mezi parametry modelu a jeho predikcemi [Hou+11; KAG19; Agg+14]:

$$\mathbb{I}(Y, \omega | x, \mathcal{D}) = \mathbb{H}(Y|x, \mathcal{D}) - \mathbb{E}_{\omega \sim \mathcal{P}(\omega|\mathcal{D})} \mathbb{H}(Y|x, \omega),$$

přičemž x a Y jsou prvek a distribuce pravděpodobností tříd tohoto prvku, $\mathcal{D}$ je trénovací datová sada (s označenými prvky), ω jsou parametry modelu s pravděpodobnostní distribucí $\mathcal{P}(\omega|\mathcal{D})$. $\mathbb{H}$ značí Shannonovu entropii a $\mathbb{E}$ střední hodnotu. Ve výběrovém přístupu je pak pro označení vybrán ten prvek, pro nějž je vzájemná informace největší: $\hat{x} = \operatorname{argmax}_x \mathbb{I}(Y, \omega | x, \mathcal{D})$. Levý člen výrazu, $\mathbb{H}(Y|x, \mathcal{D})$, lze vyložit jako entropii predikcí modelu a pravý člen jako střední hodnotou entropie predikcí přes distribuci parametrů [KAG19].

Článek původní strategie BALD se zaměřoval zejména na modely tzv. Gaussovské procesy a vyvozuje postup, jak aproximovat vzájemnou informaci pro tyto modely [Hou+11]. Novější zásahy do bayesovského aktivního učení se však zabývají již neuronovými sítěmi. U nich vzniká potřeba zavedení již zmíněné distribuce nad parametry – tu lze odhadnout například použitím Monte Carlo vynechávání. Různé výsledky modelu aplikací této techniky lze pak chápat jako realizace distribuce nad parametry [Agg+14; GIG17; GG15]. Yarin Gal a Zoubin Ghahramani [GG15] pak přicházejí s aproximací distribuce tříd s využitím MC vynechávání [GIG17]:

$$\mathcal{P}(Y = y | x, \mathcal{D}) \approx \frac{1}{N} \sum_{i=0}^{N} \mathcal{P}(Y = y | x, \hat{\omega}_i)$$

s N výsledky MC vynechávání. Výraz $\mathcal{P}(Y = y | x, \hat{\omega}_i) =: p_y^i$ značí pravděpodobnost třídy y vzorku x pro i-tý výsledek MC. Pomocí tohoto vzorce lze přepsat kritérium strategie BALD pro neuronové sítě (také označováno DBALD – Deep BALD):

$$\mathbb{I}(Y, \omega | x, \mathcal{D}) \approx -\sum_{y \in Y} \left(\frac{1}{N} \sum_i p_y^i \right) \log \left(\frac{1}{N} \sum_i p_y^i \right) + \frac{1}{N} \sum_{y \in Y, i} p_y^i \log \left(p_y^i \right).$$

Andreas Kirsch, Joost van Amersfoort a také Yarin Gal ve své práci dále posouvají bayesovské aktivní učení představením dávkové verze strategie BALD – batchBALD [KAG19].

3.6 Dávkové vzorkování

Výše uvedené kategorie strategií aktivního učení předpokládají výběr vždy právě jednoho prvku, který model vylepší nejvíce, což ale vede k frekventovanému učení. Pro některé modely tento přístup nepředstavuje problém, u modelů s vyšší náročností učení (např. neuronové sítě) by ale bylo výhodnější navzorkovat více dat a trénovat je dávkově. Již zavedené kategorie je možné zobecnit na výběr více vzorků: výběr selektivního vzorkování můžeme ukládat do mezipaměti, po jejímž naplnění model učíme na vybraných prvcích. Podobný přístup můžeme aplikovat na zbylé dvě kategorie. Tento naivní přístup však zavádí další problém: model je ve fázi vzorkování nejistý stále ve stejných vzorcích, je tedy velmi pravděpodobné, že mezipaměť bude redundantně zaplněna podobnými vzorky [Ren+20; Car+17].

Toto chování zachycuje obrázek 3.2, který porovnává výběr vzorků (černě) při použití strategie nejmenší jistoty, popsána výše, a dávkové strategie hodnocené dávkové vzorkování. Červené body reprezentují trénovací sadu a šedé zbytek datasetu. Body jsou vektorovým zastoupením vzorků datové sady IRIS (klasifikace obrázků) ve dvourozměrném prostoru. Použitým klasifikátorem je tzv. algoritmus k-nejbližších sousedů. Z obrázků je zřejmé, že dávkové vzorkování vybralo nesourodé vzorky, zatímco nejmenší jistota prvky, které si jsou mnohem podobnější.

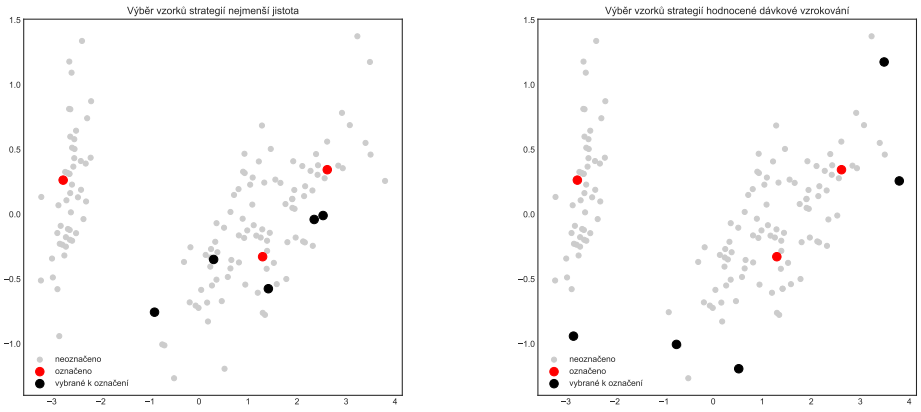
V přístupu dávkového vzorkování je tedy třeba dále zavést míru podobnosti vybíraného prvku a zbytku již vybraných vzorků.

3.6.1 Hodnocené dávkové vzorkování

Mezi významné zástupce strategií této kategorie patří *hodnocené dávkové vzorkování* (Ranked batch-mode sampling) navržený Cordosem et al. [Car+17]. Strategie se opírá o nejistotu modelu, hodnotu získanou například jednou ze strategií vzorkování založeného na nejistotě, a přidává k ní informaci o podobnosti s ostatními neoznačenými vzorky. Původní článek používá strategii nejmenší jistoty (viz sekci 3.2.1). Podobnost vzorku a označené množiny je v článku počítána jako maximum ze vzdáleností vzorku a všech prvků množiny. Vzdáleností je myšlena kterákoliv vektorová metrika, např. eukleidovská, manhattonská nebo Čebyševova. Strategie se dá zapsat pomocí vzorce:

$$s(u) = \alpha(1 - \Phi(u, \mathcal{L})) + (1 - \alpha)U(u),$$

$$\Phi(u, \mathcal{L}) = \mathbf{max}_{l \in \mathcal{L}}\{d(u, l)\},$$



(a) Vzorkování strategií nejmenší jistota

(b) Vzorkování strategií hodnocené dávkové vzorkování

Obrázek 3.2: Projekce datové sady IRIS do dvourozměrného prostoru a vyznačení některých prvků pro dávkovou a nedávkovou strategii. Obrázky pořízeny upraveným skriptem z [DH18, dokumentace knihovny – Ranked batch-mode sampling]

.

$$\alpha = \frac{|\mathcal{U}|}{|\mathcal{U}| + |\mathcal{L}|},$$

kde $\mathcal{U}$ a $\mathcal{L}$ jsou neoznačená a označená množina vzorků, U je některá ze strategií nejistoty (např. nejmenší jistota), d je funkce vzdálenosti vektorů (eukleidovská, cosinová...), u je zkoumaný vzorek, s je hodnota přidělena touto strategií a α je regulující parametr rovný poměru velikosti neoznačené sady k celkovému počtu vzorků. Místo Φ ze vzorce je možné použít jakoukoliv jinou metriku měřící vzájemnou podobnost. Nutno je podotknout, že modely učeny touto strategií také musí umět produkovat pravděpodobnostní distribuci nad třídami vzorků jako tomu bylo u strategií založených na nejistotě, jelikož je přímo používá.

3.6.2 Nejistota a diverzita

Autoři některých mladších strategií hovoří o dvou přístupech aktivního učení: nejistotě modelu a rozličnosti vzorků [YLB20; Ash+19]. Nejistota modelu může být shodná s heterogenními metodami založenými na nejistotě, diverzita je v nějakém smyslu rozdílnost vzorků (např. shluk po provedení k-means nebo k-means++ algoritmu). Strategie této kategorie, ač ne vždy takto literaturou označovány, se dají považovat za dávkové vzorkování.

3.6.2.1 Přístup zastupující množinou

O. Sener a S. Savarese ve svém článku [SS18] problém s diverzitou zkoumají nad modely konvolučních neuronových sítí. Na aktivní učení přihlížejí jako na problém hledání zastupující (reprezentativní) podmnožiny (angl. core-set problem): cílem je vybrat takovou podmnožinu vzorků, aby se celková výkonnost modelu změnila minimálně oproti celé množině. Autoři ukazují, že hledání zastupující množiny odpovídá k-Center problému, tedy výběru k středových vzorků takových, že největší vzdálenost vzorků od svého nejbližšího středu je nejmenší. Velikost k odpovídá počtu vzorků vybíraných v iteraci aktivního učení. Pro tento problém nabízí dvě řešení – žravý a robustní algoritmus. Výsledky těchto algoritmů se liší v optimalitě; k-Center je NP-těžký problém, není tedy znám algoritmus pro zjištění optima v rozumném čase. Tato strategie se nezaměřuje na heterogenitu.

3.6.2.2 Dávkové aktivní učení diverzními gradientními příznaky

Autoři strategie BADGE Jordan T. Ash et al. [Ash+19] zdůrazňují problémy metod založených jen na nejistotě a jen na diverzitě, ale také dávkových strategií. Konstatují, že úspěšnost těchto strategií závisí na velikosti dávky ve vztahu k celkovému objemu dat a na strukturovanosti datové sady. Za další problém označují závislost vybraných vzorků na výběru hyperparametrů modelu – tedy že malá změna v hyperparametrech může cílit k zásadní změně ve vybraných vzorcích k označení orákulem. S ohledem na tyto poznatky navrhují svou strategii aktivního učení určenou pro (hluboké) neuronové sítě.

Strategie se snaží aplikovat jak nejistotu modelu, tak diverzitu vzorků v dávce. Obě tyto vlastnosti se snaží vyjádřit touž metrikou – gradientem ztrátové funkce podle parametrů poslední (výstupní) vrstvy. Podobá se tak i metodám očekávané změny modelu. Gradienty je možné počítat s ohledem na libovolnou třídu, do nichž klasifikátor vzorky řadí:

$$g_x^y = \frac{\partial}{\partial W} l_{\mathrm{CE}}(f(x|\Theta), y),$$

kde W značí parametry výstupní vrstvy, x a y jsou popořadě vzorek a třída, $f(x|\Theta)$ je vektor pravděpodobností tříd podle klasifikátoru s parametry Θ pro vzorek x a l_{CE} značí ztrátovou funkci – křížovou entropii. Patrně $W \subseteq \Theta$. Z možných gradientů pro vzorek se operuje s tím pro nejpravděpodobnější třídu (podle modelu), tedy $g_x = g_x^{\hat{y}}$, $\hat{y} = \mathrm{argmax}_i f(x|\Theta)_i$. Nejistota modelu je pak vyjádřena normou tohoto gradientu a diverzitu zajišťuje výběr takových prvků, jejichž g_x vyjadřují různé směry. Aplikace algoritmu k-means++ nad těmito gradienty všech neoznačených vzorků ústí ve výběr množiny obsahující rozdílné prvky, v nichž si je model nejistý.

3.6.3 Diskriminativní aktivní učení

Výzkumníci D. Gissin a S. Shalev-Schwarz [GS19] přicházejí s netradiční strategií aktivního učení a řadí ji mezi dávkové přístupy. Autoři přinášejí do problematiky aktivního učení odlišný pohled od ostatních metod tím, že se jej snaží definovat jako problém binární klasifikace. Vycházejíce z teze, že pokud prvky neoznačené sady jsou nerozeznatelné od prvků označených, pak sada označených prvků je reprezentativní vzhledem k celé datové sadě, definují problém klasifikace původních vzorků do třídy „označen“, nebo „neoznačen“. Následně navrhují naučit (libovolný) klasifikátor – diskriminátor, aby mezi těmito přeoznačenými vzorky dokázal rozpoznávat třídu, a v každém kroku aktivního učení vybrat K prvků, u nichž si je diskriminátor nejvíce jistý jejich náležitostí do třídy „neoznačen“. Výběrem pak tyto prvky začnou náležet do třídy druhé. Výsledkem této metody by mělo být přerozdělení vzorků tak, aby si diskriminátor nebyl jistý u žádného prvku, což ústí ve splnění předpokladu teze.

3.7 Možnosti Implementace

Ve vlastních experimentech bude využita již existující implementace strategií a algoritmů aktivního učení. Tato sekce vyjmenuje některé možnosti výběru a zdůvodní volbu konkrétních.

Vlastní implementace. První možností je vlastní implementace zvolených strategií aktivního učení. Výhodou v tomto případě je možnost napsat rozhraní vyhovující skriptům experimentů s ohledem na všechny jeho parametry. Avšak vzhledem k režii tohoto přístupu a náročnosti implementace některých strategií aktivního učení nebyla tato možnost volena, neboť by přinesla příliš veliké časové náklady.

Knihovna modAL. Další možnou volbou je knihovna modAL vyvíjená doktorem Tivadarem Dankou. Knihovna je psaná v programovacím jazyce Python pro framework Scikit-learn. Obsahuje však také podporu pro Pytorch, framework pro strojové učení, který je využíván v experimentech. Knihovna poskytuje všechny výše popsané strategie založené na nejistotě, komisní vzorkování (také 3 strategie), ale také vzorkování založené na hustotě informace nebo dávkové vzorkování. Podporuje jak výběrové, tak selektivní přístupy. Kromě aktivního učení klasifikátorů také umožňuje využít AL pro regresní modely [DH18]. Vzhledem k jednoduchosti rozhraní a používání knihovny byla volena tato možnost jako jedna z implementací aktivního učení v experimentech.

Knihovna Small-Text. Tato knihovna je vyvíjená skupinou lidí z Leipzigské univerzity a Institutu aplikované informatiky v Leipzigu [Sch+21]. Narozdíl od knihovny modAL se zaměřuje zejména na klasifikaci v oblasti NLP a utilizuje state-of-the-art dotazovací strategie. Poskytuje API pro modely typu transformer ze služby HuggingFace, na něž se zaměřuje i tato práce. Pokročilé strategie, které tato knihovna poskytuje, jsou důvodem k využití i této implementace v experimentech.

3.7.1 Studený začátek

Důležitým rozhodnutím při vývoji experimentů bylo stanovit, zda v každé iteraci (epoše) aktivního učení má být již trénovaný model z předchozích epoch pouze doučen na aktivním učením vybraných vzorcích (teplý start), anebo zda má být model inicializován na počáteční stav a naučen na celé označené množině (studený začátek). Pro experimenty byl zvolen druhý přístup, jelikož bylo prokázáno, že teplý začátek může poškodit zobecňovací schopnosti modelu [AA19].

Experimenty 4

4.1 Vymezení pojmů

Pro jednoznačnost následujícího popisu experimentů bude potřeba explicitně vymezit některé další pojmy, které budou nyní definovány.

Epocha aktivního učení. Také AL epocha. Jedna obrátka procesu aktivního učení. Zahrnuje výběr vzorků, označení orákulem a trénink modelu na všech označených datech. Po tréninku je vyhodnocena přesnost modelu.

Cyklus aktivního učení. Konečná posloupnost po sobě jdoucích epoch aktivního učení. Po jeho skončení je experiment dokončen.

Krok aktivního učení. Počet vzorků vybíraných v AL epoše. Zároveň také rozdíl velikostí trénovacích datových sad dvou po sobě jdoucích epoch aktivního učení. V každé epoše aktivního učení se může lišit.

Ukončovací podmínka. Podmínka, která terminuje cyklus AL. Může jí být například pevný počet epoch aktivního učení, přesnost modelu nebo velikost označených dat. V rámci experimentů této práce je ukončovací podmínkou vyčerpání kroků AL.

Vzorkovací krok. Okamžik, kdy jsou navzorkovány některé vlastnosti modelu – například přesnost na trénovací/validační datové sadě nebo jiné metriky. V experimentech je čas reprezentován právě vzorkovacím krokem, neboť se jedná o diskrétní vzorkování nějaké funkce (metriky) v závislosti na čase.

Trénovací epocha. Epocha v běžném chápání z pohledu strojového učení. Jeden průchod tréninku modelu celou trénovací datovou sadou. Trénink typicky sestává z mnoha epoch. Během jedné epochy jsou (při učení neuronových sítí) trénovací data podávána po dávkách.

Předčasné přerušení. Metoda zastavení tréninku modelu předtím, než dojde k přeučení. Během tréninku je např. sledována jistá metrika a po určité době, během níž nedošlo ke zlepšení, je trénink zastaven. Anglicky early stopping.

Východisko. Slovem východisko je v této práci přeloženo anglické „baseline". Jedná se o základní, „nulový" nebo také referenční stav/experiment. V této práci to konkrétně jsou experimenty, na něž není aplikováno aktivní učení, případně je aplikována strategie „náhodného vzorkování", která vybírá vzorky k označení náhodně.

Následující pseudokód vyjadřuje základní algoritmus aktivního učení obsahující výše definované pojmy:

```
AL cyklus:
  pokud platí Ukončovací podmínka: skonči.
  AL epocha:
     vyber podle AL strategie tolik vzorků, kolik je Krok AL
     označ vybrané vzorky
     trénink modelu:
       Trénovací epocha:
          opakuj do pokrytí celé trénovací sady:
            natrénuj model na dávce vzorků
            navzorkuj metriky na trénovací sadě
            po několika dávkách navzorkuj na evaluační sadě
          pokud Předčasné přerušení učení neukončí:
            opakuj epochu
     navzorkuj metriky na testovací sadě
     jdi na začátek AL cyklu
```

4.2 Typy a cíle experimentů

4.2.1 Experimenty východisek

Tyto experimenty popisují chování tréninku modelů bez aktivního učení. Model je naučen na celé datové sadě. Účelem pokusů je najít graf závislosti vybraných metrik modelu na čase trénování (vyjádřený vzorkovacím krokem). Výsledný graf tedy vypovídá o průběhu učení a slouží jako východisko pro experimenty s aktivním učením. Porovnáním východiska a aktivního učení dokážeme vyvodit některé závěry – zda je AL skutečně vybírá výhodné prvky (dosahuje podobné přesnosti na menší datové sadě) atp.

Tyto experimenty jsou parametrizovány následujícími hodnotami.

- Velikost trénovacích dávek: kolik vzorků vstupuje do učícího algoritmu modelu naráz (angl. batch size). Trénovací epocha spustí algoritmus tolikrát, kolika je roven poměr velikosti celé datové sady k velikosti trénovacích dávek.
- Předčasné přerušení: hodnota ano/ne, udává, zda má být předčasné přerušení použito.
- Počet trénovacích epoch: pokud není předčasné ukončení použito, udává celkový počet trénovacích epoch.
- Poměr trénovací sady k evaluační: udává poměr v intervalu $(0, 1)$ velikostí trénovací sady a sady evaluační. Původní trénovací sada je tímto poměrem rozdělena na skutečnou trénovací sady a sadu vyhodnocovací. Pokud datová sada však obsahuje vlastní evaluační podmnožinu, je využita tato.
- učící koeficient (angl. learning rate): základní učící koeficient pro lineární plánovač.
- Datová množina: kterou datovou množinu použít. Viz sekci 4.3.
- Model a další pro něj specifické hyperparametry: Viz sekci 4.4.
- Zařízení: výběr, zda má trénování probíhat na procesoru stroje, nebo na grafické kartě.

4.2.2 Experimenty aktivního učení

Tyto experimenty sledují průběh učení modelu s využitím aktivního učení. Ve skutečnosti se však jedná o více modelů – každá epocha AL produkuje naučené modely stejného typu, pouze s různou velikostí tréninkové datové sady. Pro každý z modelů jsou vyhodnocovány metriky na tréninkové i validační sadě. Způsob tréninku modelu závisí na použité knihovně strategií AL. Nicméně, tyto experimenty můžeme parametrizovat stejnými hodnotami jako experimenty východisek, k nimž bude třeba přidat krok (případně kroky, viz dále) aktivního učení a konkrétní strategii. Ukončovací podmínka je buď označení všech prvků datové sady, nebo vyčerpání kroků aktivního učení. Počet kroků (AL epoch) a jejich velikost je určena „iterační strategií“, která může být zadána jako:

- počet AL epoch – kroky jsou pak ekvidistantní a tak veliké, aby v počtu epoch pokryly celou sadu, nebo
- velikost kroku udaná počtem vzorků nebo poměrem (procenty) k celkovému počtu vzorků – kroky jsou opět ekvidistantní a AL epoch je tolik, aby byla pokryta celá sada, nebo

- seznam velikostí kroků udaných obdobně jako výše – epoch aktivního učení je právě tolik, jak dlouhý je seznam, a velikosti kroků odpovídají pořadí v seznamu. Tímto způsobem nemusí být pokryta celá sada.

Na rozdíl od experimentů východisek máme po skončení celého AL cyklu k dispozici mnoho vyhodnoceních na testovací množině (právě tolik, kolik bylo AL epoch) a můžeme sledovat nárůst (či pokles) nějaké metriky mezi jednotlivými naučenými modely. Cílem těchto experimentů je pak graficky znázornit metriky získané na testovací sadě všech modelů v závislosti na velikosti trénovací sady. Porovnáním těchto výsledných grafů experimentů s různými strategiemi aktivního učení je možné zjistit, která ze strategií funguje lépe. Je dále možné zkoumat a odhalovat závislosti ostatních parametrů experimentu v prostředí aktivního učení. Porovnáním s východiskem je také možné odhalit přínos aktivního učení – hodnoty blížící se optimu východiska by se při aktivním učení měly v grafech objevit při menší velikosti trénovací sady. Naopak by aktivní učení nemělo dosahovat výsledků lepších, neboť informace obsažená v celé sadě je stejná.

Důležité je zmínit, že v experimentech aktivního učení činíme úkrok od reálného nasazení. V praxi je k dispozici pouze neoznačená množina dat, z níž AL vybírá prvky k označení orákulem. V experimentech ale používáme již označené celé datové sady, nicméně modelům sdělujeme správný výsledek jen na její podmnožině. Doplněk této podmnožiny modelu představujeme jako neoznačenou množinu, z níž jsou prvky vybírány. Díky tomuto postupu můžeme simulovat reálný proces, aniž bychom potřebovali při testování skutečného anotátora a zvyšovali cenu experimentů.

4.3 Datové sady

Experimenty je možné provádět nad různými datovými sadami. Sady se klasifikují podle typu problému, na který model připravuje. Vzhledem k tématu práce skripty podporují pouze textové sady. Je možné použít libovolné datové sady z databáze služby HuggingFace zadáním přesného názvu sady v databázi. Níže popsané jsou „očekávané“ sady[1].

ČSFD. Česká sada na analýzu sentimentu, texty jsou komentáře z internetových stránek Československé filmové databáze. Komentáře klasifikuje do tří tříd: pozitivní, neutrální a negativní. Trénovací část sady obsahuje přes 80 000 označených komentářů. Tato sada se nenachází v databázi HuggingFace a je dostupná v datech projektu.

[1] pro jiné není chování definováno

České novinové články. Další česká sada, která obsahuje úryvky novinových článků z šesti portálů. Datová sada obsahuje několik typů textů, které je možné klasifikovat (titulek, stručný obsah, úryvek atd.), a také údaj o datu. Klasifikovat lze v rámci několika kategorií, jako např. portál, pohlaví autorů či kategorii. V experimentech model třídí úryvky článků do kategorií, kterých je 26. Trénovací sada obsahuje přes 1 300 000 vzorků. Dříve ji bylo možné stáhnout z databáze HuggingFace, z níž však v průběhu tvorby této práce byla odstraněna. Nyní se nachází v datech projektu.

Recenze produktů z internetového obchodu Mall.cz. Tato sada obsahuje 24 000 českých (mezi nimi několik slovenských) recenzí, které řadí jako kladné, neutrální nebo negativní – jedná se tedy opět o analýzu sentimentu. V databázi HuggingFace je možné sadu dohledat jako:

```
fewshot-goes-multilingual/cs_mall-product-reviews
```

IMDB. Anglická obdoba sady ČSFD, komentáře pochází z anglické internetové databáze filmů. Obsahuje 25 000 trénovacích i testovacích vzorků, které třídí buď jako pozitivní, nebo negativní (oproti ČSFD nerozpoznává neutrální sentiment). Sada je dostupná službou HuggingFace – název je stejný, tedy `imdb`.

DBpedia ontology. Anglická sada čerpající ze znalostní sady DBpedia, verze 2014. DBpedia je strukturovaná znalostní báze tvořená články z Wikipedií mnoha jazyků [Leh+14; ZZL15]. Použitá datová sada je výňatkem čtrnácti nepřekrývajících se tříd označujících téma článku. Ke každé třídě bylo vybráno 45 000 vzorků pro učení nebo testování a každý vzorek obsahuje titulek a abstrakt článku na Wikipedii. Trénovací sada má celkem 560 000 vzorků. Sada obsahuje pouze anglické texty [ZZL15]. V databázi HuggingFace je možné sadu najít pod názvem `dbpedia_14`.

4.4 Modely

Experimenty je možné vést pro modely různých či podobných architektur. Výrazným rysem také je, zda model rozumí konkrétnímu jazyku anebo zda je jazykově nezávislý. Skripty experimentů umožňují výběr ze šesti modelů – čtyři z nichž jsou tzv. transformery. Tyto jsou popsány v následujících dvou podsekcích. Skript je však možné spustit pro libovolný model z databáze služby HuggingFace zadáním úplného názvu modelu v databázi. Správné chování skriptu však pro jiné modely není zaručeno.

U modelů z HuggingFace lze před spuštěním experimentů určit, zda se mají načíst již předtrénované, anebo neinicializované. Ostatní modely se vytváří neinicializované.

4.4.1 Modely závislé na jazyku

V experimentech je možné volit mezi čtyřmi modely závislými na jazyce. Dva z nich vycházejí z architektury ELECTRA small. Modely ELECTRA jsou odvozeny od současně nejpoužívanějších modelů – transformerů – konkr. části encoder. Tyto modely jsou trénovány přístupem známým jako GAN, volně přeloženo jako „generativní soupeřivé sítě“, při němž jsou trénovány dvě neuronové sítě. V případě modelů ELECTRA jsou jimi generátor, model trénovaný na úloze rekonstrukce maskovaného jazyka, a diskriminátor, jehož úlohou je zjistit, která slova (tokeny) byla generátorem doplněna. Proces učení se však od typického GAN dále mírně liší [Cla+20]. Výsledkem učení je diskriminátor, tedy model předučený na rozeznávání falešných (zaměněných, generátorem doplněných) slov nebo tokenů.

Zbylé dva modely jsou architekturou model RoBERTa base, taktéž transformery. RoBERTa je shodná s modelem BERT až na pár změn, kterými jsou: delší trénování modelu na delších sekvencích, není trénována pro úkol predikce další věty a dynamické obměny masky pro úlohu maskovaného modelování jazyka. Díky změnám dosahuje lepších výsledků než původní BERT [Liu+19]. V experimentech je možné volit mezi modelem z původního článku [Liu+19] a RoBERTou naučenou pouze na českých datových sadách – model RobeCzech [Str+21].

Modely ELECTRA přístupné v experimentech jsou: model z původního článku [Cla+20], předtrénovaný na anglickém korpusu, a český model Small-E-Czech, trénovaný na české datové sadě DaReCzech pro úlohu hodnocení relevance textu, vyvinutý firmou Seznam.cz [Koc+21]. Ke všem čtyřem modelům se přistupuje přes rozsáhlou databázi Huggingface Hub.

Pro validní experimentování je tedy potřeba testovat modely ELECTRA small a RoBERTa base pouze na anglických sadách. Model Small-E-Czech je teoreticky také možné testovat na sadách anglických, jelikož tento model vychází z předtrénovaného ELECTRA small modelu, ale experimenty jsou primárně určeny pro české sady. Model RobeCzech by měl být zkoumán na českých sadách.

4.4.2 Modely nezávislé na jazyku

Z modelů nezávislých na jazyku je možné použít „Embedding Averaging“ nebo také „Mean“ model přímo zakódovaný ve skriptech experimentů nad frameworkem PyTorch. Tento jednoduchý model počítá průměry vektorů vnoření slov (angl. word embedding) a průměry transformuje skrytou lineární vrstvou s aktivační funkcí ReLU a výstupní lineární vrstvou. Během učení (a při Monte-Carlo vzorkování) je také použit tzv. dropout, strategie pro zlepšení výsledků „vynecháváním“ některých neuronů. Pravděpodobnost vynechání je parametrizovatelná.

Tento model primárně slouží jako kontrast chytřejších modelů dostupných v experimentech pro srovnání přínosu aktivního učení mezi modernějšími přístupy

v NLP a primitivními.

Model je možné parametrizovat následujícími hodnotami:

- velikost vnoření slov: dimenze vektorů vnoření slov,
- pravděpodobnost vynechání: s jakou pravděpodobností dropout vrstvy vynechají určitý neuron,

Dalším modelem nezávislým na jazyku, který je v experimentech možné použít, je tzv. „náhodný les“ (Random forest). Tento model agreguje několik rozhodovacích stromů (v experimentech 200), které náhodně naučí na vstupní sadě dat. Třída vzorku při inferenci je rozhodnuta na základě výstupu každého stromu. Způsobů, jak náhodně učit stromy a jak zvážit rozhodnutí stromů při predikci, je několik; v experimentech je použita implementace knihovny scikit-learn [Ped+11]. Učení stromů probíhá tzv. bootstrapováním (výběrem s opakováním) z množiny všech trénovacích vzorků a rozhodnutí o třídě inferovaného vzorku odpovídá průměru výstupů ze stromů lesa. Model primárně slouží k porovnání výsledků aktivního učení mezi neuronovými sítěmi a modelem zcela odlišné architektury. K extrakci příznaků z textu je použita metoda TF-IDF.

4.4.3 Trénink

K učení modelů se v experimentech využívá primárně dvou přístupů pro modely Mean a transformery (dále torch modely) a dvou pro náhodný les. Konkrétní použitý přístup je dán typem experimentu nebo volbou strategie aktivního učení. Experimenty východisek učí torch modely pomocí Trainer API služby HuggingFace. Jedná se o vysoce parametrizovatelný modul usnadňující učení modelů – hlavně transformerů, ale funguje i pro modely frameworku PyTorch, neboť jej HuggingFace transformery také využívají.

V experimentech aktivního učení záleží na volbě aktivní strategie a typu modelu. Knihovna modAL (viz sekci 3.7) neposkytuje z pohledu autora této práce vyhovující podporu pro torch modely – učení modelu je tedy přenecháno opět Trainer API. Modely frameworku Scikit-learn však podporovány jsou, náhodné lesy použité v těchto experimentech se strategiemi knihovny modAL jsou učeny prostředky této knihovny. Na druhé straně knihovna Small-Text podporuje Scikit-learn i torch modely, a tak je učení ponecháno implementaci v knihovně.

Torch modely jsou učeny s optimizátorem AdamW a lineárním plánovačem s počátečním učícím koeficientem $2 \cdot 10^{-5}$.

4.5 Návrh Experimentů

Zkoumané experimenty aktivního učení zahrnují z výše popsaných modelů transformery (Small-E-Czech, RoBeCzech, ELECTRA small, RoBERTa base) a náhodné lesy. Mean modely nebyly zahrnuty, jelikož v současnosti není jejich architektura obvyklá. Z datových sad byly vybrány IMDB, ČSFD, recenze na Mall.cz a DBpedia Ontology. Jazykově závislé modely – tedy všechny transformery – byly trénovány jen na sadách „svého“ jazyka (ČSFD a recenze / IMDB a DBpedia) a náhodné lesy na všech čtyřech. Ze strategií byl vybrán vzorek dostupných z použitých knihoven (viz sekci 3.7) s ohledem na rozšíření, pokročilost a zastoupení různých kategorií popsaných v kapitole 3. Z knihovny modAL byly zkoumány strategie vzorkování na základě entropie a největšího rozdílu (pod klíčovými slovy `entropy` a `margin`) a hodnocené dávkové vzorkování (`unc_batch` odvozené z Uncertainty Batch sampling, jak se metoda v knihovně nazývá). Tyto strategie jsou testovány pouze na náhodných lesích (viz dále). Výběr z knihovny Small-Text zahrnuje výběr vzorků zastupující množinou (`greedy_coreset`), strategie BADGE a BALD, diskriminativní aktivní učení (`discriminative`) a znovu entropii pro transformer modely. Jako východisko byla přidána také strategie náhodného vzorkování (`random`), která vybírá prvky k označení náhodně.

Některé strategie navíc požadují od modelů zvláštní vlastnosti – z použitých to jsou bayesovské aktivní učení (BALD; vyžaduje Monte Carlo vynechávání), výběr vzorků zastupující množinou a dávkové aktivní učení diverzními gradientními vnořeními (BADGE; poslední dvě strategie vyžadují vektorovou reprezentaci z poslední vrstvy neuronové sítě). Těmito vlastnostmi však náhodné lesy neoplývají a proto pro ně nebyly měřeny. Místo nich jsou však testovány vybrané strategie knihovny modAL.

U strategie diskriminativního aktivního učení byl za diskriminátor volen stejný model jako cílový. Liší se pouze počtem tříd (vzhledem k povaze úlohy dvě) a tréninkem na 150 epoch bez předčasného přerušení – aby většinu času experimentu nespotřeboval jeho trénink.

Některé parametry experimentů jsou nastaveny pro všechny stejné: předčasné přerušení je zapnuto, transformery se vytvářejí předtrénované a poměr trénovací sady k validační je 90 % – až na sadu Recenzí produktů Mall.cz, jež poskytuje vlastní validační prvky. Velikosti trénovacích dávek (angl. batch sizes) závisí na velikosti modelu – u transformerů typu ELECTRA je nastavena na 40 a pro RoBERTy na 10 vzorků. Transformery jsou vždy trénovány s využitím GPU, zatímco náhodné lesy jsou učeny pouze na procesorech. Iterační strategie se v průběhu zavádění experimentů měnila. Zpočátku byl počet epoch aktivního učení nastaven na 48 s ekvidistantním krokem AL. Později pro lepší navzorkování kritické části – na malých velikostech počtu vzorků – byla uplatněna strategie vzorkování po 1 % sady ve 12

krocích, dalších 6 kroků po 5 % a poslední tři kroky po 10 %. Celkové navzorkování pozdějších experimentů je tedy 72 % trénovací sady.

Experimenty byly na metacentru spouštěny dvojím způsobem: celý trénink a aktivní učení v jedné úloze, nebo přes několik úloh, kde každá zahrnuje trénink na již označených datech, evaluaci modelu a výběr dalších vzorků. Ve druhém způsobu je počet úloh roven počtu kroků aktivního učení. Všechny experimenty s iterační strategií 48 ekvidistantních kroků a několik experimentů s druhou popsanou iter. strategií byly spuštěny prvním způsobem, zbytek experimentů (všechny s druhou iterační strategií) druhým způsobem.

Výsledná data byla nahrána do systému Weights and Biases, pomocí nějž byly vygenerovány výsledné grafy experimentů.

4.6 Zaznamenané potíže

Během vývoje experimentů nastaly některé komplikace, jež není možné nebo je nepraktické ošetřit.

První opakující se komplikací je chyba připojení ke službě Weights and Biases, kvůli níž experiment selže již na začátku. Tato chyba však nenastává ve většině případů a neobjevuje se konzistentně s nastavením experimentů – problém se nachází v připojení stroje k síti. Řešením může být opakované spuštění experimentu, vypnutí služby (resp. přepnutí do offline režimu – experimenty je pak třeba synchronizovat příkazem manuálně) nebo spouštění na strojích s dobrým připojením. V různých fázích spouštění experimentů byla volena všechna řešení.

Experimenty transformerů někdy předčasně končí kvůli nedostatečné paměti grafické karty – zejm. modely RoBERTa a RoBeCzech. K vyřešení problému může dopomoct menší velikost dávky při tréninku, tento přístup však nepřinesl výsledky vždy[2]. Jiné řešení může tkvět v lepší grafické kartě – a jelikož se experimenty počítají v gridovém prostředí metacentra, má smysl se tímto řešením zabývat. Před zařazením úlohy do čekací fronty je možné zažádat o specifický cluster (skupina strojů s podobnými vlastnostmi) a lze tedy požadovat, aby byla úloha spuštěna na výkonnějším stroji. Toto však přináší další komplikace: výkonné stroje jsou často obsazené a fronty na ně dlouhé – ke spuštění úlohy může dojít i za několik dní.

Omezujícím faktorem je také limit doby běhu úlohy na metacentru pro GPU uzly. Úlohy mohou žádat nejvýše o 24 hodin běhu. Pro některé výpočetně náročnější strategie aktivního učení však jeden den není dostačující a úloha je ze systému odstraněna před navzorkováním všech kroků AL. Tento problém částečně řeší spouštění jednoho experimentu ve více úlohách (viz poslední odstavec sekce 4.5).

[2] pro RoBERTa modely byla testována dávka jen o 5 vzorcích – chyba však stále přetrvávala.

Znepříjemňující komplikací bylo odstranění datové sady České novinové články z databáze HuggingFace. Místo ní byla vybrána nová datová sada – Recenze produktů Mall.cz. Sadu novinových článků se autorovi této práce později podařilo obnovit, do experimentů však opět zahrnuta nebyla.

Výsledky a Diskuze

5

5.1 Grafy

Výsledné grafy byly vygenerovány pomocí služby Weights & Biases. Křivky byly vyhlazeny exponenciálním klouzavým průměrem s faktorem $\alpha = 0,4$. Grafy často zobrazují pouze část navzorkovaných dat z důvodu přehlednosti a lepší čitelnosti. Rozsah části grafu byl volen tak, aby zbytek celkového grafu odpovídal posledním několika vzorkům (tj. nepříliš se od nich měnil). Důležité je také poznamenat, že svislá osa (znázorňující přesnost modelu) nezačíná v nule ze stejného důvodu – čitelnosti. Rozsah osy je volen tak, aby na zvolené části grafu obsáhl všechny navzorkované přesnosti. U této osy je však dále důležité zaměřit se i na rozptyl hodnot. Podobné platí i u vodorovné osy (velikost trénovací sady), jelikož každá datová sada obsahuje rozdílný počet dat a tedy i experimenty mají rozdílný krok aktivního učení.

Grafy zobrazují střední hodnotu (lomená čára) a směrodatnou chybu (poloprůhledná oblast) z více realizací shodných experimentů.

Grafy se nacházejí v příloze A. Interaktivní grafy je možné zobrazit ve webovém rozhraní služby Weights and Biases jak pro experimenty východisek[1], tak pro experimenty aktivního učení[2].

5.2 Diskuze

Příloha A předkládá grafy, které poskytují informace, z nichž můžeme vyvodit některé závěry.

5.2.1 Výpočetní složitost některých strategií

Jedním z fenoménů, jež je možné na grafech experimentů východisek odhalit, jsou křivky navzorkované na méně velikostech trénovací sady oproti zbytku křivek v grafu. Nejvíce konzistentně se takto chovají strategie hodnoceného dávkového

[1] `https://api.wandb.ai/links/jamlady/2871w8u7`
[2] `https://api.wandb.ai/links/jamlady/rzr1w66v`

vzorkování (`unc_batch`, červeně) a diskriminativní aktivní učení (`discriminative`, šedě). Toto chování však není neočekávané, jelikož prostřední metacentra umožňuje maximální dobu úlohy jeden den a obě metody jsou výpočetně náročné – hodnocené dávkové vzorkování v každém dotazování porovnává všechny vzorky v neoznačené množině a diskriminativní učení trénuje celý model (shodný s výsledným). Z posledního odstavce sekce 4.5 plyne, že pouze pro trénink modelu a výběr dalších vzorků k označení je potřeba více než den výpočetního času, nikoliv pro celý proces. Tento poznatek dále dotváří představu o složitosti těchto strategií. Při detailnějším pohledu na strategii diskrimintaivního učení lze vypozorovat, že výpočetní náročnost (ve smyslu počtu navzorkovaných kroků AL) je vyšší pro větší modely (RoBERTa, RoBeCzech) a větší datové sady (DBpedia, ČSFD). V některých případech dokonce ani nedošlo na dokončení prvního výběru (RoBeCzech – ČSFD, RoBERTa – DBpedia), zatímco na menších modelech a sadách (např. Electra – IMDB) stačí pokrýt skoro celý zobrazený rozsah. Obdobný jev lze pozorovat také u strategie hodnoceného dávkového vzorkování, v rámci experimentů ji lze však porovnávat pouze na různých sadách, jelikož model je neměnný.

Diskriminativní aktivní učení navíc téměř nikdy na navzorkovaném rozsahu nepřekoná jiné strategie. V některých případech je porovnatelná s náhodným vzorkováním a v jiných horší. Tyto závěry však není možné obecně aplikovat na tuto strategii, neboť její chování může výrazně ovlivňovat volba diskriminátoru. Strategie hodnoceného dávkového vzorkování většinou poskytuje lepší výsledky oproti náhodnému vzorkování, nepřekonává však jiné strategie.

V některých grafech vykazuje „neupočitatelnou“ náročnost také strategie zastupující množinou (`greedy_coreset`, tmavě modrá), v případě experimentu Electra – DBpedia navíc tato strategie nebyla z důvodu čitelnosti zahrnuta do grafu. Neupočitatelné jsou pak téměř všechny strategie v případě experimentu RoBERTa – DBpedia, a to i při prvním dotazování. Lze se domnívat, že příčinou je velikost modelu i sady.

5.2.2 Přínos aktivního učení

V dalším se zaměřme na celkový přínos aktivního učení. Aktivní učení se napříč všemi datovými sadami a modely chová nekonzistentně; např. v některých případech překonává náhodné vzorkování, jinde nelze o přínosu rozhodnout. Je tedy potřeba učinit rozbor menších skupin. Experimenty rozčleňme podle datových sad; budeme tak schopni rozlišovat účinnost pro různé modely ve stejném prostoru dat.

Sada DBpedia Ontology je největší použitou sadou v experimentech. V případě malého transformer modelu Electra poskytuje aktivní učení nestálé výsledky, zejména strategie entropie, jejíž nestálost lze vypozorovat i u zbylých dvou modelů a také v dalších grafech. Strategie BALD (zeleně) a BADGE (žlutě) jsou mnohem kon-

zistentnější, avšak nedosahují takových přesností jako náhodné vzorkování, které je navíc ještě stabilnější než BALD. U modelu RoBERTa není možné kriticky zvážit přínos aktivního učení, jelikož skoro všechny strategie AL trvaly příliš dlouhou dobu. U náhodného lesa na několika prvních krocích AL není možné rozhodnout o úspěšnosti, po několika krocích je však stabilně nejpřínosnější strategie založená na největším rozdílu. Pro tento model je aktivní učení přínosné. Všechny modely se však značně blíží přesnostem svých experimentů východisek i na zobrazených počtech anotovaných vzorků, které dosahují zlomku celého rozsahu datové sady.

U malé datové sady IMDB ve všech případech křivky strategií převyšují náhodné vzorkování, je však vhodné zaměřit se na další úkazy. U modelu Electra jsou rozdíly mezi strategiemi patrné již po málo krocích (s ohledem na směrodatnou chybu). Graf však zachycuje poměrně úzký pás hodnot přesnosti oproti jiným, absolutní rozdíl mezi křivkami tedy není tolik markantní. Model RoBERTa je omezen ještě užším rozsahem přesnosti, avšak aktivní učení na něm modelu spěje. Modely se přesností pohybují u maxim experimentů východisek zhruba v polovině datové sady (nejsou horší o více než 2 % na přesnosti).

Sada recenzí produktů na Mall.cz je obdobně veliká jako sada IMDB, narozdíl od ní však třídí vzorky do tří kategorií. Vliv aktivního učení na model Small-E-Czech u této sady nelze kvůli velikostem směrodatných chyb diskutovat. Na posledních zobrazených velikostech označené sady náhodné vzorkování dokonce poskytuje lepší výsledky. U modelu RoBeCzech zhruba po 4000 označených vzorcích mírně překonává náhodné vzorkování strategie BALD, rozdíly však po 9000 označených vzorcích mizí. Oproti předchozím dvěma sadám transformer modely na zobrazeném rozsahu (zhruba polovina sady) nedosahují tak dobrých výsledků v porovnání s experimenty východisek. Liší se až o 4 % v přesnosti.

U datové sady ČSFD je přínos AL poznatelný u všech modelů. Pro model Small-E-Czech je zprvu přínosnější náhodné vzorkování, avšak po označení asi 7000 vzorků dosahuje vyšších přesností strategie BADGE, od 16000 označených převyšují náhodné vzorkování také ostatní strategie. Model RoBeCzech těží z aktivního učení podobně jako Small-E-Czech, lepší přesnosti jsou však u AL (strategie BALD) pozorovatelné dříve. U obou modelů se u pozdějších velikostí označené sady náhodné vzorkování s ostatními strategiemi setkávají. Grafy transformer modelů objímají zhruba třetinu velikosti datové sady, níž se oproti maximu z experimentů východisek liší až o 6 %. Model náhodný les v této skoro třetině díky aktivnímu učení dosahuje přesnosti pouze o 0,5 % horší, bez něj (náhodné vzorkování) až o 3 %.

Dále se můžeme zaměřit na jednotlivé strategie použité pro modely typu transformer (vzhledem k tématu práce). Strategie založená na zastupující množině je každém experimentu převýšena ostatními strategiemi, podobně jako diskriminativní aktivní učení popsané v minulé podsekci. Zbylé tři strategie – BADGE, BALD a založená na entropii – se v různých experimentech chovají různě. BALD a entropie

dokáží náhodné vzorkování znatelně převýšit (RoBERTa – IMDB, Small-E-Czech – ČSFD, RoBeCzech – obě sady), ale také podhodnotit (Small-E-Czech – Mall reviews). U strategie BADGE je možné pozorovat obdobné s tím rozdílem, že se zdá více držet podobných hodnot jako náhodné vzorkování a tedy neúspěch této strategie není tolik penalizující jako u předchozích strategií (Small-E-Czech – obě sady, Electra – IMDB).

5.2.3 Zhodnocení

Z grafů a diskuze vyplývá, že úspěch aktivního učení silně závisí na datové sadě a modelu. Největšího úspěchu dosahuje pro model náhodného lesa, avšak toto není pro cíle práce významné a pouze potvrzuje, že aktivní učení může být přínosné.

Nejpřínosnějším závěrem z diskuze by byla odpověď na otázku: kterou strategii aktivního učení volit při trénování modelu? Tato odpověď však z experimentů neplyne, ba dokonce nedokážeme ani odpovědět, zda je aktivní učení vůbec vhodné použít. To silně závisí na kontextu učení – zejména na zmíněné sadě a modelu. Nejen v tomto ohledu je potřeba dalšího výzkumu efektivity aktivního učení. Z výsledků této studie tedy nelze vyvodit přesný postup k volbě použití aktivního učení. Pokud obstarání anotací dat není příliš nákladné, nejvhodnějším způsobem se zdá být nepoužití aktivního učení, tj. anotovat všechna trénovací data, nebo jejich podmnožinu a simulovat tak náhodné vzorkování. Cena za anotace tak může být (logicky) vyšší než při použití náhodného vzorkování v paradigmatu AL (inkrementální přidávání anotací do ukončovací podmínky – např. požadovaná přesnost), ta ale může být převážena cenou za čas a prostředky k učení modelu, neboť aktivní učení trénuje mnoho modelů (počet AL epoch). Díky porovnávání s experimenty východisek také můžeme vyvodit, že výhodné může být použití náhodného vzorkování v paradigmatu aktivního učení s hrubým krokem AL (dotazování mnoha vzorků v každé epoše), jelikož se v experimentech modely na některých sadách dokázaly naučit klasifikovat obstojně. Výhodou tohoto přístupu je možnost ukončení před označením celé množiny a potlačení učení tolika modelů (s větším krokem AL je trénováno méně modelů). Také je možné použít i jiné strategie s hrubým krokem, jejich přínos oproti náhodnému vzorkování by však potřeboval lépe prověřit, jelikož experimenty této práce testovaly menší počty dotazovaných vzorků.

V případě, v němž je známo, že cena všech anotací převyšuje cenu učení mnoha modelů, má již jistě smysl uvažovat o rozsáhlejším aktivním učení. Volba strategie však tak jednoznačná není. Z výsledků experimentů této práce vyplývá, že vhodnými kandidáty mohou být náhodné vzorkování, jelikož dotazování není výpočetně náročné a v případech, kdy jej převyšují ostatní strategie, nezaostává o více než 2 % v přesnosti a také dokáže být stabilnější. Dále se nabízí strategie založená na entropii, která také není náročná, ale hrozí horšími výsledky. Podobně se chovají také

strategie BADGE nebo BALD, ty ale spotřebují mnohem více výpočetních zdrojů. Strategie založená na největším rozdílu by zasluhovala lépe prověřit, jelikož u modelu náhodného lesa téměř nikdy nepodhodnotila strategii založenou na entropii. Dalším zajímavým přístupem může být kombinace existujících strategií (např. různé části z nově dotazovaných vzorků vybrat jinou strategií), avšak žádná taková studie se nedostala do autorova povědomí.

Závěr 6

Úkolem této bakalářské práce bylo prozkoumat existující metody aktivního učení zejm. pro moderní modely používané ve strojovém učení se zaměřením na zpracování přirozeného jazyka, vybrané metody otestovat na modelech, změřit jejich úspěšnost oproti náhodnému výběru vzorků pomocí automatizovaných experimentů a kriticky jejich výsledky zhodnotit.

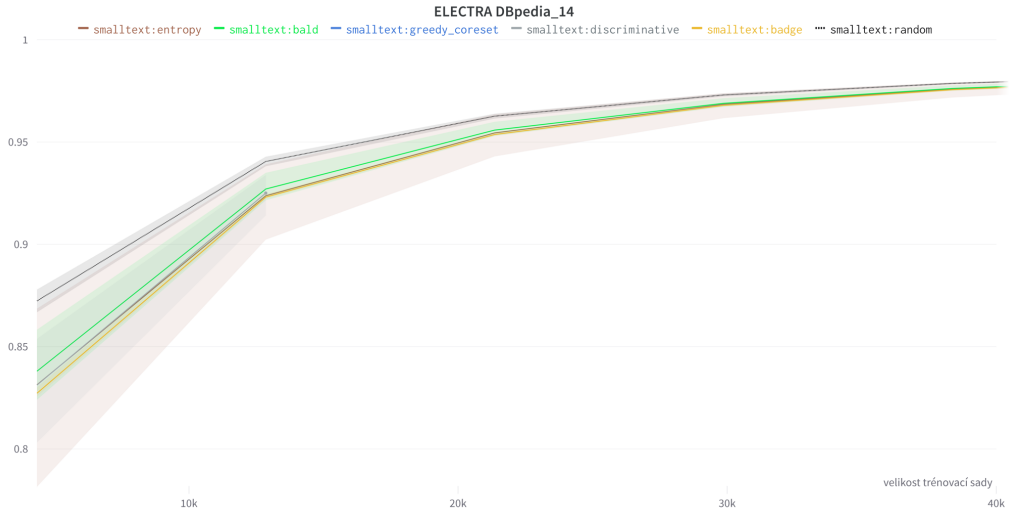
První kapitola práce po úvodu pojednává o souvisejících pracích a studiích s touto bakalářskou prací. Kapitola druhá uvádí problematiku aktivního učení a motivace pro něj. Následně uvádí taxonomii metod používaných v aktivním učení k obdržení vzorků, tzv. strategií aktivního učení, podloženou literaturou a v rámci této taxonomie představuje některé existující strategie – od původních až po nejmodernější zaměřené na neuronové sítě. Následující kapitola pojednává o experimentech v rámci práce – typy experimentů, dostupné datové sady a modely umělé inteligence a konkrétní návrh samotných experimentů. V další kapitole byly představeny výsledky experimentů ve formě požadovaných grafů.

Experimenty odhalily nekonzistentnost jednotlivých vybraných strategií aktivního učení – strategie vykazují odlišné chování pro různé modely a datové sady. V některých případech poskytuje náhodné vzorkování lepší výsledky než jiné strategie, v jiných naopak. Důvody neúspěchů strategií aktivního učení v této práci nebyly konstatovány, avšak existují studie zabývající se tímto tématem. Porovnání mezi dvěma typy představených experimentů ale přineslo důležité poznání. Ve všech případech byly zkoumané modely schopny dosáhnout podobné přesnosti při učení na části datové sady jako při učení na celé. Poměr velikosti části sady a velikosti celé sady a také rozdíl v přesnosti získané na dané části s přesností na celé sadě je proměnlivý podle datové sady, modelu a zvolené strategie, i přesto největší zjištěný rozdíl přesností činí 6 % se zhruba třetinou datové sady.

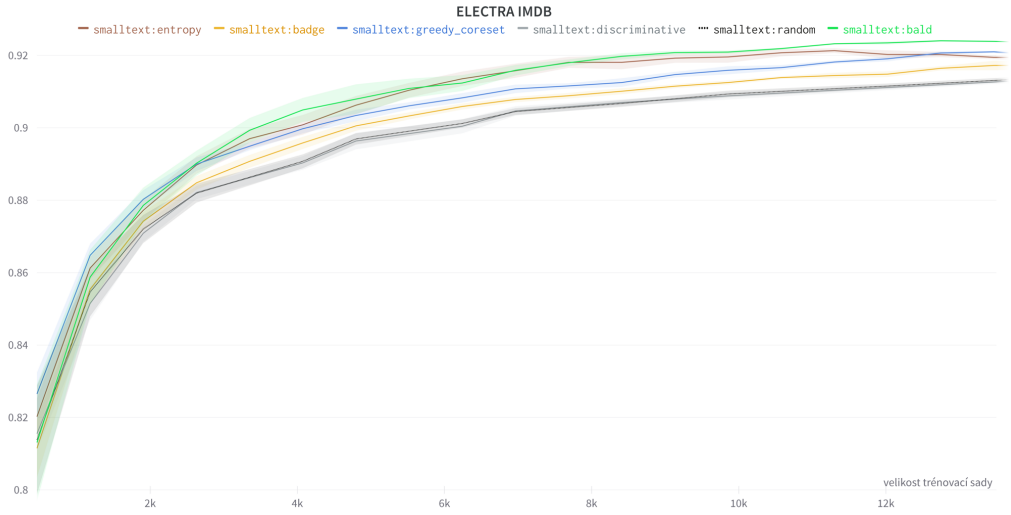
Grafy experimentů

A

A.0.1 Experimenty aktivního učení

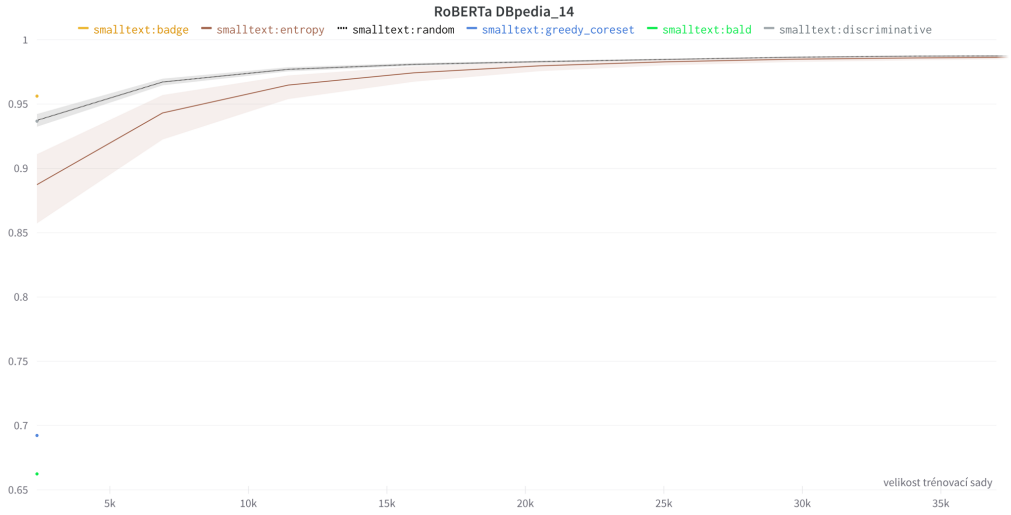


(a) Sada DBpedia Ontology

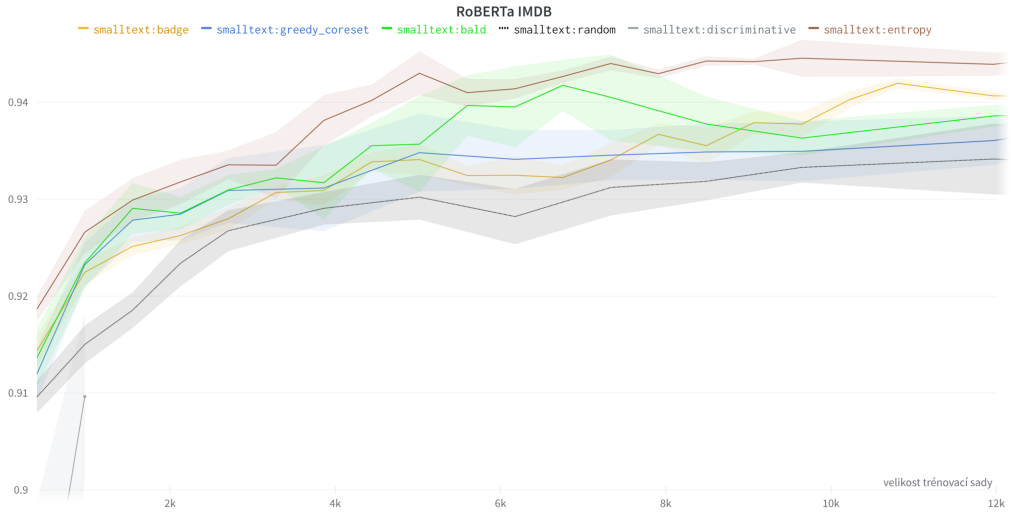


(b) Sada IMDB

Obrázek A.1: Přínos aktivního učení modelu Electra-Small

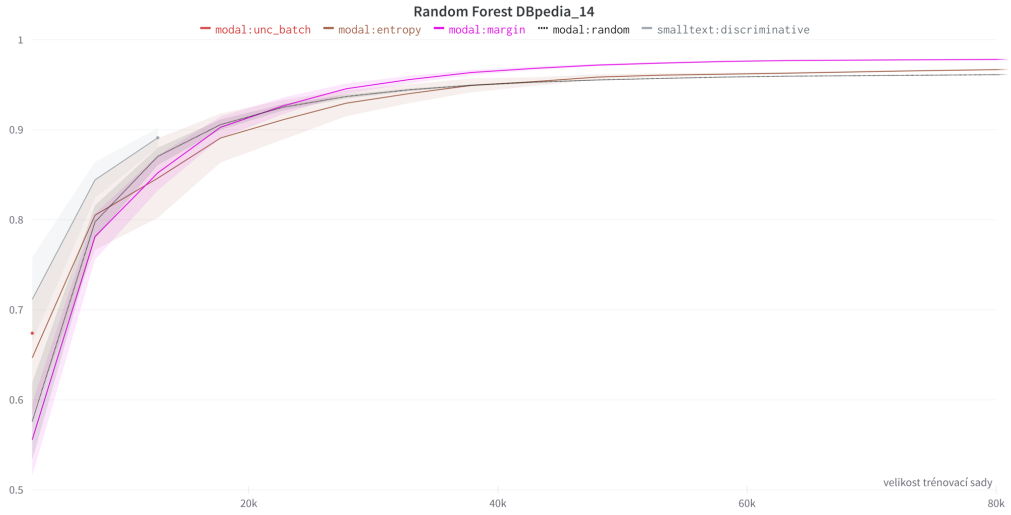


(a) Sada DBpedia Ontology

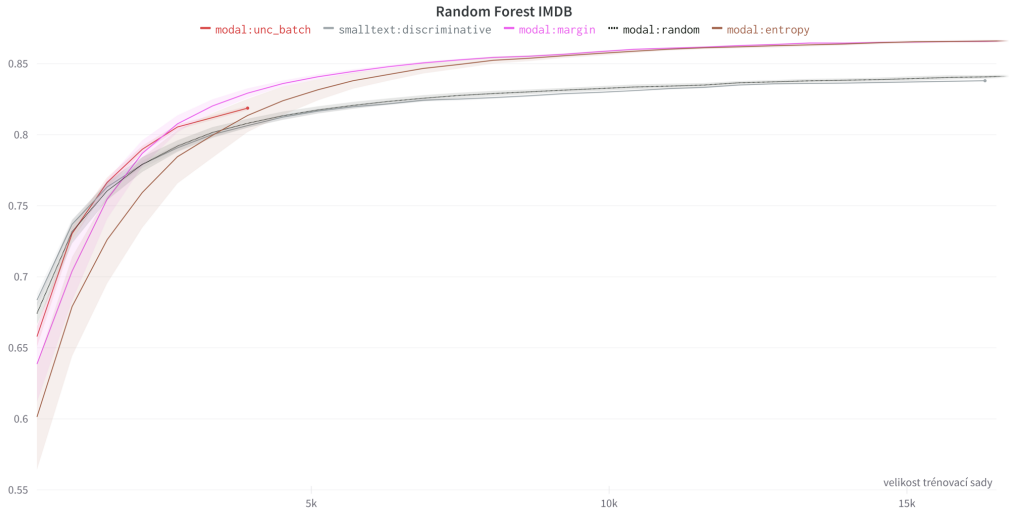


(b) Sada IMDB

Obrázek A.2: Přínos aktivního učení modelu RoBERTa-Base

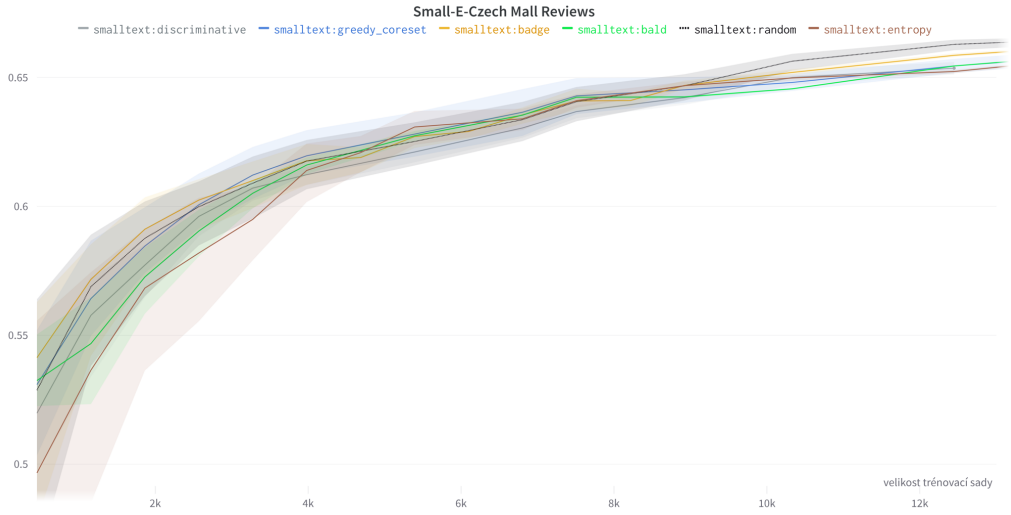


(a) Sada DBpedia Ontology

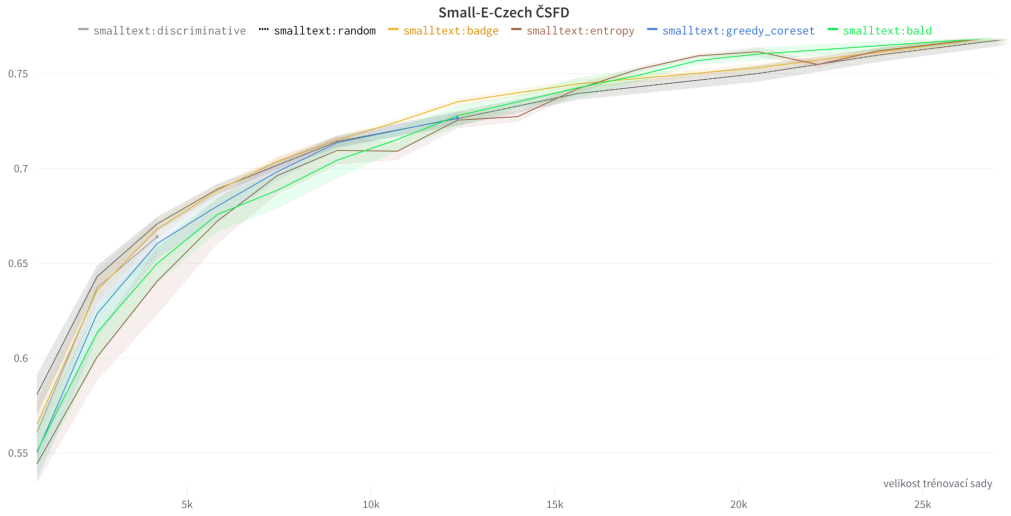


(b) Sada IMDB

Obrázek A.3: Přínos aktivního učení modelu Náhodný les na anglických sadách

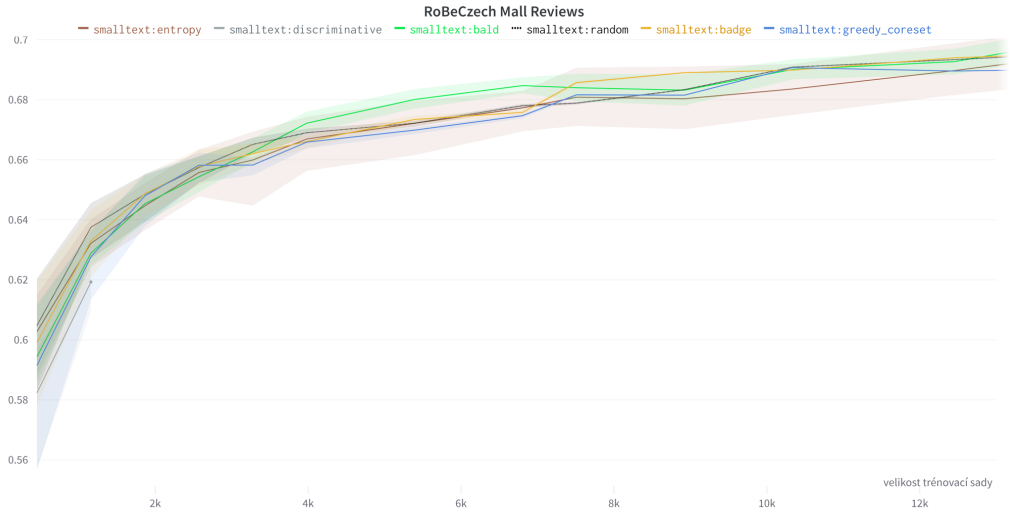


(a) Sada Mall.cz recenze

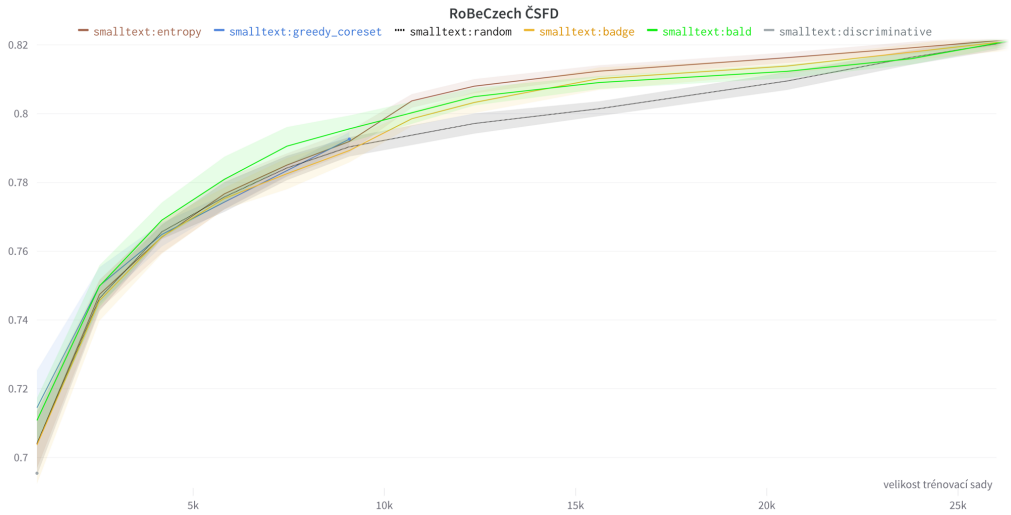


(b) Sada ČSFD

Obrázek A.4: Přínos aktivního učení modelu Small-E-Czech

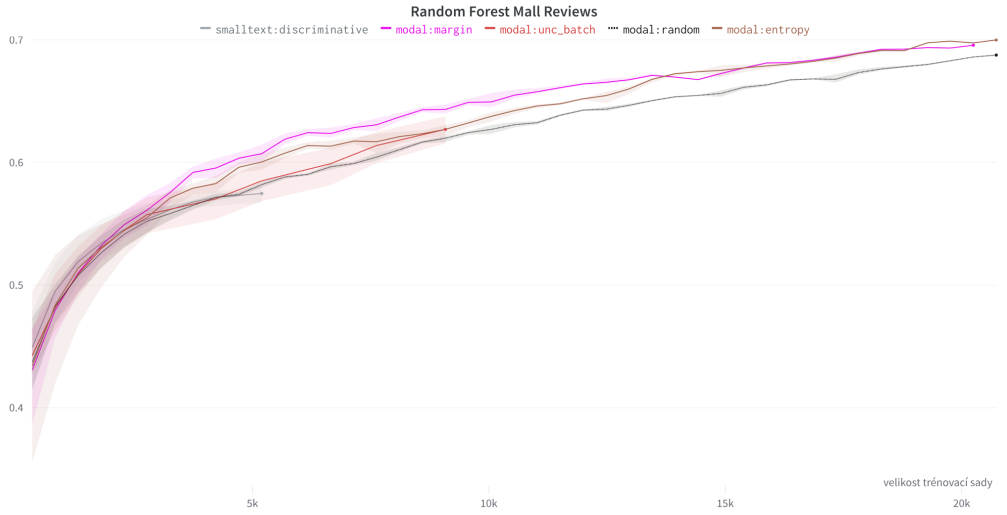


(a) Sada Mall.cz recenze

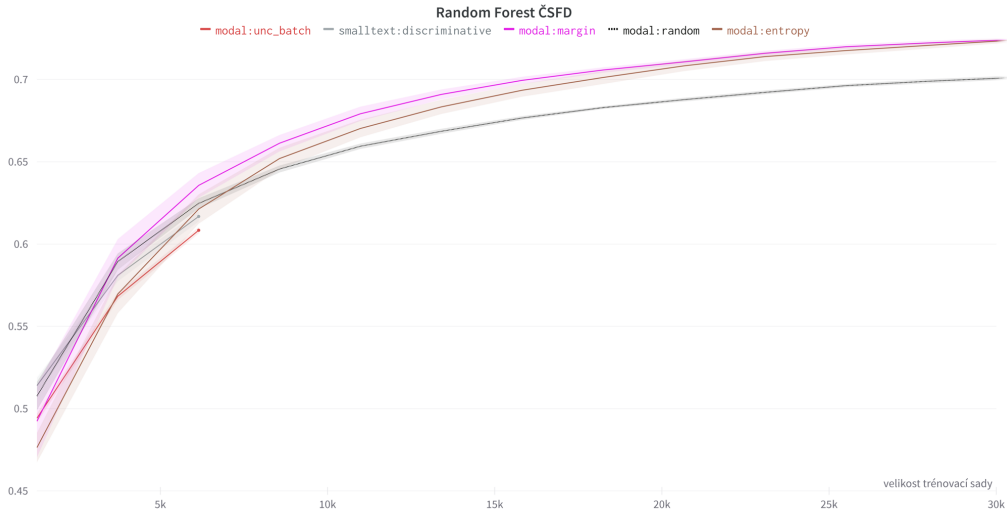


(b) Sada ČSFD

Obrázek A.5: Přínos aktivního učení modelu RoBeCzech



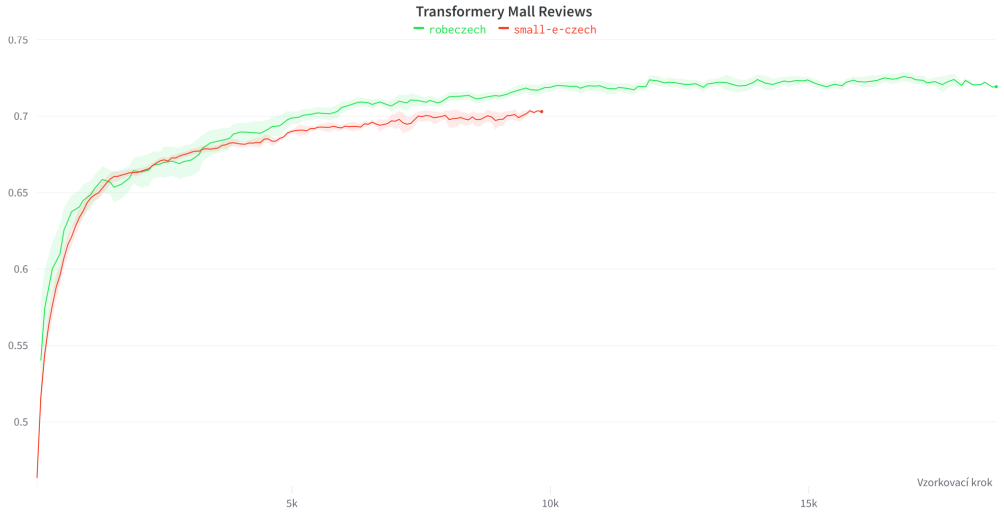
(a) Sada Mall.cz recenze



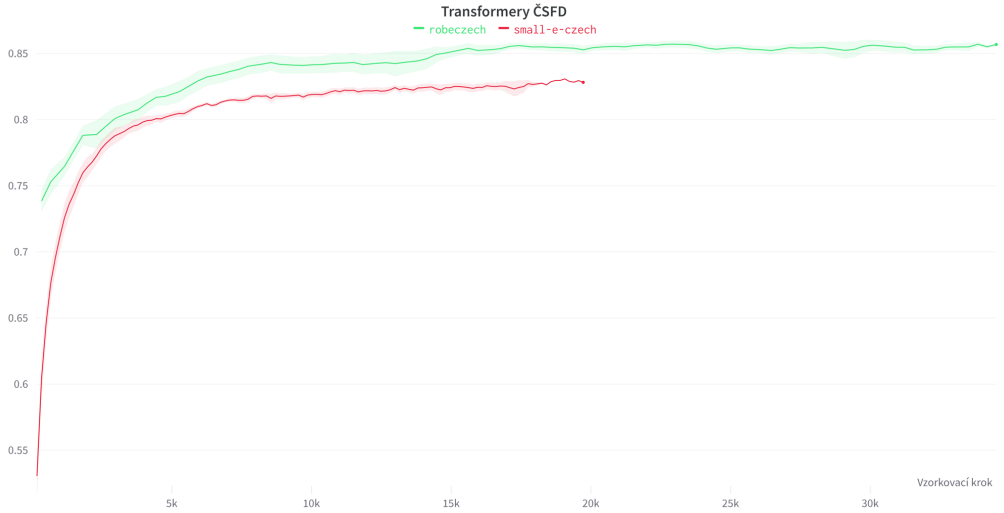
(b) Sada ČSFD

Obrázek A.6: Přínos aktivního učení modelu Náhodný les na českých sadách

A.0.2 Experimenty východisek

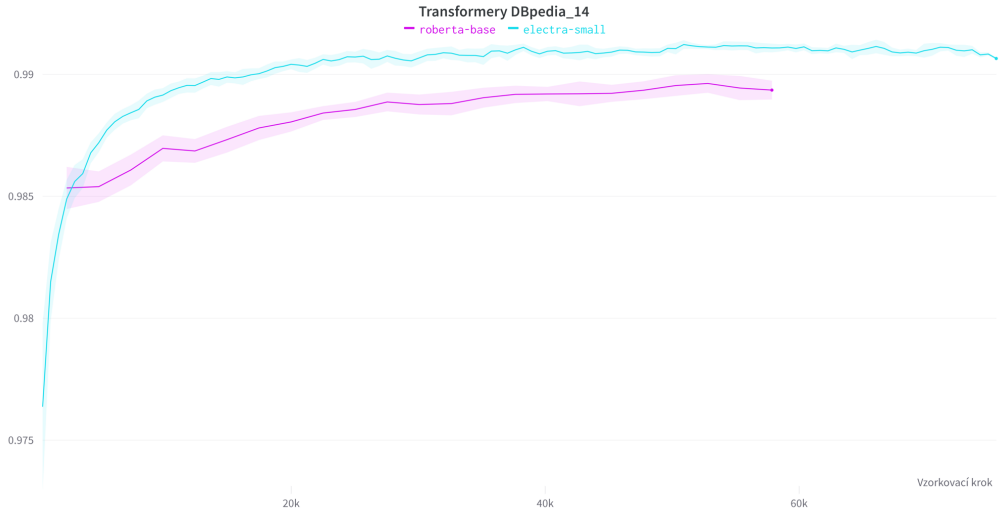


(a) Sada Mall.cz recenze

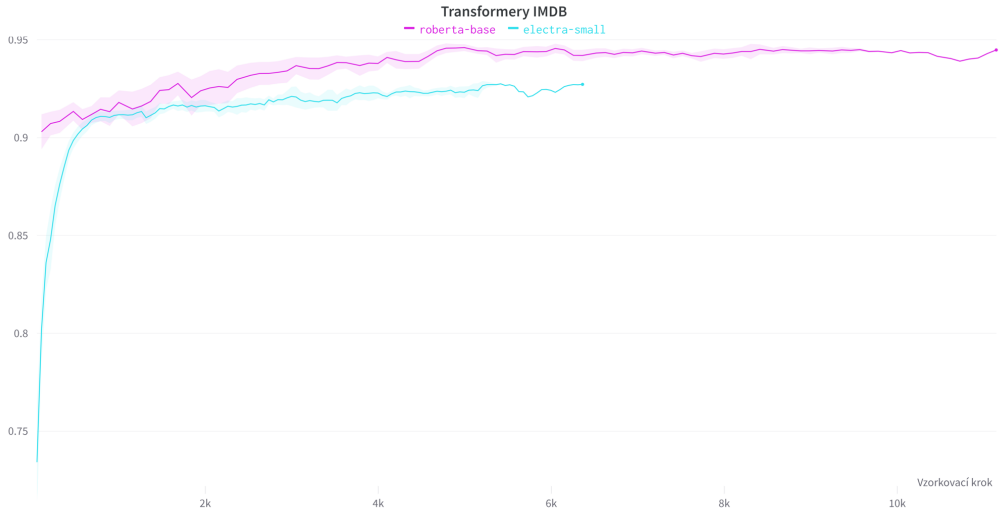


(b) Sada ČSFD

Obrázek A.7: Průběh učení transformer modelů na českých datových sadách

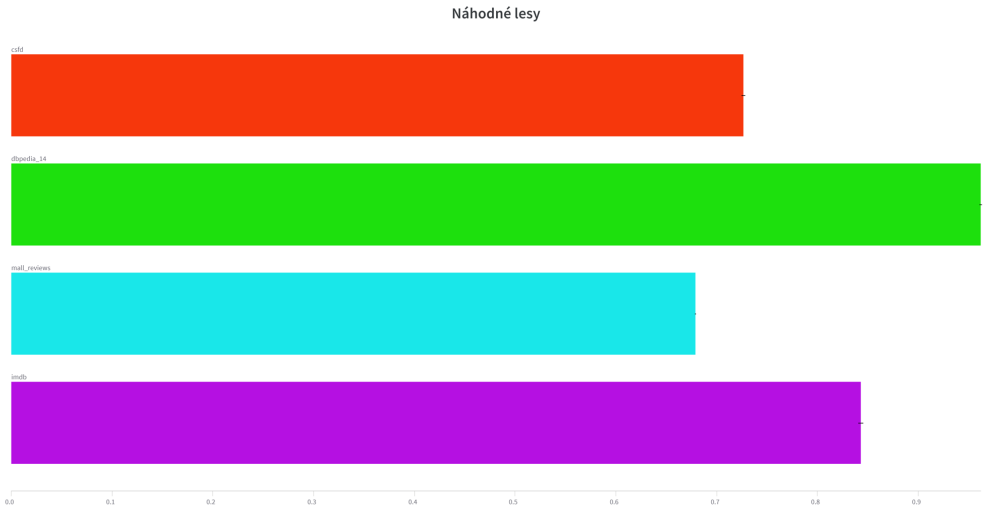


(a) Sada DBpedia Ontology



(b) Sada IMDB

Obrázek A.8: Průběh učení transformer modelů na anglických datových sadách



Obrázek A.9: Přesnost modelu Náhodný les na všech datových sadách

Bibliografie

[Agg+14] AGGARWAL, C.; KONG, X.; GU, Q.; HAN, J.; YU, P. Active Learning: A Survey. *Data Classification: Algorithms and Applications*. 2014. Dostupné také z: http://charuaggarwal.net/active-survey.pdf.

[AA19] ASH, Jordan T.; ADAMS, Ryan P. On the Difficulty of Warm-Starting Neural Network Training. *CoRR*. 2019, roč. abs/1910.08475. Dostupné z arXiv: 1910.08475.

[Ash+19] ASH, Jordan T.; ZHANG, Chicheng; KRISHNAMURTHY, Akshay; LANGFORD, John; AGARWAL, Alekh. Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds. *CoRR*. 2019, roč. abs/1906.03671. Dostupné z arXiv: 1906.03671.

[Car+17] CARDOSO, Thiago N.C.; SILVA, Rodrigo M.; CANUTO, Sérgio; MORO, Mirella M.; GONÇALVES, Marcos A. Ranked batch-mode active learning. *Information Sciences*. 2017, roč. 379, s. 313–337. ISSN 0020-0255. Dostupné z DOI: https://doi.org/10.1016/j.ins.2016.10.037.

[Cla+20] CLARK, Kevin; LUONG, Minh-Thang; LE, Quoc V.; MANNING, Christopher D. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *CoRR*. 2020, roč. abs/2003.10555. Dostupné z arXiv: 2003.10555.

[DD22] D'ARCY, Mike; DOWNEY, Doug. *Limitations of Active Learning With Deep Transformer Language Models*. 2022. Dostupné také z: https://openreview.net/forum?id=Q8OjAGkxwP5.

[DH18] DANKA, Tivadar; HORVÁTH, Péter. modAL: A modular active learning framework for Python. *CoRR*. 2018, roč. abs/1805.00979. Dostupné z arXiv: 1805.00979.

[Gal16] GAL, Yarin. *Uncertainty in Deep Learning*. 2016. Dis. pr. University of Cambridge.

[GG15] GAL, Yarin; GHAHRAMANI, Zoubin. Bayesian convolutional neural networks with Bernoulli approximate variational inference. *arXiv preprint arXiv:1506.02158*. 2015.

[GIG17] GAL, Yarin; ISLAM, Riashat; GHAHRAMANI, Zoubin. Deep Bayesian Active Learning with Image Data. *CoRR*. 2017, roč. abs/1703.02910. Dostupné z arXiv: 1703.02910.

[GS19] GISSIN, Daniel; SHALEV-SHWARTZ, Shai. Discriminative Active Learning. *CoRR*. 2019, roč. abs/1907.06347. Dostupné z arXiv: 1907.06347.

[Hou+11] HOULSBY, Neil; HUSZÁR, Ferenc; GHAHRAMANI, Zoubin; LENGYEL, Máté. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*. 2011.

[Kar+21] KARAMCHETI, Siddharth; KRISHNA, Ranjay; FEI-FEI, Li; MANNING, Christopher D. Mind Your Outliers! Investigating the Negative Impact of Outliers on Active Learning for Visual Question Answering. *CoRR*. 2021, roč. abs/2107.02331. Dostupné z arXiv: 2107.02331.

[KAG19] KIRSCH, Andreas; AMERSFOORT, Joost van; GAL, Yarin. BatchBALD: Efficient and Diverse Batch Acquisition for Deep Bayesian Active Learning. *CoRR*. 2019, roč. abs/1906.08158. Dostupné z arXiv: 1906.08158.

[Koc+21] KOCIÁN, Matej; NÁPLAVA, Jakub; STANCL, Daniel; KADLEC, Vladimír. Siamese BERT-based Model for Web Search Relevance Ranking Evaluated on a New Czech Dataset. *CoRR*. 2021, roč. abs/2112.01810. Dostupné z arXiv: 2112.01810.

[Leh+14] LEHMANN, Jens et al. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal*. 2014, roč. 6, s. 2. Dostupné z DOI: 10.3233/SW-140134.

[Liu+19] LIU, Yinhan et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*. 2019, roč. abs/1907.11692. Dostupné z arXiv: 1907.11692.

[Ped+11] PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 2011, roč. 12, s. 2825–2830.

[Ren+20] REN, Pengzhen et al. A Survey of Deep Active Learning. *CoRR*. 2020, roč. abs/2009.00236. Dostupné z arXiv: 2009.00236.

[SS18] SENER, Ozan; SAVARESE, Silvio. *Active Learning for Convolutional Neural Networks: A Core-Set Approach*. 2018. Dostupné z arXiv: 1708.00489 [stat.ML].

[Set09] SETTLES, Burr. *Active Learning Literature Survey*. 2009. Computer Sciences Technical Report, 1648. University of Wisconsin–Madison.

[Sch+21] SCHRÖDER, Christopher; MÜLLER, Lydia; NIEKLER, Andreas; POTTHAST, Martin. Small-Text: Active Learning for Text Classification in Python. 2021. Dostupné z arXiv: 2107.10314 [cs.LG].

[Str+21] STRAKA, Milan; NÁPLAVA, Jakub; STRAKOVÁ, Jana; SAMUEL, David. RobeCzech: Czech RoBERTa, a Monolingual Contextualized Language Representation Model. In: EKŠTEIN, Kamil; PÁRTL, František; KONOPÍK, Miloslav (ed.). *Text, Speech, and Dialogue*. Cham: Springer International Publishing, 2021, s. 197–209. ISBN 978-3-030-83527-9.

[YLB20] YUAN, Michelle; LIN, Hsuan-Tien; BOYD-GRABER, Jordan L. Cold-start Active Learning through Self-supervised Language Modeling. *CoRR*. 2020, roč. abs/2010.09535. Dostupné z arXiv: 2010.09535.

[ZZL15] ZHANG, Xiang; ZHAO, Junbo Jake; LECUN, Yann. Character-level Convolutional Networks for Text Classification. *CoRR*. 2015, roč. abs/1509.01626, s. 6. Dostupné z arXiv: 1509.01626.

Seznam obrázků

3.1 Vizualizace výběru vzorků jednotlivými metodami založenými na nejistotě pro úlohu ternární klasifikace. Osy zobrazují pravděpodobnost dvou tříd, pravděpodobnost třetí třídy se dopočítá jako $1-p_1-p_2$. Grafy pak vyjadřují skóre jednotlivých metod, vzorky s pravděpodobnostní distribucí odpovídající tmavým bodům jsou lepší kandidáti podle dané metody. Obrázek z [DH18, dokumentace knihovny] 9
3.2 Projekce datové sady IRIS do dvourozměrného prostoru a vyznačení některých prvků pro dávkovou a nedávkovou strategii. Obrázky pořízeny upraveným skriptem z [DH18, dokumentace knihovny – Ranked batch-mode sampling] 14

A.1 Přínos aktivního učení modelu Electra-Small 38
A.2 Přínos aktivního učení modelu RoBERTa-Base 39
A.3 Přínos aktivního učení modelu Náhodný les na anglických sadách . . . 40
A.4 Přínos aktivního učení modelu Small-E-Czech 41
A.5 Přínos aktivního učení modelu RoBeCzech 42
A.6 Přínos aktivního učení modelu Náhodný les na českých sadách 43
A.7 Průběh učení transformer modelů na českých datových sadách 44
A.8 Průběh učení transformer modelů na anglických datových sadách . . . 45
A.9 Přesnost modelu Náhodný les na všech datových sadách 46