**Detekce osobních údajů pomocí metod hlubokého učení v nestrukturovaném textu**

**David Ondrášek**

**Úvod**

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.

U vyvinutého nástroje bude vhodně evaluována jeho spolehlivost a použitelnost.

**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 projektu**

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.

**Rešerše literatury**

**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.

**Metody použité k dosažení cílů**

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.

**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**

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.

**Zdroje**

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>