Vysoká škola ekonomická v Praze

Fakulta informatiky a statistiky

****

Detekce osobních údajů pomocí metod hlubokého učení v nestrukturovaném textu

diplomová PRÁCE

Studijní program: Informační systémy a technologie

Specializace: Vývoj informačních systémů

Autor: Bc. David Ondrášek

Vedoucí diplomové práce: Ing. Josef Doležal

Praha, [měsíc rok]

Poděkování

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Aenean id metus id velit ullamcorper pulvinar. In convallis. Sed convallis magna eu sem. Maecenas lorem. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Ut tempus purus at lorem. Phasellus et lorem id felis nonummy placerat. Mauris elementum mauris vitae tortor. Aliquam id dolor. Vivamus porttitor turpis ac leo.

Abstrakt

Klíčová slova

amet, consectetuer, dolor, Lorem ipsum, sit.

JEL klasifikace

amet, consectetuer, dolor, Lorem ipsum, sit.

Abstract

Keywords

amet, consectetuer, dolor, Lorem ipsum, sit.

JEL Classification

amet, consectetuer, dolor, Lorem ipsum, sit.

Obsah

[Úvod 13](#_Toc97157005)

[1.1 Vymezení problému 14](#_Toc97157006)

[1.1.1 Absence jednoznačné definice osobního údaje 14](#_Toc97157007)

[1.1.2 Limitující závislost na jazyce trénovacích datasetů 15](#_Toc97157008)

[1.1.3 Nekomplexní implementace existujících modelů 15](#_Toc97157009)

[1.2 Účel a cíle práce 15](#_Toc97157010)

[1.3 Omezení projektu 16](#_Toc97157011)

[1.4 Význam a přínos práce 16](#_Toc97157012)

[2 Rešerše 17](#_Toc97157013)

[2.1 Rešeršní strategie 17](#_Toc97157014)

[2.2 Absence jednoznačné definice osobního údaje 18](#_Toc97157015)

[2.3 Limitující závislost na jazyce trénovacích datasetů 19](#_Toc97157016)

[2.4 Nekomplexní implementace existujících modelů 19](#_Toc97157017)

[2.5 Shrnutí 20](#_Toc97157018)

[3 Metodika práce 21](#_Toc97157019)

[4 Modelování osobních údajů 22](#_Toc97157020)

[4.1 Osobní údaje v kontextu Veřejného registru smluv 23](#_Toc97157021)

[4.1.1 Zákon o registru smluv 23](#_Toc97157022)

[4.1.2 Registr smluv 23](#_Toc97157023)

[4.1.3 Nástroj pro anonymizaci dokumentů 24](#_Toc97157024)

[4.2 Typy vstupních dokumentů 25](#_Toc97157025)

[4.3 Kategorizace osobních údajů 26](#_Toc97157026)

[4.3.1 Kompletní seznam 28](#_Toc97157027)

[4.4 Mapa kategorií 29](#_Toc97157028)

[4.4.1 Jmenné identifikátory 29](#_Toc97157029)

[4.4.2 Číselné identifikátory 29](#_Toc97157030)

[4.4.3 Znakové identifikátory 30](#_Toc97157031)

[4.4.4 Lokalizační identifikátory 30](#_Toc97157032)

[4.4.5 Ostatní identifikátory 30](#_Toc97157033)

[4.4.6 Shrnutí 30](#_Toc97157034)

[5 Použité techniky hlubokého učení 31](#_Toc97157035)

[5.1 Strojové učení 31](#_Toc97157036)

[5.2 Hluboké učení 31](#_Toc97157037)

[5.3 Natural Language Processing 31](#_Toc97157038)

[5.4 Named Entity Recognition 31](#_Toc97157039)

[5.5 Transfer learning 31](#_Toc97157040)

[5.6 Multilingvní modely 31](#_Toc97157041)

[5.7 Context-based modely 31](#_Toc97157042)

[5.8 Sekvenční modely 31](#_Toc97157043)

[5.9 Meta klasifikátory 31](#_Toc97157044)

[6 Dataset 32](#_Toc97157045)

[6.1 Získání trénovacích dat 32](#_Toc97157046)

[6.2 Průzkum dat 32](#_Toc97157047)

[6.3 Příprava dat 32](#_Toc97157048)

[6.3.1 Úpravy na základě příznakového modelu osobních údajů 32](#_Toc97157049)

[6.4 Reprezentace dat 32](#_Toc97157050)

[6.4.1 Word Embeddings 32](#_Toc97157051)

[6.4.2 Word2vec 32](#_Toc97157052)

[7 Výběr technologií 33](#_Toc97157053)

[7.1 Výběr NLP/NER frameworku 33](#_Toc97157054)

[7.2 DL knihovny 33](#_Toc97157055)

[7.3 Vývojové prostředí 33](#_Toc97157056)

[7.4 Techniky průběžné evaluace 33](#_Toc97157057)

[8 Vývoj modelu 34](#_Toc97157058)

[8.1 Architektura 34](#_Toc97157059)

[8.2 Rozdělení dat 34](#_Toc97157060)

[8.3 Ztrátová funkce 34](#_Toc97157061)

[8.4 Fine-tuning 34](#_Toc97157062)

[9 Evaluace modelu 35](#_Toc97157063)

[9.1 Hodnocení naměřených kvalitativních indikátorů 35](#_Toc97157064)

[9.2 Automatická evaluace 35](#_Toc97157065)

[9.3 Evaluace člověkem 35](#_Toc97157066)

[9.4 Zhodnocení výsledků a porovnání 35](#_Toc97157067)

[Závěr 37](#_Toc97157068)

[Použitá literatura 38](#_Toc97157069)

[Přílohy I](#_Toc97157070)

[Příloha A: Název první přílohy I](#_Toc97157071)

[Příloha B: Název druhé přílohy I](#_Toc97157072)

Seznam obrázků

Seznam tabulek

[Tabulka 1: Mapa kategorií osobních údajů 30](#_Toc97230177)

Seznam výpisů programového kódu

Seznam zkratek

NER Named Entity Recognition

NLP Natural Language Processing

OCR

GRU

LSTM

RNN

BRNN

TODO rozšířit dle již hotového textu

Úvod

GLOBAL TODO

* Bohužel v češtině kvůli nedostatku anotovaných datasetů se asi k kontext-citlivému rozpoznávání stejně moc nedostanu
* Model bude muset být modulární – meta-classifier, přičemž základními modely budou ty co umějí základní NER – jména, bydliště, atd.
* Czech Named Entity Corpus
* Kontext-citlivé rozpoznávání
  + Anonymizovat ne jen jedno slovo, ale celou větu kde se více identifikačních údajů nachází
  + Určovat kontext na základě více vyhledaných sekundárních identifikátorů

Není to tak dávno, kdy si pod pojmem automatizace šlo představit pouze zavádění robotů a automatických procedur do továren a tento termín byl sám o sobě spojován pouze s průmyslovým sektorem. Dnes je s příchodem digitálních technologií vše jinak a stejný proces automatizace můžeme pozorovat i v oblasti služeb. Tam dochází k automatizaci rutinních kancelářských činností a celkově vzato i většiny nekreativních prací.

Jednou z těchto prací je i detekce a klasifikace osobních údajů v různých textech. Tato práce je ve velké části případů vykonávána kvůli povinnosti anonymizace těchto údajů na základě různých zákonných opatření, týkajících se ochrany osobních údajů, v čele se známým evropským nařízením GDPR.

Problematika detekce identifikátorů v textu je součástí vědecké domény s názvem Named Entity Recognition (NER), která se soustavně vyvíjí již mnoho let. Samotný termín Named Entity byl definován R. Grishmanem již v roce 1996 (Grishman, 1996). Od té doby došlo k objevu řady významných metod detekce identifikátorů, jako jsou některé slovníkové metody, lookup tabulky nebo regulární výrazy (Nadeau, 2007). Poslední dobou se však jako nejvhodnější způsob pro detekci identifikátorů v textu jeví využití hlubokého učení, zejména pak technik Natural Language Processing (NLP). To totiž umožňuje nastavit detekční pravidla, která se vyznačují vyšší abstrakcí a dokážou tak zachytit i případy, které se v textu vymykají běžným slovním vyjádřením. To v důsledku znamená, že techniky hlubokého učení mají v NER výsadní postavení a v současné době vykazují nejlepší výsledky (Xu et al., 2021).

Detekce a klasifikace osobních údajů je tedy typově specifickou aplikací NER a potažmo NLP na konkrétní sadu dokumentů, obsahujících osobní údaje. Problém v tomto případě vzniká nejasnou definicí toho, co to osobní údaj vlastně je, respektive tím, že se osobním údajem mohou stát i informace, které nejsou explicitně unikátní (Mihulková, 2018). V některých případech je tedy nutné tuto informaci (zda se jedná o osobní údaj) získat přímo z textu na základě širšího kontextu. To existující zahraniční vědecké práce, které se detekcí osobních údajů již zabývají spíše ignorují a soustředí se více na kvalitativní zpřesnění detekce identifikátorů na základě pokročilejších statistických modelů (Silva et al., 2020).

Ač se některé současné práce problematikou NER s ohledem na širší kontext textu zabývají, je tato funkčnost modelů hlubokého učení stále problematická a složitá (Williams, 2017). Místo hledání obecného řešení, které by dokázalo rozlišovat kontext ve všech typech textů, je v této situaci vhodnějším řešením tuto funkčnost aplikovat pouze na konkrétní typ dokumentu, čímž se sníží nutný stupeň abstrakce potřebný pro zpracování textu. Aplikací tohoto principu v případě detekce osobních údajů by tak mělo být možné získat kvalitnější výsledky.

Tato diplomová práce zkoumá možnost využití kontext-citlivé NER v oblasti detekce osobních údajů, a to na sadě dokumentů, které jsou běžně nahrávány do veřejného registru smluv. Pomocí state-of-the-art metod hlubokého učení je v ní navržen a implementován nástroj, který je volným rozšířením anonymizačního nástroje “Nástroj pro anonymizaci dokumentů” dostupném na Portálu veřejné správy (<https://anonymizace.gov.cz/crossroad>). Pomocí tohoto nástroje lze anonymizovat nahrané dokumenty, lze to však provádět pouze manuálně. Nástroj, navržený v této práci, by tuto činnost po integraci do původního nástroje dokázal provádět i automaticky.

TODO rozšířit podle budoucího stavu

## Vymezení problému

Subjekty, které mají povinnost nahrávat smlouvy do veřejného registru smluv čelí při tomto procesu řadě těžkostem, které dokonce mohou v některých případech zamezit vkládání smluv pracovníkům bez dostatečného právnického a technického vzdělání.

Jedním z těchto problémů je povinnost anonymizovat osobní údaje v nahraných dokumentech a přílohách. Tato povinnost je stanovena dle 160/1999 Sb. Zákona o svobodném přístupu k informacím a následně i 101/200 Sb. Zákona o ochraně osobních údajů, který upravuje předpisy GDPR. Samotná činnost detekce osobních údajů pro jejich následnou anonymizaci je nejen jednotvárná a nezáživná, ale existuje zde velké riziko lidské chyby, ať už přehlédnutím důležitého údaje nebo i jeho špatnou klasifikací a interpretací.

Následující problémy jsou společné pro jakoukoliv detekci osobních údajů v nestrukturovaném textu. Tímto pojmem (nestrukturovaný text) se označuje jakýkoliv text, který nemá dopředu pevně stanovenou strukturu v podobě předem definovaných kapitol či povinných textových částí. Neobsahuje také žádná povinná textová pole, do kterých by byly zapisovány konkrétní informace (např. pole pro jméno a příjmení).

### Absence jednoznačné definice osobního údaje

Dle Zákona o zpracování osobních údajů lze termín osobní údaj definovat jako jakoukoli informaci, která se týká identifikované nebo identifikovatelné žijící osoby (Mihulková, 2018). Tato definice je z hlediska implementace anonymizačního nástroje velmi široká. Neexistuje žádný souhrnný seznam, který by definoval, co všechno se dá považovat za osobní údaj. Je proto třeba definovat jednotlivé osobní, případně citlivé osobní údaje pro konkrétní typ dokumentů, který bude nástroj zpracovávat.

### Limitující závislost na jazyce trénovacích datasetů

Pro úspěšné využití jakékoliv techniky hlubokého učení je potřeba mít k dispozici kvalitní dataset, na jehož datech může být vznikající model natrénován. V oblasti NLP (Natural Language Processing) je tak klíčové mít buď k dispozici dataset v požadovaném jazyce nebo předtrénovaný multilingvní model, který by se dal následně metodou transfer learning přetrénovat pro požadované využití (Moberg, 2020). V případě detekce osobních údajů se navíc v existujících datových sadách obsahujících ukázky různých smluv logicky žádné osobní údaje nevyskytují, protože tyto smlouvy již anonymizací musely projít.

### Nekomplexní implementace existujících modelů

Jednotlivé již existující modely, určené pro detekci identifikátorů v nestrukturovaném textu nejsou dostatečně komplexní, aby dokázaly rozlišit osobní údaje i na základě širšího kontextu analyzovaného textu. Zároveň jsou tyto modely často vytvářeny pouze jako akademické koncepty a necílí na žádnou konkrétnější doménu do které by měly spadat zpracovávané dokumenty. Také vzhledem k jejich abstraktnímu využití nejsou tyto modely veřejně dostupné.

## Účel a cíle práce

Hlavním cílem této práce je navrhnout a implementovat prototyp nástroje, který pomocí algoritmů hlubokého učení dokáže v nestrukturovaném textu klasifikovat osobní údaje a umožní tak rychlejší a přesnější zpracování dokumentů v případě potřeby anonymizace těchto údajů. Konkrétní implementace bude přizpůsobena povinné anonymizaci osobních údajů při nahrávání dokumentů do veřejného registru smluv. Bude tedy sloužit jako rozšíření existujícího nástroje “Nástroj pro anonymizaci dokumentů” dostupném na Portálu veřejné správy (<https://anonymizace.gov.cz/crossroad>) a pomáhat může všem subjektům povinným v tomto registru smlouvy zveřejňovat. Nástroj bude následně evaluován na vhodně vybrané sadě dokumentů.

Dílčím cílem práce je analyzovat zákony a nařízení, které je potřeba respektovat při anonymizaci osobních údajů při nahrávání dokumentů do veřejného registru smluv. Na základě této analýzy bude vypracován model, který bude vyjadřovat míru potřeby anonymizace konkrétního identifikátoru v nestrukturovaném textu. Tento model bude sloužit jako podklad pro tvorbu požadavků pro vyvíjený prototyp nástroje.

Dalším dílčím cílem je provést rešerši existujících metod hlubokého učení, využitelných pro klasifikaci osobních údajů v nestrukturovaném textu, zejména pak metod, které umožňují tuto klasifikaci v textech v českém jazyce. Dojde také k analýze možností kombinace těchto metod za účelem dosažení co nejlepšího výsledku detekce a klasifikace údajů.

TODO tahle kapitola možná nebude potřeba?

## Omezení projektu

Vzhledem ke komplexnosti celé problematiky ochrany osobních údajů a kvůli specifickému charakteru osobních údajů v různých druzích textů je navržený nástroj omezen jen na jednu konkrétní sadu dokumentů, a to na sadu textových dokumentů, které jsou běžně nahrávány do veřejného registru smluv.

Dále je práce omezena limitujícím počtem českých datasetů vhodných pro trénování vyvíjeného modelu. Vytváření nutně komplexnějších datových sad není součástí této práce. To může znamenat menší přesnost detekce než u modelů vyvíjených zahraničními výzkumníky a trénovaných na rozsáhlejších a lépe připravených textových datech v jiném jazyce.

Práce se také nesnaží zkoumat nové algoritmy hlubokého učení. Místo toho v ní jde o výběr state-of-the-art metod hlubokého učení a jejich aplikaci na danou problematiku. Znamená to, že přesnost detekce vyvíjeného nástroje bude s největší pravděpodobností maximálně stejně velká, jako přesnost detekce u existujících metod hlubokého učení, z kterých bude vycházet. Stejně tak i funkcionalita nástroje v oblasti rozpoznávání širšího kontextu textu bude omezená současným stavem poznání v této oblasti.

## Význam a přínos práce

Přínosem této práce je samotný nástroj na detekci osobních údajů v nestrukturovaných dokumentech běžně nahrávaných do veřejného registru smluv, který slouží jako proof of concept využitelnosti algoritmů hlubokého učení a kontext-citlivé NER při zpracování osobních údajů.

Vyvinutý nástroj pomůže všem subjektům, povinným v tomto registru smlouvy zveřejňovat, v rychlejší a přesnější detekci osobních údajů v nahrávaných dokumentech za účelem jejich následné anonymizace. V ideálním případě bude moci být tento nástroj použit jako rozšíření existujícího nástroje “Nástroj pro anonymizaci dokumentů”, dostupném na Portálu veřejné správy (<https://anonymizace.gov.cz/crossroad>). Přínosem se dá také označit fakt, že nástroj bude detekovat osobní údaje v textech v českém jazyce.

Dalším přínosem pak může být modularita vytvořeného modelu, která by měla umožňovat ostatním výzkumníkům převzít část modelu a upravit ji pro jiný typ dokumentů. Jinak řečeno tedy využít stejný postup pro vytváření požadavků jako v této práci a následně model přetrénovat na nové sadě dokumentů. Model pak bude vhodný k detekci osobních údajů i v jiném případě užití, jako je například anonymizace osobních údajů při digitalizaci veřejně přístupného archivu.

TODO opravit podle výsledků

# Rešerše

## Rešeršní strategie

Systematická část rešerše vhodné literatury proběhla s využitím databáze ACM, Google Scholar a vyhledávačem Univerzity Karlovy UKAŽ. Na ACM a Google Scholar byly vyhledávány práce s technickým zaměřením, týkající se nejnovějších poznatků NLP ve spojení s klasifikací a detekcí identifikátorů v textu nebo využití multilingvních modelů v praktických aplikacích.

V databázi UKAŽ pak byly vyhledávány práce zabývající se ochranou osobních údajů a jejich identifikací, kterých nebylo na předchozích databázích nalezeno dostatečné množství. Nebyl vybrán žádný časový filtr, ale bylo nastaveno řazení článků od nejnovějších ke starším a následně i řazení dle relevance. Při každém hledání bylo analyzováno prvních 20 článků, přičemž největší důraz byl při tom kladen na praktické využití technik NLP při detekci textových identifikátorů co nejvíce podobných osobním údajům.

Při vyhledávání docházelo ke spojením několika klíčových slov logickými operátory.

**Klíčová slova využitá ve vyhledávacích řetězcích ACM a Google Scholar**

*NLP, Natural Language Processing, NER, Named Entity Recognition, multilingual, PII, personally identifiable information, Cognitive Data Capture, privacy, data privacy, private, personal, anonymization, pseudonymization, data security*

**Příklady vyhledávacích řetězců pro ACM a Google Scholar**

* (NLP OR Natural Language Processing) AND multilingual
* (NER OR Named Entity Recognition) AND (PII OR personally identifiable information)
* (NER OR Named Entity Recognition) AND (PII OR personally identifiable information) AND (NLP OR Natural Language Processing)
* (NLP OR Natural Language Processing) AND (Cognitive Data Capture)
* (NLP OR Natural Language Processing) AND privacy
* (NLP OR Natural Language Processing) AND data privacy
* (NLP OR Natural Language Processing) AND private
* (NLP OR Natural Language Processing) AND personal
* (NLP OR Natural Language Processing) anonymization

**Klíčová slova využitá ve vyhledávacích řetězcích UKAŽ**

*Ochrana osobních údajů, automatizovaný, GDPR*

**Příklady vyhledávacích řetězců pro UKAŽ:**

* jakékoliv pole obsahuje Ochrana osobních údajů
* jakékoliv pole obsahuje Ochrana osobních údajů AND jakékoliv pole obsahuje automatizovan
* jakékoliv pole obsahuje GDPR
* jakékoliv pole obsahuje GDPR AND jakékoliv pole obsahuje automatizovan

V nesystematické části rešerše došlo k prohledávání internetu se zaměřením na firmy, zabývající se digitalizací dokumentů a na konkrétní informace týkající se ochrany osobních údajů (například definice samotného osobního údaje). Pozorně prostudována byla například webová stránka Evropské komise, která srozumitelnou formou čtenáře seznamuje se základními pojmy v oblasti ochrany údajů v Evropské unii. (EK). Došlo také na odborné konzultace, ve kterých byla doporučena odborná literatura týkající se ochrany osobních údajů.

## Absence jednoznačné definice osobního údaje

Termín osobní údaj, definovaný již v kapitole Vymezení problému, může mít ze své podstaty mnoho významů. Informace je totiž za osobní údaj uvažovaná až ve chvíli, kdy vede k přímé a jednoznačné identifikaci jednotlivce. Často se tedy stává, že osobním údajem se údaj stává až ve chvíli, když se vyskytuje v textu společně s dalšími údaji, které dohromady dávají ucelenou informací o konkrétní osobě (Mihulková, 2018).

Do tohoto vymezení se mísí i koncepce práva na soukromí, která je ukotvena v evropských, kontinentálních právních kulturách. V této koncepci se jako osobní údaj uvažuje i jakákoliv informace, kterou jednotlivec sám o sobě nechce sdílet ve veřejném prostoru (Kubica, 2019).

Jako jednotlivá podkategorie osobních údajů se rozlišují tzv. zvláštní osobní údaje. Tyto údaje mohou být buď citlivějšího charakteru a vypovídat o rasovém původu, politických názorech nebo sexuálním vyznání a celkově vzato jsou klasifikovány tak, že mohou subjekt poškodit ve společnosti. Mezi tyto údaje potom patří i biometrické údaje, jakožto jednoznačné fyzické identifikátory subjektu (MVČR). Zpracování údajů spadajících do této podkategorie se řídí přísnějšími pravidly a při implementaci anonymizačního nástroje to musí být bráno v potaz.

Problém vzniká ve státní správě, kde může kolidovat 106/1999 Sb. Zákon o svobodném přístupu k informacím a 101/200 Sb. Zákon o ochraně osobních údajů. Občan má totiž právo přístupu k informacím, které nejsou omezené nařízeními o ochraně osobních údajů. Tyto informace potom může dostat v anonymizované formě. Ve chvíli, kdy ale údaj, jako je například křestní jméno není považován za osobní údaj, nespadá pod zákon o ochraně osobních údajů a měl by se dostat k žadateli v plné, neanonymizované podobě (Gealfow et al., 2019).

Na základě těchto tezí můžeme tušit, že pojem osobní údaj se liší podle konkrétní aplikace využití informací, a i podle sektoru, ve kterém ke zpracování dat dochází. Ostatně i například v (EK) a (Mihulková) se uvedené příklady osobních údajů liší.

## Limitující závislost na jazyce trénovacích datasetů

V oblasti strojového učení, a zejména pak v podoblasti NLP je často adresován problém nelokalizovaných vstupních dat. Vzhledem k dnešnímu stavu vědeckého světa se totiž výzkumné práce píšou ve většině případech v angličtině a v důsledku toho existuje i největší množství předtrénovaných modelů, které jsou založeny právě na anglických korpusech (Pod pojmem korpus se rozumí množina slov, která může být využita pro trénink modelu. Většinou jsou slova v korpusu spojená vazbami k doméně, ke které je korpus plánovitě vytvořen). Angličtina je také výhodná z pohledu Named Entity Recognition (NER), protože například oproti češtině se v ní nevyskytují ve větším množství pády, rody či větší množství nepravidelných tvarů množných čísel podstatných jmen.

Možným řešením problému s nedostatkem kvalitních korpusů různých jazyků mohou být moderní modely, jako je například model ELECTRA, který je vhodný pro trénování sítě transformerů i menším výpočetním výkonem a korpusem menší velikosti (Clark et al., 2020). Populárním řešením je také využití revolučního modelu BERT, publikovaného týmem výzkumníků z Google v roce 2018 (Devlin et al., 2019), zejména pak jeho multilingvní varianty M-BERT, která v současné době podporuje 104 jazyků, včetně češtiny (BERT GITHUB). Problém v tomto případě může být imbalance korpusu, na kterém byl model trénován (celosvětové záznamy Wikipedie). Kvůli rozdílnému poměru celkového množství textu každého jazyka v trénovacím datasetu se tak u modelu M-BERT dají u méně zastoupených jazyků čekat horší výsledky (Chau et al., 2020).

Jako další state-of-the-art multilingvní modely se dají označit modely XLM a jeho upravená varianta XLM-R, které mají ve srovnání provedeném Johnem Mobergem na 15 jazycích o několik procent lepší úspěšnost (Moberg, 2020).

## Nekomplexní implementace existujících modelů

Jak již bylo dříve zmíněno, pro anonymizaci osobních údaje je vhodné zvolit metody NLP, které jsou součástí technik hlubokého učení (Qu el al., 2021; Silva et al., 2020; Ellman, 2018). Jednotlivé metody z této oblasti lze pro dosažení optimálních výsledků potom různě kombinovat. Většina existujících prací se však zabývá pouze zkoumáním a implementací těchto metod zvlášť.

Silva ve své práci využívá NER a porovnává v ní vhodné NLP toolkity, které se hodí pro klasifikaci osobních identifikátorů v smlouvách. S největším F1 score mu v jeho případě funguje toolkit Stanford CoreNLP (Silva et al., 2020). Jeho výsledek pak rozporuje např. Mendels, kterému v jeho případě nejlepší F1 score vychází u toolkitu Flair (Mendels, 2020).

Mendels sám uvádí další vhodné metody vhodné pro detekci identifikátorů, jako je využití regulárních výrazů uvnitř klasifikačních vrstev nebo vytvoření blacklistů s textovými řetězci s větší pravděpodobností výskytu hledaného výrazu.

Zajímavá je práce Mathiase Ellmanna, který se pomocí NLP snaží detekovat duplikáty „issue trackerů”. Narozdíl od předchozích autorů jde více do hloubky (vzhledem k tomu, že pracuje pouze s binárním klasifikátorem si to může dovolit) a tím dokáže lépe zpracovat samotnou extrakci příznaků z textu na základě sémantické analýzy textu originálního „issue“. Tím potom dochází k rozšířenému chápání širšího kontextu a menšímu zaměření na samotné identifikátory (Ellman, 2018).

Širším kontextem se zabývá také Jake Williams, který ve své práci vytváří „interní“ a „externí“ model. Interní model je určen pro lokální klasifikaci identifikátorů jako je název města nebo jméno. Externí model potom pracuje s již předzpracovaným textem připraveným interním modelem. Tento text obsahuje klasifikované identifikátory nahrazené klíčovými slovy, které by měly být z hlediska širšího kontextu relevantními pro lepší pochopení textu (Williams, 2017).

Chen Qu se zabývá konceptem typických architektur anonymizačních modelů a pokládá otázku, jak s anonymizovanými daty dokážou pracovat další vrstvy hlubokých neuronových sítí, které byly předučeny na neanonymizovaných datech (Qu el al., 2021).

## Shrnutí

Z provedené rešerše se dá odvodit pravděpodobný další postup při implementaci anonymizačního nástroje. Nejprve je na základě provedené analýzy potřeba definovat požadavky na vývoj nástroje, zejména potom definovat osobní údaje pro konkrétní aplikaci anonymizace dokumentů nahrávaných do veřejného registru smluv, které se stanou příznaky ve fázi feature extraction při trénování modelu.

Vychází z ní také na co je třeba dávat si pozor při výběru modelu a při vytváření architektury nástroje a kombinaci různých NLP metod. Je také potřeba vybrat vhodný dataset, který musí reflektovat nejen jazyk analyzovaného textu při detekci osobních údajů, ale i vybraný model hluboké neuronové sítě, který tomu musí být přizpůsoben.

# Metodika práce

Vzhledem k charakteru této práce byla pro implementační část vybrána lehká metodika ML Project Checklist, vytvořená Aurélienem Géronem jako součást knihy *Hands-on Machine Learning with Scikit-Learn and TensorFlow,* která definuje fáze vývoje kvalitního modelu hluboké neuronové sítě v souvislosti s ideálním výběrem vhodných technik využitelných při vývoji požadované aplikace (Géron, 2017).

Samotný vyvinutý prototyp bude evaluován pomocí nejznámějších metrik využívaných při evaluaci modelů hlubokých neuronových sítí, jako je F1 score, Precision, Recall, aj. (Mishra, 2018). Evaluace bude probíhat na sadě dokumentů, které se běžně nahrávají do veřejného registru smluv, typicky tedy na jednotlivých smlouvách.

Analýza principů hlubokého učení, využitelných pro klasifikaci osobních údajů a také derivace požadavků na vyvíjený nástroj ze zákonů a zákonných opatření bude probíhat cestou systematické rešerše.

TODO rozšířit

TODO napsat k jednotlivým kapitolám jakou část metodiky plní

# Modelování osobních údajů

TODO zatím jen pracovní název

* Název by měl být jiný, nejde totiž o příznaky, jako spíš o kategorie pro hiearchical text classification
* Vypracovat nějaký vzorec pro určení toho kolik jednotlivých osobních údajů potřebuji k identifikaci osoby (nejednoznačných osobních údajů)

Paralela s ML checklistem – vlastně tím plníme první kapitolu

TODO

* Napsat na portál veřejného registru a poptat se jich s čím se tam potýkají
  + co je nejsložitější na zpracování a detekci
  + jaké dokumenty je možné nahrávat
  + jaká existují rizika v případě že nebudou osobní údaje správně anomymizovány
  + s jakými PII se nejčastěji setkávají a jaké se dají považovat za obtížněji detekovatelné
  + existují nějaké speciální PII, které jsou vyloženě spojené s registrem smluv?
  + Poslat jim můj seznam a zeptat se jestli by ho dokázali rozšířit
  + přiložit komunikaci do příloh
  + napsat i dalším tvůrcům anonymizačních nástrojů na jakém principu je jejich nástroj založen…jestli používají techniky DL
* co hrozí při špatném zpracování osobních údajů
* definovat základní pojmy zpracování údajů – správce, zpracovavatel, atd.
* srovnání již existujících řešení na českém trhu – nástroje na anonymizace údajů při správě dokumentů
  + důraz na to, které z těchto rešení nabízí přímo auto klasifikaci údajů
  + důraz na to kdo dokáže číst i dle kontextu

TODO předmluva

* oslí můstek k tomu abych se dostal k veřejnému registru smluv
* mělo by to být psané tak, aby to bylo více obecné a registr smluv je jenom konkrétní aplikace použitých technik

## Osobní údaje v kontextu Veřejného registru smluv

### Zákon o registru smluv

340/2015 Sb. Zákon o registru smluv, který nabyl platnosti 1. 7. 2016 a který byl 2x novelizován, nejprve v roce 2017 (zákonem 249/2017 Sb.), dále v roce 2019 (zákonem 177/2019 Sb.), hovoří o zvláštních podmínkách účinnosti některých smluv a o uveřejňování těchto smluv ve veřejném registru smluv (Bouda, 2019).

Tento zákon stanovuje povinnost k uveřejňování soukromoprávních smluv a smluv o poskytnutí dotace nebo návratné finanční výpomoci, jejichž alespoň jednou stranou je některý z okruhu vymezených povinných subjektů. Tímto subjektem je běžně míněn některý veřejnoprávní subjekt, který je definován v uzavřeném výčtu dle §2 tohoto zákona. Jsou zde také definované podmínky pro povinnost uveřejnění smlouvy, jako je například minimální výše hodnoty plnění smlouvy 50 000 Kč nebo plnění smlouvy převážně na území České republiky. Subjekt musí dle tohoto zákona také splňovat podmínky platnosti uveřejnění, jako je například nutnost uveřejnění 30 dnů od jejího uzavření nebo (Kraus, 2022).

TODO možné rozšíření…uvézt i některé výjimky uveřejnění?

### Registr smluv

TODO možné rozšíření…jaké smlouvy do registru smluv patří

Registr smluv je informačním systém veřejné správy, který slouží k uveřejňování smluv dle platného zákona. Správcem tohoto IS je Ministerstvo vnitra, které ale žádným způsobem neodpovídá za správnost nahraných smluv. Samotná kontrola správnosti smlouvy a především kontrola znečitelnění obsahu (neboli anonymizace či úplné odstranění některých údajů) spadá na subjekt, který smlouvu uveřejňuje (Kraus, 2022).

Smlouvy je možné do registru zasílat přes grafické rozhraní skrz elektronický formulář, dostupný na Portálu veřejné správy nebo přes systém datových schránek.

Prostřednictvím registru smluv není možné zveřejňovat informace, které nelze poskytnout dle 106/1999 Sb. Zákona o svobodném přístupu k informacím. Těmi jsou v kontextu této práce a možností algoritmického zpracování dat pomocí hlubokého učení osobní údaje a obchodní tajemství, jehož detekce však není cílem této práce.

Práce s osobními údaji je dále omezená 110/209 Sb. Zákonem o zpracování osobních údajů a potažmo tedy i nařízením Evropského parlamentu a rady EU GDPR. Vzhledem k tomu, že při uveřejnění smlouvy není k dispozici oprávněný zájem ani jiný právní titul, který by tuto operaci ospravedlňoval (ÚOOÚ, 2016) a získání souhlasu by znamenalo další interakci se subjektem smlouvy, je nutné všechny osobní údaje dle jejich platné definice v uveřejňovaných smlouvách řádně anonymizovat. V případě neexistující či nekompletní anonymizace může Úřad pro ochranu osobních údajů (ÚOOÚ) udělit odpovědnému subjektu pokutu nebo se případně poškozená fyzická osoba může domáhat náhrady způsobené újmy (Kraus, 2022).

### Nástroj pro anonymizaci dokumentů

Ministerstvo vnitra poskytuje přes Portál veřejné správy pro účely anonymizace textů uveřejňovaných smluv Nástroj pro anonymizaci dokumentů, který však mohou využívat pouze orgány veřejné moci. Tento nástroj je dostupný na adrese <https://anonymizace.gov.cz/crossroad/>.

Tento nástroj umožňuje řádným způsobem (znečitelněním všech vrstev textu) anonymizovat nahraný dokument. Dle textového popisu pro nepřihlášené uživatele však nenabízí automatickou detekci klíčových identifikátorů, a to ani v podobě jejich vyhledávání přes předdefinované šablony bez znalosti širšího kontextu textu. Model vyvíjený v další části této práce by se tedy mohl stát dalším rozšířením tohoto nástroje a pomoci tak pracovníkům orgánů veřejné moci lépe a rychleji anonymizovat nahrávané dokumenty.

Na obrázku 1 lze vidět část úvodní intrografiky dostupné na webu Portálu veřejné správy. Lze z něho vyčíst, že strojové zpracování textu, neboli Optical Character Recognition (OCR) při nahrání textu v obrazovém formátu nástroj již zvládá. Začernění osobních údajů a obchodního tajemství se ale stále musí provádět ručně.



Obrázek : Nástroj pro anonymizaci dokumentů – intrografika

TODO

* Omezení tohoto nástroje
* Vyjádření o nástroji dle mailu
* Existující externí nástroje? To bych možná mohl dát spíš někam jinam

## Typy vstupních dokumentů

Smlouvy smí do registru smluv být nahrávané pouze v elektronické podobě v několika povolených formátech. Elektronický obraz textového obsahu smlouvy se do registru zveřejňuje v otevřeném a strojově čitelném formátu, případně v otevřeném formátu umožňujícím úplně strojové zpracování textového obsahu (Kraus, 2022).

Otevřený a strojově čitelný formát definuje zákon 106/1999 Sb. o svobodném přístupu k informacím.

Otevřený formát nesmí být závislý na konkrétním technickém a programovém vybavení. To obvykle splňují formáty, které jsou udržovány neproprietárními standardizačními organizacemi a které nemají právní omezení pro jejich používání (např. PDF, ODT).

Strojově čitelný formát obsahuje takovou vnitřní strukturu, která umožňuje při dalším elektronickém zpracování extrahovat ze souboru čitelný a nijak nepozměněný text od originálu.

Do 31. prosince 2023 je možné do registru smluv nahrát i dokumenty, které sice umožňují úplné strojové zpracování textového obsahu, ale které v sobě neobsahují další informace o vnitřní struktuře dat, jako jsou informace o nadpisech, podnadpisech nebo poznámkách pod čarou. Vždy je ale nutné, aby dokument obsahoval čitelnou textovou vrstvu.

Tato informace je důležitá pro implementaci modelu hlubokého učení, který bude text zpracovávat. Znamená totiž, že metadata u jednotlivých textů nemusí být vždy dostupná a nelze se na ně tedy spolehnout.

Do registru smluv se nahrává smlouva ve stejném formátu, v jakém byla v originále podepsána, a to včetně všech jejích dodatků, které by měly být součástí jednoho dokumentu. Subjekt, který do registru smlouvu nahrává musí vyplnit i metadata. Ta by měla obsahovat základní informace o smlouvě, jako je identifikace smluvních stran, vymezení předmětu smlouvy, datum jejího uzavření a případně, pokud jí lze určit, i cenu předmětu vyjednávání. Informace uvedené v metadatech však musí splňovat stejná nařízení o nezveřejnitelných informacích jako samotný text smlouvy.

Součástí nahrání smlouvy mohou být i přílohy. Obrazové přílohy obvykle nutnosti nahrávání nenaplňují. Textové přílohy musí splňovat stejné formální požadavky pro správný formát jako samotná smlouva.

## Kategorizace osobních údajů

Na základě předpokládaného výskytu osobních údajů, běžně se vyskytujících ve smlouvách nahrávaných do registru smluv podle Metodického návodu k aplikaci Zákona o registru smluv (Kraus, 2022) a na základě úvahy autora byl vytvořen seznam specifických osobních údajů, které se pravděpodobně budou vyskytovat ve zpracovávaných dokumentech (TODO tohle by mělo být lépe podloženo, mělo by to být minimálně ještě na základě e-mailové komunikace s portálem veřejné správy a na základě analýzy smluv v datasetu. Taky můžu udělat nějakou analýzu návrhových vzorů pro smouvy).

Údaje jsou roztříděné do několika základních kategorií primárně dle formátu, ve kterém se údaj v textu vyskytuje (sekvence čísel, znaků, apod.).

**Jmenné identifikátory**

* Jméno
* Příjmení
* Příjmení za svobodna
* E-mail osobní
* E-mail pracovní
* Přihlašovací údaje (login)

**Číselné identifikátory**

* Rodné číslo/číslo pojištěnce
* Číslo občanského průkazu (OP)
* Číslo pasu
* Číslo řidičského průkazu
* Číslo platební karty
* Číslo jiné osobní karty (např. členské karty)
* Daňové identifikační číslo (DIČ)
* Číslo bankovního účtu
* Sdružené inkaso plateb obyvatelstva (SIPO)
* IBAN
* Telefonní číslo osobní
* Telefonní číslo pracovní
* Datum narození
* Věk

**Znakové identifikátory**

* Státní poznávací značka (SPZ)
* MAC adresa
* IP adresa
* Internetová doména
* Podpis
* Elektronický podpis
* Elektronický klíč
* Biometrická data
* Identifikační číslo vozidla (VIN)
* Parcelní číslo
* Sériové číslo
* GPS souřadnice

**Nezařazené**

* Adresa
* Pohlaví
* Rasa
* Náboženské vyznání
* Sexuální vyznání
* Titul
* Pracovní zařazení
* Politická příslušnost
* Státní příslušnost
* Obchodní tajemství
* Cena zakázky

**Identifikátory nepodléhající anonymizaci**

* Název firmy
* Identifikační číslo osoby (IČO)
* Id datové schránky

Tento základní seznam ale není k dalšímu zpracování dostatečný, protože nereflektuje zvláštní situace, kdy některý údaj povinné anonymizaci může a nemusí podléhat. Zároveň nevystihuje jednotlivé vazby údajů mezi sebou a také v něm chybí rozpad na konkrétnější identifikátory nižší úrovně.

### Kompletní seznam

V další fázi tedy byla po hlubší analýze vytvořena tabulka, která obsahuje podrobnější seznam jednotlivých identifikátorů, rozepsaných více dopodrobna.

Každému identifikátoru bylo přiděleno unikátní číslo ID, název, příznak, zda je identifikátor primární nebo sekundární, seznam závislých identifikátorů a příznak, zda je identifikátor specifický využitím v dokumentech typu smlouva.

Pole **Typ identifikátoru** vyjadřuje, zda je daný identifikátor v kontextu vždy osobním údajem (v tomto případě je označen jako primární) nebo se osobním údajem stává až při vzniku vazby na další identifikátor (v tomto případě je označen jako identifikátor sekundární). To si lze jednoduše představit na jednoduchém příkladu křestního jména a příjmení. Samo o sobě není křestní jméno tak silným identifikátorem, aby dokázalo identifikovat konkrétní osobu na kterou odkazuje. Ve spojení s příjmením se ovšem množina lidí, která může být tímto identifikátorem označena značně zmenšuje a v tomto kontextu lze tedy již prohlásit, že se jedná o osobní údaj dle platné definice.

Existují samozřejmě výjimky, kdy se jedním celým jménem (křestní jméno + příjmení) dá označit více osob a naopak křestní jméno nebo příjmení může být tak unikátní, že je přiřazeno jen jedné osobě. Hranice může být v tomto případě velmi tenká. Tato práce se však ve své základní podobě zabývá standartními situacemi.

V poli **Závislé identifikátory (ID)** jsou v případě sekundárních identifikátorů uvedeny odkazy na další identifikátory, na kterých je v kontextu textu daný identifikátor závislý ve smyslu toho, že pokud daná vazba v textu existuje, stává se kombinace těchto identifikátorů osobním údajem.

V poli **Identifikátor specifický pro dokument typu smlouva** se nachází informace, zda je nutná detekce tohoto identifikátoru specifická přímo pro využití detektoru v kontextu smluv. Například identifikátor, který odkazuje na adresu bydliště je v kontextu běžného textu téměř vždy osobním údajem. V kontextu dokumentu typu smlouva zde však existuje závislost na tom, jestli se jedná o adresu-bydliště fyzické osoby či o adresu-sídlo právnické osoby. V druhém jmenovaném případě totiž tento údaj nepodléhá povinné anonymizaci.

To, jestli je identifikátor specifický jen pro smlouvy, je také důležité rozdělení jednotlivých detektorů do vhodných modulů tak, aby se v případě využití modulárního přístupu k vývoji aplikace daly některé obecné části znovu využít i v případě přepracování modelu na detekci osobních údajů v jiném kontextu.

V případě, kdy identifikátor nepodléhá povinné anonymizace, nejsou vyplněny hodnoty v žádném dalším sloupci kromě ID a název identifikátoru.

Kompletní seznam identifikátorů lze naleznout v příloze XXX. TODO zjistit jak správně vložit kompletní tabulku do přílohy a jak na ní správně odkazovat

## Mapa kategorií

V kompletním seznamu se pracuje s počtem 53 identifikátorů. Tento počet je v další fázi redukován pomocí postupného logického zjednodušování, kdy je bráno v potaz konkrétní využití aplikace na detekci osobních údajů na dokumentech typu smlouva. Také zde dochází k odstranění některých duplicitních příznaků u jednotlivých identifikátorů.

Kvůli své podstatě sekundárních identifikátorů vede tato mapa k následnému vytvoření n-rozměrné vazební matice binárních hodnot, kdy n je rovno celkovému počtu zpracovávaných identifikátorů.

Zjednodušenou mapu kategorií osobních údajů lze najít v tabulce 1.

|  |  |  |  |
| --- | --- | --- | --- |
| **Mapa kategorií osobních údajů** | | | |
|  |
| **ID** | **Identifikátor** | **Typ identifikátoru** | **Závislé identifikátory (ID)** |  |
| **Jmenné identifikátory** | | | |  |
| 1 | Křestní jméno | sekundární | 2, 7, 9, 53 |  |
| 2 | Příjmení | sekundární | 1, 7, 9, 53, 56 |  |
| 4 | Jméno + Příjmení | primární | 1, 2 |  |
| 6 | E-mail s unikátním jménem | primární |  |  |
| 7 | E-mail pracovní s obecným jménem | sekundární | 1, 2, 9, 53 |  |
| 9 | Přihlašovací údaje (login) | sekundární | 1, 2, 7, 10, 32, 33 |  |
| 10 | Název společnosti | - | - |  |
| **Číselné identifikátory** | | | |  |
| 12 | Rodné číslo/číslo pojištěnce | primární |  |  |
| 13 | Číslo OP | primární |  |  |
| 14 | Číslo pasu | primární |  |  |
| 15 | Číslo platební karty | primární |  |  |
| 16 | Číslo jiné karty | sekundární | vydavatel karty, 10 |  |
| 17 | Číslo bankovního účtu | sekundární | 59 |  |
| 19 | IBAN | sekundární | 59 |  |
| 21 | Telefonní číslo | primární |  |  |
| 23 | Číslo řidičského průkazu | primární |  |  |
| 24 | Věk (sec) | sekundární | 1, 2, 3, 53 |  |
| 25 | SIPO | sekundární | 59 |  |
| **Znakové identifikátory** | | | |  |
| 30 | SPZ | sekundární | 1, 2, záznam o jízdě |  |
| 31 | MAC adresa | primární |  |  |
| 32 | IP adresa | sekundární | 9, 10, 53 |  |
| 33 | Doména | sekundární | 9, 10, 53 |  |
| 38 | VIN | sekundární | 1, 2, záznam o jízdě |  |
| 40 | Sériové číslo | sekundární | 1, 2, 10 |  |
| 42 | Daňové identifikační číslo (DIČ) | sekundární | 59 |  |
| **Lokalizační identifikátory** | | | |  |
| 45 | Adresa – bydliště | primární |  |  |
| 46 | Adresa – sídlo PO nebo FOP | - | - |  |
| **Ostatní identifikátory** | | | |  |
| 49 | Náboženské vyznání | sekundární | nepřímá závislost |  |
| 52 | Titul | sekundární | nepřímá závislost |  |
| 53 | Pracovní zařazení | sekundární | 1, 2, 7, 10, 32, 33 |  |
| 54 | Politická příslušnost | sekundární | nepřímá závislost |  |
| 55 | Státní příslušnost | sekundární | nepřímá závislost |  |
| 56 | Datum narození | sekundární | 2, 3, 10 |  |
| 59 | Právnická osoba (PO)/Fyzická osoba podnikající (FOP) | - | - |  |

Tabulka : Mapa kategorií osobních údajů

### Jmenné identifikátory

V této skupině došlo ke sjednocení všech identifikátorů vyjadřujících jméno osoby do 3 kategorií: křestní jméno, příjmení a celé jméno, které vzniká spojením 2 předchozích. Kvůli tomu, že všechna celá jména budou v konečném důsledku stejně podléhat anonymizaci, identifikátor označující jednatele/zástupce firmy nedává v samostatném významu smysl.

E-mail byl zařazen do 2 kategorií: E-mail s unikátním jménem (ať už pracovní nebo osobní) a e-mail s obecným jménem, který si lze představit například jako info@vse.cz. Ten se dá za osobní údaj uvažovat jen pokud jsou známé dodatečné informace jako například příjmení osoby.

Název společnosti sám o sobě anonymizaci nepodléhá. Je ho ovšem potřeba brát v potaz kvůli vazbám dalších sekundárních identifikátorů, které jsou na tomto údaji závislé.

### Číselné identifikátory

Primární identifikátory uvedené v kompletním seznamu zůstávají principiálně ve stejné podobě, jen dochází ke sjednocení osobního a pracovního telefonního čísla a především k vytvoření vazby na identifikátoru *Právnická osoba (PO)/Fyzická osoba podnikající (FOP)*.

Identifikátory označující identifikační číslo osoby (IČO) byly zcela odstraněny, protože tento údaj v daném kontextu nepodléhá anonymizaci.

### Znakové identifikátory

Z této skupiny byl odstraněn identifikátor vyjadřující podpis. Vyvíjený program totiž nezpracovává obrazová data a k vyfiltrování podpisu v obrazové podobě by tak mělo dojít ještě v předchozím zpracování. Pokud je podpis uveden jen v textovém formátu, spadá již pod identifikátor vyjadřující jméno osoby.

Stejně tak byla vyřazena i textová reprezentace elektronického podpisu a klíče. U těchto údajů se totiž tento formát zobrazení nepředpokládá. Stejně tak jsou vyjmuta i biometrická data, která by se ve smlouvě vůbec neměla v žádné formě vyskytovat.

### Lokalizační identifikátory

Adresa obecně vždy podléhá anonymizaci, pokud není v kontextu smlouvy zároveň sídlem PO nebo i FOP. Z toho se dá tedy odvodit, že tento identifikátor je ve své základní podobě primární, avšak dá se na něj pohlížet i jako na sekundární, přičemž jako sekundární anonymizaci nepodléhá.

### Ostatní identifikátory

Identifikátor *Rasa* byl z této skupiny odstraněn kvůli velmi nepravděpodobnému výskytu ve smlouvách. To stejné platí i o identifikátoru *Sexuální vyznání*. *Pohlaví* bylo odstraněno kvůli malé identifikační vypovídající hodnotě a zároveň kvůli složitostem při detekci pohlaví v českém jazyce, kde se dá pohlaví určit jen podle tvaru slova.

Cena zakázky by měla být dle metodických doporučení uvedena přímo v předmětu smlouvy a anonymizaci tak podléhá jen ve velmi specifických případech, které jsou velmi složité a značně nekonkrétní. Cena může souviset i s obchodním tajemstvím, jehož detekce není v této verzi aplikace umožněna.

### Shrnutí

Po redukci kompletního seznamu identifikátorů se podařilo snížit počet identifikátorů z 53 na 34. Do vytvořené mapy se dají v budoucnu při rozšíření stávajícího využití nebo při novém stanovení typu analyzovaných dokumentů přidávat i další identifikátory, které mohou být navázané na již stávající. Díky modulární struktuře vyvíjeného modelu by mělo být možné relativně jednoduše rozšířit stávající funkcionalitu vyvíjeného klasifikátoru.

TODO Udělat nějakou paralelu s ML checklistem jako závěrečné zhodnocení kapitoly

# Použité techniky hlubokého učení

Úvod k hlubokému učení

* Obecná teorie od základu ML po větší zúžení a uvedení do oblasti NLP a NER.
* Zaměření na multilingvní modely a na modely zpracovávající kontext textu
* Model by měl být vytvořen z více modelů závislých na typu údaje a využití tak, aby byla možnost ho rozšiřovat o další modely více závislé na kontextu

## Hluboké učení

Hlubokým učením (deep learning) se nazývá podoblast strojového učení (machine learning). Hluboké neuronové sítě, tedy neuronové sítě patřící právě do oblasti hlubokého učení vynikají především větším stupněm abstrakce, která je způsobena větším počtem skrytých vrstev (hidden layers).

Známým příkladem, který ukazuje, jak větší počet skrytých vrstev dokáže ovlivňovat složitost požadované operace, je trénink neuronové sítě tak, aby dokázala provádět operaci XOR. Zatímco binární operace jako AND a OR jsou lineárně separovatelné v dvourozměrném prostoru jednou polorovinou, u XOR už toto tvrzení neplatí a pro úspěšné vyjádření této funkce je potřeba do neuronové sítě přidat skrytou vrstvu.

Zároveň je dokázáno, že jakákoliv funkce vyjádřená ve dvourozměrném prostoru může být neuronovou sítí simulována dostatečným počtem výpočetních jednotek (computation units) v pouze jedné skryté vrstvě, přičemž toto tvrzení vytváří paralelu s Fourierovou transformací. Tento přístup je však pro praktické použití nevýhodný, protože s narůstající přesností simulace funkce roste exponenciálně i počet výpočetních jednotek ve skryté vrstvě a tím pádem i celková výpočetní náročnost (Wang, 2017).

TODO obrázek ukázky vyšší abstrakce, možná přímo ten XOR?

Jako hluboké neuronové sítě jsou definovány takové neuronové sítě, které mají 3 a více skrytých vrstev a pracují s více úrovněmi reprezentace informací. Princip většího počtu skrytých vrstev je bližší fungování lidského mozku, který také tíhne k vytváření větší abstrakce a postupnému spojování dílčích informací (Bengio a Delalleau, 2011).

Dá se tedy říci, že zatímco mělké neuronové sítě (s jednou skrytou vrstvou) se budou při úloze detekce automobilu v obrázku snažit hledat automobil jako celek, hluboké neuronové sítě budou v jednotlivých vrstvách nejdříve detekovat dílčí části automobilu, jako jsou kola, karoserie, a podobně.

Eldan a Shamir ve své práci *The Power of Depth for Feedforward Neural Netoworsk* (2015) zjistili a dokázali, že více skrytých vrstev neuronové sítě dosahuje z hlediska výpočetní náročnosti v poměru s přesností modelu lepší výsledky. Zároveň však mohou nastat situace, kdy je přidávání dalších vrstev spíše na škodu, protože model může v důsledku ve vyšších vrstvách vytvářet i abstraktní vazby takové úrovně, že z hlediska řešeného problému nemusejí dávat smysl.

Zároveň obecně platí, že čím je model neuronové sítě hlubší, tím je i náročnější na trénování algoritmem zpětné propagace chyby (backpropagation algorithm), protože trénování se značně zpomaluje a zároveň se zvyšuje problém uváznutí v lokálních minimech kvůli špatné volbě počátečních vah. Také je v těchto případech pro trénování potřeba většího množství trénovacích dat (Hinton, 2006).

Tento problém adresoval v roce 2006 ve své práci *A fast learning algorithm for deep belief nets* Geoffrey Hinton, který přišel s greedy algoritmem, který umožňuje trénovat jednotlivé vrstvy odděleně, a tím otevřel cestu k řadě dalším výzkumům v oblasti hlubokého učení pomocí jemného ladění (fine tuning) ve vyšší vrstvách, bez přístupu k obrovskému množství trénovacích dat.

* Co to je, jak to navazuje na strojové učení, co se v něm využívá, co to vlastně je, jaký je matematický princip a jaké jsou běžné typy modelů – svm, neuronovka,…

## Natural Language Processing

Techniky patřící do oblasti Natural Language Processing (dále NLP) se zabývají strojovým zpracováním jazyka. Jedná se tedy o činnost, která je pro člověka relativně jednoduchá (ve většině případech), ale v případě strojového zpracování zde věda naráží na obtížnost vyjádření myšlenky, kontextu či emocí pomocí matematických pravidel (Johri et al., 2021).

NLP samo o sobě není přímou součástí hlubokého učení. Výzkum strojového zpracování jazyka začal ještě dávno před tím, než se objevily první záznamy o hlubokém učení a o hlubokých neuronových sítích. Nejprve se svět NLP dlouho zabýval vytvořením automatického překladače, který by dokázal spolehlivě automaticky překládat mezi různými jazyky. To vedlo k vytvoření časově náročného projektu, který si dal za cíl vytvořit překladové strojové slovníky. To se podařilo Georgesi Artstrounimu, který si nechal roku 1933 patentovat zařízení s názvem „Mechanical brain“, které sloužilo jako mechanický překladač (machinetranslate.org, nedatováno).

Tento přístup se ale z praktického používání v reálném světě ukázal jako nepoužitelný, protože tradiční slovníkový přístup si nedokáže poradit s gramatikou. V roce 1957 přišel Noam Chomsky s teorií o lingvistických strukturách, která měla obecně definovaným setem pravidel definovat matematické vztahy mezi jednotlivými slovy ve větách. Později se ale ukázalo, že tento přístup nedokáže nikdy komplexně gramatiku vystihnou kvůli rozdílům v jednotlivých jazycích (Johri et al., 2021).

Dalším evolučním krokem byla tokenizace slov ve větě, jejich klasifikace a zpracovávání kontextu dle pořadí tokenů ve stromové struktuře představující větu. Hojně se při tom využívala rekurze, kdy docházelo k postupnému zpracovávání všech tokenů, dokud nebyl rozpoznán význam textu. Taková technika NLP se nazývá Augmented Transition Network (ATN) a byla popsána roku 1970 Williamem Woodsem (Woods, 1970).

Všechny výše popsané metody měly jednu zásadní nevýhodu: všechna pravidla pro zpracování textů musela být přesně definována a algoritmicky popsána, což z tohoto úkolu dělalo nesmírně časově a intelektuálně náročnou činnost (nehledě na to, že způsob, jak některá pravidla programaticky vyjádřit ani nebyl nalezen).

V 80. letech 20. století se oblast NLP začala pomalu prolínat se světem strojového učení, které pomocí například rozhodovacích stromů dokázalo přesněji interpretovat jednotlivá gramatická a syntaktická pravidla na základě tréninku modelu z reálných dat. Stále zde ale existovala velká oblast nevyřešených problémů s určování významu slov na základě kontextu textu nebo rozdílné chápání textu na základě vztahu pisatele.

Nejnovějšími technikami NLP se tedy staly techniky hlubokého učení, které umožňují právě vyšší abstrakci chápání textu.

Klíčovou technikou se stalo vytváření matic reprezentujících vztahy mezi jednotlivými tokeny (těmi můžou být jak jednotlivá slova, tak například i dvojice slov, atd.). Samotné vytvoření této matice se provádí trénováním modelu na korpusu slov, které by se měly týkat stejné či příbuzné problematiky jako na kterou bude využitý i výsledný NLP model.

Běžnými technikami jsou například TF-IDF, One-Hot Encodings, Word Embeddings, Word2Vec a další. Cílem je vytvořit takovou n-rozměrnou matici, aby slova která jsou si významem a využitím ve větách příbuzná, měla k sobě co nejbližší vzdálenost (Nigam, 2021).

Základních typů neuronových sítí, které jsou v současné době považované jako nejvhodnější pro použití v NLP je několik. V první řadě jde o rekurentní neuronovou síť (Recurrent Neural Network – RNN), která dokáže zpracovat vstupní data (jedno slovo/jeden token) s ohledem na data zpracovaná při předchozím průchodu. Tím pádem dokáže při zpracování např. věty brát v potaz i kontext předchozích slov. Problém RNN spočívá v tom, že fungují pouze jednosměrně. To znamená, že slova následující ve větě později už nemohou ovlivnit zpracovaný význam slov předcházejících dříve ve větě. Zároveň zde existuje problém takzvaného mizejícího gradientu, kdy v delších textech postupně mizí vazba na předchozí zpracovaná data (Goldberg, 2017).

V současné době se tedy spíše využívají modifikované verze RNN, které tyto problémy adresují. Nejznámějšími jsou modely Gated Recurrent Unit (GRU) a Long Short Term Memory (LSTM), které řeší problém mizejícího gradientu představením takzvaných bran (gate), které umožňují lepší kontrolu nad oddělenou pamětí, představující kontext předchozích tokenů (Goldberg, 2017).

Problém nemožnosti vytvoření vazby na tokenech zpracovávaných později ve větě je řešen obousměrným RNN (Bidirectional RNN – BRNN). V tomto případě probíhá zpracovávání tokenů obousměrně, takže při zpracování konkrétního tokenu BRNN pracuje s error gradienty obou RNN (Goldberg, 2017).

* Nějaký obrázek který představuje třeba ve stromu rozložení věty
* Jaké jsou běžné problémy které se v této oblasti řeší
* Jaké jsou běžné úlohy
* Více rozvést vytváření word embeddings,tf-idf,…

## Named Entity Recognition

* Popsat jak je NER součástí NLP
* Najít nějaký paper, který dobře popisuje NER a jeho současný stav

## Multilingvní modely

## Context-based modely

* Nejprve by zde měla být část o tradiční detekci údajů bez hlubokého učení a tím i bez možnosti zpracování kontextu
* Je součástí NER – zaměřit se na modely, které dokáží udržet delší kontext, ne jen třeba 1 větu

## Transfer learning

* Napsat o závislosti na deep learningu – viz Deep Learning book, str. 526 Greedy Layer-Wise …

## Sekvenční modely

## Meta klasifikátory

# Dataset

Tohle asi moc nemá cenu psát moc dopředu. Existuje spousta způsobů jak se k datům dostat a jak s nimi dále pracovat. Možná bude potřeba vygenerovat vlastní dataset.

Víceméně jde o kapitoly 2-4 v ML checklistu.

## Získání trénovacích dat

## Průzkum dat

## Příprava dat

### Úpravy na základě příznakového modelu osobních údajů

## Reprezentace dat

### Word Embeddings

### Word2vec

Atd.

# Výběr technologií

## Výběr NLP/NER frameworku

## DL knihovny

## Vývojové prostředí

## Techniky průběžné evaluace

# Vývoj modelu

Tady si taky myslím že nemá moc cenu psát něco dopředu. Spíš to bude záznam procesu vývoje a vysvětlení zajímavých funkcí či ukázka zajímavých událostí v průběhu vývoje. Následné názvy kapitol jsou vyloženě jen ilustrační.

## Architektura

## Rozdělení dat

## Ztrátová funkce

## Fine-tuning

Atd.

# Evaluace modelu

TODO

* Tady by mělo dojít i k porovnání výsledků vůči ostatním modelům, které nejsou založené na NLP a tudíž nemohou zpracovávat i kontext

## Hodnocení naměřených kvalitativních indikátorů

## Automatická evaluace

## Evaluace člověkem

## Zhodnocení výsledků a porovnání

Závěr

Použitá literatura

BERT GITHUB, bert/multilingual.md at master · google-research/bert. *GitHub* [online] [vid. 2022-01-09]. Dostupné z: <https://github.com/google-research/bert>

CHAU, Ethan C., Lucy H. LIN a Noah A. SMITH, 2020. Parsing with Multilingual BERT, a Small Corpus, and a Small Treebank. *arXiv:2009.14124 [cs]* [online]. [vid. 2022-01-09]. Dostupné z: <http://arxiv.org/abs/2009.14124>

CLARK, Kevin, Minh-Thang LUONG, Quoc V. LE a Christopher D. MANNING, 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *arXiv:2003.10555 [cs]* [online]. [vid. 2022-01-09]. Dostupné z: <http://arxiv.org/abs/2003.10555>

DEVLIN, Jacob, Ming-Wei CHANG, Kenton LEE a Kristina TOUTANOVA, 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]* [online]. [vid. 2022-01-09]. Dostupné z: <http://arxiv.org/abs/1810.04805>

EK, Ochrana údajů. *Evropská komise - European Commission* [online] [vid. 2022-01-09]. Dostupné z: <https://ec.europa.eu/info/law/law-topic/data-protection_en>

ELLMANN, Mathias, 2018. Natural language processing (NLP) applied on issue trackers. In: *Proceedings of the 4th ACM SIGSOFT International Workshop on NLP for Software Engineering* [online]. New York, NY, USA: Association for Computing Machinery, s. 38–41 [vid. 2022-01-09]. NL4SE 2018. ISBN 978-1-4503-6055-5. Dostupné z: doi:[10.1145/3283812.3283825](https://doi.org/10.1145/3283812.3283825)

GEALFOW, John Altair a Christian MAY, 2019. Anonymizace osobních údajů v soudních rozhodnutích. *Revue pro právo a technologie* [online]. **10**(19), 3–39. ISSN 1805-2797. Dostupné z: doi:[10.5817/RPT2019-1-1](https://doi.org/10.5817/RPT2019-1-1)

GRISHMAN, Ralph a Beth SUNDHEIM, 1996. Message Understanding Conference-6: a brief history. In: *Proceedings of the 16th conference on Computational linguistics - Volume 1* [online]. USA: Association for Computational Linguistics, s. 466–471 [vid. 2022-01-09]. COLING ’96. Dostupné z: doi:[10.3115/992628.992709](https://doi.org/10.3115/992628.992709)

GÉRON, Aurelien, 2017. *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. 1st vyd. B.m.: O’Reilly Media, Inc. ISBN 978-1-4919-6229-9.

KUBICA, Jan, 2019. *Vybrané problémy technologické realizace evropské ochrany osobních údajů* [online]. B.m. [vid. 2022-01-09]. Univerzita Karlova. Dostupné z: <http://invenio.nusl.cz/record/397848>

MENDELS, Omri, 2020. NLP approaches to data anonymization. *Medium* [online] [vid. 2022-01-09]. Dostupné z: <https://towardsdatascience.com/nlp-approaches-to-data-anonymization-1fb5bde6b929>

MIHULKOVÁ, Jitka, 2018. *Co je, co není a co bude osobní údaj podle GDPR - Frank Bold Advokáti* [online] [vid. 2022-01-09]. Dostupné z: <https://www.fbadvokati.cz/cs/clanky/541-co-je-co-neni-a-co-bude-osobni-udaj-podle-gdpr>

MISHRA, Aditya, 2018. Metrics to Evaluate your Machine Learning Algorithm. *Medium* [online] [vid. 2022-01-09]. Dostupné z: <https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234>

MOBERG, John, 2020. A deep dive into multilingual NLP models. *Peltarion* [online] [vid. 2022-01-09]. Dostupné z: <https://peltarion.com/blog/data-science/a-deep-dive-into-multilingual-nlp-models>

MVČR. *Zvláštní kategorie osobních údajů - Ochrana osobních údajů* [online] [vid. 2022-01-09]. Dostupné z: <https://www.mvcr.cz/gdpr/clanek/zvlastni-kategorie-osobnich-udaju.aspx>

NADEAU, David a Satoshi SEKINE, 2007. A survey of named entity recognition and classification. *Lingvisticæ Investigationes* [online]. **30**(1), 3–26. ISSN 0378-4169, 1569-9927. Dostupné z: doi:[10.1075/li.30.1.03nad](https://doi.org/10.1075/li.30.1.03nad)

QU, Chen, Weize KONG, Liu YANG, Mingyang ZHANG, Michael BENDERSKY a Marc NAJORK, 2021. Natural Language Understanding with Privacy-Preserving BERT. In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* [online]. New York, NY, USA: Association for Computing Machinery, s. 1488–1497 [vid. 2022-01-09]. ISBN 978-1-4503-8446-9. Dostupné z: <http://doi.org/10.1145/3459637.3482281>

SILVA, Paulo, Carolina GONÇALVES, Carolina GODINHO, Nuno ANTUNES a Marilia CURADO, 2020. Using natural language processing to detect privacy violations in online contracts. In: *Proceedings of the 35th Annual ACM Symposium on Applied Computing* [online]. New York, NY, USA: Association for Computing Machinery, s. 1305–1307 [vid. 2022-01-09]. ISBN 978-1-4503-6866-7. Dostupné z: <http://doi.org/10.1145/3341105.3375774>

WILLIAMS, Jake a Giovanni SANTIA, 2017. Context-Sensitive Recognition for Emerging and Rare Entities. In: *Proceedings of the 3rd Workshop on Noisy User-generated Text* [online]. Copenhagen, Denmark: Association for Computational Linguistics, s. 172–176 [vid. 2022-01-09]. Dostupné z: doi:[10.18653/v1/W17-4423](https://doi.org/10.18653/v1/W17-4423)

XU, Hanchen, Zhenxiang CHEN, Shanshan WANG a Xiaoqing JIANG, 2021. Chinese NER Using ALBERT and Multi-word Information. In: *ACM Turing Award Celebration Conference - China ( ACM TURC 2021)* [online]. New York, NY, USA: Association for Computing Machinery, s. 141–145 [vid. 2022-01-09]. ISBN 978-1-4503-8567-1. Dostupné z: <http://doi.org/10.1145/3472634.3472667>

BOUDA, Petr, 2019. *Aktuálně k registru smluv | Právní prostor* [online] [vid. 2022-02-13]. Dostupné z: <https://www.pravniprostor.cz/clanky/spravni-pravo/aktualne-k-registru-smluv>

KRAUS, Martin, 2022. *Metodický návod k aplikaci ZRS*. 5. leden 2022. B.m.: MVČR.

ÚOOÚ, 2016. *Ochrana osobních údajů v souvislosti se zákonem o registru smluv: Názory a rozhodnutí Úřadu: Úřad pro ochranu osobních údajů* [online] [vid. 2022-02-14]. Dostupné z: <https://www.uoou.cz/ochrana-osobnich-udaju-v-souvislosti-se-zakonem-o-registru-smluv/d-20322/p1=1099>

WANG, Haohan a Bhiksha RAJ, 2017. On the Origin of Deep Learning. *arXiv:1702.07800 [cs, stat]* [online]. [vid. 2022-03-03]. Dostupné z: <http://arxiv.org/abs/1702.07800>

BENGIO, Yoshua a Olivier DELALLEAU, 2011. On the Expressive Power of Deep Architectures. In: Jyrki KIVINEN, Csaba SZEPESVÁRI, Esko UKKONEN a Thomas ZEUGMANN, ed. *Algorithmic Learning Theory* [online]. Berlin, Heidelberg: Springer, s. 18–36. Lecture Notes in Computer Science. ISBN 978-3-642-24412-4. Dostupné z: doi:[10.1007/978-3-642-24412-4\_3](https://doi.org/10.1007/978-3-642-24412-4_3)

ELDAN, Ronen a Ohad SHAMIR, 2015. The Power of Depth for Feedforward Neural Networks [online]. [vid. 2022-03-05]. Dostupné z: <https://arxiv.org/abs/1512.03965v4>

HINTON, Geoffrey E., Simon OSINDERO a Yee-Whye TEH, 2006. A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation* [online]. **18**(7), 1527–1554. ISSN 0899-7667, 1530-888X. Dostupné z: doi:[10.1162/neco.2006.18.7.1527](https://doi.org/10.1162/neco.2006.18.7.1527)

JOHRI, Prashant, Sunil Kumar KHATRI, Ahmad AL-TAANI, Munish SABHARWAL, Shakhzod SUVANOV a Avneesh CHAUHAN, 2021. Natural Language Processing: History, Evolution, Application, and Future Work. In:  [online]. s. 365–375. ISBN 9789811597114. Dostupné z: doi:[10.1007/978-981-15-9712-1\_31](https://doi.org/10.1007/978-981-15-9712-1_31)

machinetranslate.org [b.r.]. Georges Artsrouni. *Machine Translate* [online] [vid. 2022-03-13]. Dostupné z: <https://machinetranslate.org/georges-artsrouni>

WOODS, W. A., 1970. Transition network grammars for natural language analysis. *Communications of the ACM* [online]. **13**(10), 591–606. ISSN 0001-0782. Dostupné z: doi:[10.1145/355598.362773](https://doi.org/10.1145/355598.362773)

NIGAM, vibhor, 2021. Natural Language Processing: From Basics, to using RNN and LSTM. *Analytics Vidhya* [online]. [vid. 2022-03-13]. Dostupné z: <https://medium.com/analytics-vidhya/natural-language-processing-from-basics-to-using-rnn-and-lstm-ef6779e4ae66>

GOLDBERG, Yoav, 2017. Neural Network Methods for Natural Language Processing. *Synthesis Lectures on Human Language Technologies* [online]. **10**(1), 1–309. ISSN 1947-4040. Dostupné z: doi:[10.2200/S00762ED1V01Y201703HLT037](https://doi.org/10.2200/S00762ED1V01Y201703HLT037)

Poznámka: Za koncem Použité literatury nutné nesmazat *konec oddílu*, aby fungovalo různé číslování hlavního textu práce a příloh.

Přílohy

1. Název první přílohy
2. Název druhé přílohy