Anonymizace osobních údajů pomocí metod hlubokého učení

David Ondrášek

Úvod

Není to tak dávno, co pojem automatizace představoval jen zavádění robotů a automatických procedur do továren a byl spojen pouze s průmyslovým sektorem. Dnes si bez těchto technologií svět nedokážeme ani představit. V celém odvětví díky automatizace skokově vzrostla efektivita práce a klasické ruční manufaktury, které předtím existovaly, už téměř nikde v moderním světě nenajdeme.

Poslední dobou můžeme pozorovat nový trend: Stejný proces automatizace, který jsme dříve mohli vidět v průmyslovém sektoru přechází do oblasti služeb, kde se snaží nahradit rutinní kancelářské činnosti a nekreativní, nekvalifikované práce. Nezůstává ale jenom u toho. Velkou automatizační výzvou jsou profese, které potřebují k jejich vykonávání lidskou kreativitu, ale i ty jsou dnes už do jisté míry nahrazovány, za předpokladu, že činnost je dostatečně repetitivní nebo běžná po celém světě.

Celkově dnešní doba automatizaci velmi nahrává. Kvůli velké nezaměstnanosti je sehnat zaměstnance velmi obtížné a hlavně drahé. A i když se to povede, existuje neustálé riziko, že zrovna když se zaměstnanec naučí způsobům dané firmy a začne být ve vykonávání své práce opravdu efektivní, dostane nějakou jinou zajímavější nabídku jeho vysněné práce za vysněnou mzdu a firma, která si zaměstnance vycvičila může zase začít celý cyklus znovu od začátku.

Klíčovou činností, kterou je třeba často v automatizačních úlohách vykonávat, je převod nestrukturovaných dokumentů do digitální podoby. Tato operace se často provádí pomocí algoritmů strojového učení. I v Česku vznikají startupy, které se touto nebo podobnou problematikou zabývají a které aspirují na nejrychleji rostoucí české startupy v historii. Konkrétně společnost Rossum, založená výzkumníky v oblasti umělé inteligence z ČVUT, vybrala v prvním kole financování 2,2 miliardy korun (Tmejová 2021).

Tato diplomová práce zkoumá z hlediska ochrany osobních údajů důsledky, které pramení z automatizovaného zpracování nestrukturovaných dokumentů. Na základě toho je v ní potom navržen a implementován nástroj, který pomocí metod hlubokého učení umožňuje v dokumentu anonymizovat údaje, které byly v předchozí analýze vyhodnoceny jako údaje osobní. Implementovaný nástroj je dle předchozí analýzy upraven pro konkrétní vhodně vybranou aplikaci a je evaluována jeho spolehlivost a použitelnost.

KATEDRA INFORMAČNÍCH TECHNOLOGIÍ
Projekt diplomové práce v programu Informační systémy a technologie se specializací Vývoj IS

Vymezení problému

GDPR (General Data Protection Regulation), potažmo Zákon o zpracování osobních údajů, který tyto předpisy Evropské unie v Česku upravuje se po jeho příchodu v roce 2019 stal strašákem většiny podniků, které ke svojí práci potřebují zpracovávat větší i menší množství osobních údajů svých klientů. Dnes už se podniky tomuto zákonu dokázaly přizpůsobit do té míry, že jim ve většině případů nehrozí žádné finanční postihy.

Problém nastává v případě, kdy podnik potřebuje s ohledem na ochranu osobních údajů zpracovat nestrukturovaný dokument (například email nebo smlouvu), ve kterém není jasné, která část textu se dá považovat za osobní údaj a která ne. V případě zpracování tohoto dokumentu automatizovanými nástroji pro zpracování textu (například Rossum) je potom složité dodržet zásady pro ochranu osobních údajů a do automatizovaného procesu potom musí vstoupit kvalifikovaný pracovník, který tyto údaje dohledá.

Vymezení termínu osobní údaj

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 určit jednotlivé osobní, případně citlivé osobní údaje pro konkrétní aplikaci pro kterou bude nástroj upraven.

Jazyková a technická omezení

Pro úspěšné využití jakékoliv techniky hlubokého učení je potřeba mít k dispozici kvalitní dataset, na kterém se může vznikající model učit. 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)

Nástroje a techniky pro anonymizaci osobních údajů

Navzdory tomu, že v oblasti NLP je úloha zachování důvěry jednou ze současných klíčových důležitých výzev (Qu, 2021), jednotlivé organizace se prosazováním zásad ochrany osobních údajů nastavených ve zbytku firmy v automatizačních nástrojích pro zpracování textu moc nezabývají (Hua Li, 2006). Tato problematika tak často bývá řešena spíše na úrovni vhodného nastavení procesních pravidel organizace a možné informatické řešení, které by bylo vhodnější z důvodů menších omezení zaměstnanců a menšího času zaměstnanců stráveného ruční úpravou dokumentů tak zůstává nevyužito.

Účel a cíle projektu

Hlavním cílem této práce je navrhnout a implementovat prototyp nástroje, který pomocí algoritmů hlubokého učení dokáže v nestrukturovaném dokumentu najít osobní údaje, které byly předem vybrány a určeny analýzou skupiny nařízení o ochraně osobních údajů. Nástroj bude zprovozněn a evaluován na vhodně vybrané aplikaci.

Dílčím cílem práce je analyzovat zákony a nařízení, které podnikům mohou bránit ve zpracování firemních dokumentů a dále identifikovat údaje, které je potřeba anonymizovat a vypracovat model přístupnosti těchto údajů oprávněným osobám dle platného zákona.

Dalším dílčím cílem je výběr konkrétní aplikace, která vhodně poslouží pro správné vyzkoušení a evaluaci nástroje. Součástí tohoto cíle je také analýza využitelnosti nástroje na konkrétních byznys příkladech.

Rešerše literatury

Rešeršní strategie

Systematická část rešerše vhodné literatury proběhla převážně s využitím databáze ACM, Google Scholar a vyhledávačem Univerzity Karlovy UKAŽ. Na ACM a Google Scholar byly převážně vyhledávány práce více technického rázu, související s nejnovějšími poznatky NLP. V databázi UKAŽ pak byly vyhledávány práce související s ochranou osobních údajů, 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 jsme prohlédli prvních 20 článků a nejdříve podle názvu, následně podle abstraktu vybrali články, které byly svým obsahem nejblíže zkoumané problematice.

Vyhledávací řetězce pro ACM a Google Scholar:

- (NLP OR Natural Language Processing) AND multilingual
- (NLP OR Natural Language Processing) AND RPA
- (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
- (NLP OR Natural Language Processing) pseudonymization
- (NLP OR Natural Language Processing) data security

Vyhledávací řetězce 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 vyhledávání na webu 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.

Vymezení termínu osobní údaj

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, 2021).

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 (May, 2019).

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

Jazyková a technická omezení

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 píší ve většině případech v angličtině a v důsledku toho existuje i největší množství předtrénovaných modelů, které jsou založeny právě na anglických korpusech. Angličtina je 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, 2020). Populárním řešením je také využití revolučního modelu BERT, publikovaného týmem výzkumníků z Google v roce 2018 (Devlin, 2018), zejména pak jeho multilingvní varianty M-BERT, která v současné době podporuje 104 jazyků, včetně češtiny (BERT GITLAB). Problém v tomto případě může být imbalance datasetu, 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, 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).

Nástroje a techniky pro anonymizaci osobních údajů

Jak již bylo dříve zmíněno, pro anonymizaci osobních údaje je vhodné zvolit metody NLP, které jsou součástí technik hlubokého učení (Qu, 2021)(Silva, 2020)(Ellman, 2018). Jednotlivé metody z této oblasti lze pro dosažení optimálních výsledků potom různě kombinovat.

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, 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 anonymizaci, jako je využití regulárních výrazů uvnitř klasifikačních vrstev nebo vytvoření blacklistů s textovými řetězci s větši 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 (Ellman, 2018).

Chen Qu se zabývá konceptem typických architektur anonymizačních modelů a pokládá otázku, jak s anonymizovanými daty dokáží pracovat další vrstvy hlubokých neuronových sítí, které byly předučeny na neanonymizovaných datech (Qu, 2020).

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, 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. Vychází z ní výběr vhodného datasetu, jazyka a způsobu implementace pro danou aplikaci.

V neposlední řadě pak podle ní lze určit jednotlivé techniky, kterými bude samotná anonymizace údajů implementována.

Metody použité k dosažení cílů

Vzhledem k charakteru této práce byla pro implementační část vybrána metodika MMSP (Metodika pro Malé Softwarové Projekty) (Buchalcevová a Stanovská, 2013), která se hodí pro implementaci menších softwarových děl a tedy i této diplomové práce.

Samotný vyvinutý prototyp bude evaluován pomocí nejznámějších metrik využívaných při evaluaci deep learning modelů, jako je F1 score, Precision, Recall, aj. (Mishra, 2018).

Omezení projektu

Vzhledem ke komplexnosti celé problematiky ochrany osobních údajů a kvůli specifickému charakteru osobních údajů u různých typů aplikací je navržený nástroj omezen jen na jednu konkrétní vhodně vybranou aplikaci. Výběr této aplikace je jedním z dílčích cílů práce.

Dále je práce omezena limitujícím počtem českých datasetů vhodných pro trénování vyvíjeného modelu. Vytváření nutně komplexnějších datových sad není součástí této práce.

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.

Význam a přínos

Přínosem této práce je samotný nástroj na detekci osobních údajů v nestrukturovaných dokumentech, který slouží jako proof of concept využitelnosti algoritmů hlubokého učení při zpracování osobních údajů. Nástroj může být přizpůsoben různým aplikacím, jako například:

- vyfiltrování osobních údajů při digitalizaci firemního archivu
- zlepšení nástroje “Nástroj pro anonymizaci dokumentů” dostupném na Portálu veřejné správy (https://anonymizace.gov.cz/crossroad)
- umožnění přístupu k dokumentům obsahujícím osobní údaje neověřeným zaměstnancům

Dalším přínosem pak může být i samotná analýza zákonů a opatření o ochraně osobních údajů, která může posloužit organizacím v lepší orientaci na co se zaměřit, pokud by si daná organizace chtěla vyvinutý nástroj upravit nebo implementovat sama.

Zdroje

TMEJOVÁ, Kristýna. *Big deal! Český fakturový robot Rossum ulovil dvoumiliardovou investici* [online]. 20. 10. 2021 [cit. 2021-10-26]. Dostupné z: https://forbes.cz/big-deal-cesky-fakturovy-robot-rossum-ulovil-dvoumiliardovou-investici/

MIHULKOVÁ, Jitka. *Co je, co není a co bude osobní údaj podle GDPR* [online]. 10. 2. 2018 [cit. 2021-10-26]. Dostupné z: https://www.fbadvokati.cz/cs/clanky/541-co-je-co-neni-a-co-bude-osobni-udaj-podle-gdpr

QU, Chen, Weize KONG a kol. *Natural Language Understanding with Privacy-Preserving BERT* [online]. 2021 [cit. 2021-10-26]. Dostupné z: doi:https://doi-org.zdroje.vse.cz/10.1145/3459637.3482281

HUA LI, Yin, Hye-Young PAIK a kol. *Formal consistency verification between BPEL process and privacy policy* [online]. 9 2006 [cit. 2021-10-26]. Dostupné z: doi:https://doi-org.zdroje.vse.cz/10.1145/1501434.1501466

MOBERG, John. *A deep dive into multilingual NLP models* [online]. 24. 2. 2020 [cit. 2021-10-26]. Dostupné z: https://peltarion.com/blog/data-science/a-deep-dive-into-multilingual-nlp-models

EK. Oficiální internetová stránka Evropské unie [online]. [cit. 2021-11-05]. Dostupné z: https://ec.europa.eu/info/law/law-topic/data-protection_cs

KUBICA, Jan. *Vybrané problémy technologické realizace evropské ochrany osobních údajů*. Praha, 2021. Rigorózní práce. Univerzita Karlova, Právnická fakulta, Katedra evropského práva.

MVČR. *Ministerstvo vnitra České republiky* [online]. [cit. 2021-11-05]. Dostupné z: https://www.mvcr.cz/gdpr/clanek/zvlastni-kategorie-osobnich-udaju.aspx

MAY, Christian a John GEALFOW. Anonymizace osobních údajů v soudních rozhodnutích. *Revue pro právo a technologie* [online]. Masarykova univerzita, 2019, 10/2019 [cit. 2021-11-05], s.3-39, ISSN 1804-5383

SILVA, Paulo, Carolina GONCALVES a kol. *Using natural language processing to detect privacy violations in online contracts* [online]. 2020 [cit. 2021-11-06]. Dostupné z: doi:https://doi-org.zdroje.vse.cz/10.1145/3341105.3375774

ELLMAN, Mathias. *Natural language processing (NLP) applied on issue trackers* [online]. 2018 [cit. 2021-11-06]. Dostupné z: doi:https://doi-org.zdroje.vse.cz/10.1145/3283812.3283825

MENDELS, Omri. *Custom NLP Approaches to Data Anonymization* [online]. 2020, 8. 1. 2020 [cit. 2021-11-06]. Dostupné z: https://towardsdatascience.com/nlp-approaches-to-data-anonymization-1fb5bde6b929

BUCHALCEVOVÁ, Alena a Iva STANOVSKÁ, 2013. Příklady modelů analýzy a návrhu aplikace v UML. Praha: Oeconomica.

MISHRA, Aditya. *Metrics to Evaluate your Machine Learning Algorithm* [online]. 2018 [cit. 2021-11-06]. Dostupné z: https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

CLARK, Kevin, Minh-Thang LUONG a kol. *ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators* [online]. 2020 [cit. 2021-11-06]. Dostupné z: doi:arXiv:2003.10555v1

BERT GITLAB, Jacob a kol. *BERT Github repository* [online]. [cit. 2021-11-06]. Dostupné z: https://github.com/google-research/bert/blob/master/multilingual.md

CHAU, Ethan C., Lucy H. LIN a Noah A. SMITH. *Parsing with Multilingual BERT, a Small Corpus, and a Small Treebank* [online]. 2020 [cit. 2021-11-06]. Dostupné z: doi:arXiv:2009.14124v2

MOBERG, John. *A deep dive into multilingual NLP models* [online]. 2020 [cit. 2021-11-06]. Dostupné z: https://peltarion.com/blog/data-science/a-deep-dive-into-multilingual-nlp-models