Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Gašper Hüll

Anonimizacija sodnih odločb z metodami strojnega učenja

diplomsko delo

univerzitetni študijski program
prve stopnje
računalništvo in informatika

Mentor: prof. dr. Marko Robnik Šikonja

Ljubljana, 2020

Besedilo je oblikovano z urejevalnikom besedil LaTeX.

Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Tematika naloge:

Sodne odločbe vrhovnih sodišč vzpostavljajo sodno prakso, zato je potrebna njihova javna objava. Njihova anonimizacija služi zakrivanju občutljivih podatkov posameznikov in organizacij ter je zakonsko zapovedana. Za avtomatizacijo postopka anonimizacije se uporabljajo metode strojnega učenja, v zadnjem času so na tem področju najuspešnejše metode globokega učenja.

Preučite področje avtomatske anonimizacije in s pomočjo podatkovnih množic za prepoznavanje imenskih entitet izdelajte nekaj različnih modelov strojnega učenja za to nalogo. Vašo rešitev empirično ovrednotite.

Zahvaljujem se prof. dr. Marku Robniku Šikonji za mentorstvo pri diplomski nalogi. Njegovi nasveti in usmerjanje so bili nepogrešljivi pri nastajanju diplomskega dela.

Vsem, ki mi stojijo od strani

Kazalo

Povzetek

Abstract

1 Uvod **1**

2 Ozadje in pregled sorodnih del **3**

2.1 Ozadje 3

2.2 Tacita 4

2.3 Sorodna dela 5

3 Metodologija **7**

3.1 Podatkovne množice 7

3.2 Predprocesiranje 10

3.3 Vektorske vložitve 10

3.4 Razpoznavanje anonimiziranih enot 11

4 Ovrednotenje **13**

4.1 Metrike uspešnosti 13

4.2 Rezultati 14

4.3 Čas izvajanja 19

5 Zaključek **21**

Literatura **23**

Povzetek

Naslov: Anonimizacija sodnih odločb z metodami strojnega učenja

Avtor: Gašper Hüll

Anonimizacija sodnih odločb služi zakrivanju in zaščiti podatkov posameznika v primeru, da bi mu njihovo razkritje lahko škodovalo. V skladu z zakonodajo se morajo podatki, preko katerih lahko enolično določimo posameznika, anonimizirati.

Sodne odločbe so pretežno sestavljene le iz prostega besedila. Razpoznavanje entitet v njih zato zahteva razumevanje jezika in vsebine besedila, pomemben pa je tudi kontekst, v katerem so posamezne besede uporabljene. Anonimizacija sodnih odločb je zaradi tega težavna. V delu se osredotočam prav na razpoznavanje entitet, ki so potrebne anonimizacije.

Podatke sem pridobil iz portala sodne prakse IUS-INFO, za njihovo obdelavo pa sem uporabil globoko nevronsko mrežo izdelano po zgledu modela BERT. Besede sem glede na njihovo vektorsko vložitev klasificiral kot "anonimiziraj"oziroma "ne anonimiziraj".

Obstoječi sistemi anonimizacije za predstavitev besed uporabljajo ročno pripravljene vektorje značilk. V delu sem pokazal, da je anonimizacija uspešnejša z uporabo vektorskih vložitev modela BERT, saj je bila uspešna že z uporabo majhne učne množice namenjene razpoznavanju imenskih entitet. Še boljše rezultate sem dosegel z uporabo učne množice zgrajene iz označenih sodnih odločb.

Ključne besede: strojno učenje, anonimizacija, sodna odločba, model BERT.

Abstract

Title: Anonyimization of case law with machine learning

Author: Gašper Hüll

Anonymization of court decisions conceals and protects the information of an individual if its disclosure could be harmful. In accordance to the legislation, all data which enables unique identification of an individual, must be anonymized.

Court decisions are mostly textual. Identifying entities that need anonymization therefore requires an understanding of the language and content of the text, where context in which individual words are used is also important. This makes anonymization of court decisions is therefore difficult. In my thesis I focus on identification of entities that need anonymization.

I obtained the data from the IUS-INFO case-law portal and used a deep neural network based on the BERT model to process it. I classified words as "anonymize" or "do not anonymize".

Existing anonymization systems use manually extracted features. I show that anonymization is more successful using the vector inputs of the BERT model, which were successful using only of a small learning set designed to identify named entities. Anonymization was even better using the learning set built from annotated court decisions.

Keywords: machine learning, anonymization, case law, BERT model.

Poglavje 1

Uvod

Z vse obsežnejšimi zahtevami in določili po varovanju osebnih podatkov postaja njihova uporaba vse strožja. Posebej velik problem nastane, ko se osebni podatki pojavljajo v prostem besedilu, za katerega bi si želeli, da je dostopno širši publiki ali uporabljeno v razvoju rešitev, ki pri svojem delovanju take podatke uporabljajo. Pri tem moramo skrbno paziti, da vse podatke pravnih in fizičnih oseb, ki bi z razkritjem udeležencu povzročili škodo, anonimiziramo.

Anonimizacija poleg pravilne obdelave osebnih podatkov omogoča tudi nezmožnost identifikacije (deidentifikacije) posameznika, na katerega se podatki nanašajo. Po zaključeni anonimizaciji (po določilih Splošne uredbe EU, člen 26 GDPR) anonimizirani podatki niso več osebni podatki, zato se lahko uporabljajo v razvoju, javnih objavah in drugje.

Najbolj razširjeno področje strojne anonimizacije so medicinske kartoteke pacientov, ki se morajo po določilih GDPR (General Data Protection Regulation) ali ameriškega HIPAA (Health Insurance Portability and Accountability Act) pred kakršnokoli nadaljnjo uporabo anonimizirati in deidentifcirati tako, kot je določeno v omenjenih predpisih.

Primer prostih besedil so sodne odločbe, ki morajo biti pred objavo ali vnosom v bazo sodne prakse temeljito pregledane in anonimizirane. Anonimizacija se izvaja pri imenih pravnih in fizičnih oseb, prav tako pa tudi

v povezavi z nepremičninami (naslovi, parcele itd.), drugimi osebnimi podatki ali podatki, preko katerih bi se lahko ugotovila identiteta vpletenih. V določenih primerih morajo biti anonimizirane tudi poslovne skrivnosti.

Naloga obravnava problem zamudne ročne anonimizacije dokumentov, pri čemer je glavni cilj naloge priprava postopka, ki bo s pomočjo strojnega učenja anonimiziral osebne podatke vpletenih v prostem besedilu. Pri tem si bom pomagal z zadnjimi trendi na področju obravnave naravnega jezika in sicer z vektorskimi vložitvami besed. Pridobil jih bom s pomočjo jezikovnega modela BERT[6], ki ga je Google objavil oktobra 2018.

Za vrednotenje uspešnosti postopka, bom predlagano rešitev testiral na sodnih odločbah Vrhovnega sodišča Republike Slovenije. Anonimizirano različico besedila bom predstavil s pomočjo grafičnega vmesnika, preko katerega bo mogoče videti, kako dobro je besedilo anonimizirano.

Delo je sestavljeno iz petih poglavij. Drugo poglavje vsebuje pregled sorodnih del na področju anonimizacije in deidentifikacije prostih besedil, pri katerih se uporablja strojno učenje. V tretjem poglavju je opisan potek anonimizacije in predstavljen sistem, ki ga v delu predlagam. Način vrednotenja in vrednotenje z rezultati je predstavljeno v četrtem razdelku. V zadnjem, petem poglavju, delo povzamem in predstavim prihodnje izboljšave sistema.

Poglavje 2

Ozadje in pregled sorodnih del

Anonimizacija je ena izmed ključnih postopkov v varovanju podatkov, ki za javnost ne smejo biti objavljeni. Leta 1981 je ameriška znanstvenica Dorothy E. Denning v enem izmed svojih člankov zapisala, da popolna zaščita občutljivih podatkov verjetno nikoli ne bo mogoča. To je leta 2009 v svojem delu [13] zaključil tudi Paul Ohm, ki je takrat zapisal, da so podatki lahko zelo uporabni ali pa popolno anonimizirani, nikoli pa oboje.

Znan primer slabe anonimizacije je iz leta 2008, ko sta znanstvenika Arvind Naranayan in Vitaly Shmatikov deanonimizirala Netflixove anonimizirane podatke in tako pokazala, da anonimizacija nikakor ni popolna, tudi v primeru gigantov, kot je Netflix.

Prav zaradi varovanja podatkov, k čemur so zavezani vsi, ki te podatke hranijo, je anonimizacija danes toliko bolj pomembna.

2.1 Ozadje

Glavno področje uporabe anonimizacije je obdelava velike količine medicinskih kartotek, pri čemer morajo biti deidentificirani vsi podatki, iz katerih bi se lahko razkrila identiteta pacienta. Podatki so lahko zakriti ali zamenjani z naključnimi podatki. V večini je zaradi občutljivosti podatkov anonimizacija še vedno prepuščena ročnemu delu, vendar se zaradi napredka na področju

obdelave naravnega jezika (ang. Natural Language Processing ali NLP) pojavlja vedno več programske opreme, ki proces deidentifikacije avtomatizira. K temu prispeva tudi organizacija i2b2 (Informatics for Integrating Biology and the Bedside), ki s svojimi izzivi spodbuja inovacije tudi na tem področju (izziva leta 2006 in 2014)[16][9]. Večina sistemov za anonimizacijo je bilo razvitih predvsem za besedila v angleščini, z nekaj primeri sistemov, ki delujejo večjezično.

Problem deidentifikacije besedila lahko v splošnem prepoznamo kot problem razpoznavanja entitet (ang. Named Entitiy Recognition ali NER). Problema se sistemi lotijo na različne načine. Začetni sistemi so uporabljali velike količine pravil in slovarjev, s pomočjo katerih so iz besedil izločili entitete, ki predstavljajo osebne podatke [15]. Večina sistemov je tak pristop zaradi spreminjanja besedil, ogromnega števila kombinacij in načinov zapisov, ki jih ne moremo predvideti, opustila in se poslužila strojnega učenja.

Globoko učenje je pri obdelavi in razumevanju besedil naredilo ogromen preskok, še posebej z vpeljavo vektorskih vložitev, ki vsako besedo predstavijo s številskim vektorjem, s katerim nevronske mreže nato operirajo [14]. Modeli, ki vložitve računajo vnaprej na nivoju besed (npr. GloVe) so uporabni, vendar se ne ozirajo na to, v kakšnem kontekstu je bila beseda uporabljena. Upoštevanje konteksta v katerem so besede uporabljene v vektorski vložitvi (npr. ELMo, BERT), izboljša rezultate modelov, ki jih uporabljajo. Tudi deidentifikacijski sistemi z globokim učenjem so v zadnjih letih na področju deidentifikacije in NER problemov najuspešnejši in dosegajo najboljše rezultate na i2b2 podatkih [5].

2.2 Tacita

Sistem za anonimizacijo slovenskih sodb že obstaja. Leta 2017 so na Institutu Jožef Štefan (naprej IJS) razvili prvo orodje za pomoč pri anonimizaciji sodb imenovano Tacita, ki je za napovedovanje uporabljalo metodologije strojnega učenja. Orodje so predstavili leta 2018 [1].

Učna množica, na kateri so zgradili svoj sistem je bila sestavljena iz 2480 sodb Vrhovnega in Višjega sodišča v Ljubljani, ki so jih pred uporabo obdelali in poiskali dele besedila, ki so bili anonimizirani. Uporabili so pristop, kjer so iz oblike in konteksta besede v besedilu napovedali, ali jo je potrebno anonimizirati ali ne.

Za predstavitev besede so besede predstavili z značilkami (ang. features). Značilke so definirali sami in jih razdelili v tri sklope: osnovne, kontekstne in razredne značilke. Za primer izbire značilk so podali vprašanja, na katere so značilke odgovarjale, npr. "Je beseda v slovarju imen ali priimkov?, Je beseda v slovarju krajev?, Kakšna je oblika besede?".

Kontekst besede so določili s predstavitvijo besedila v načinu vreče besed (ang. bag of words), pri čemer so kot kontekst besede upoštevali okolico štirih besed pred dano besedo in dveh za njo. Prav tako so med značilke dodali klasifikacijo prejšnjih treh besed. Z uporabo vektorja značilk in logistične regresije (ang. logistic regression) so klasificirali besedo v dva razreda (anonimiziraj, ne anonimiziraj). Z naučenim modelom so pri 5-kratnem prečnem preverjanju dosegli preciznost 72,8%, in pri tem uspešno anonimizirali 91,8% besed (priklic, ang. recall) in bili s tem neuspešni v 8,2%.

V svojem delu sem uporabil podoben pristop, kot ga ima orodje Tacita. Razlikujeta se v gradnji in uporabi značilk, saj jih v mojem delu ne bom določil sam, temveč bom za vektor značilk uporabil vektorske vložitve iz jezikovnega modela. V delu nameravam z manjšo učno množico izboljšati uspešnost orodja Tacita.

2.3 Sorodna dela

V nadaljevanju opišem nekaj avtomatiziranih sistemov za anonimizacijo. Prvi avtomatski sistem za deidentifikacijo, Scrum, je bil razvit leta 1996. Deloval je na podlagi pravil in slovarja, predstavil pa ga je Sweeny [15]. Sledil mu je zmagovalec i2b2 izziva iz leta 2006, MITRE [4], ki uporablja strojno učenje z dvema NER modeloma. Prvi model uporablja markovska pogojna

naključna polja (ang. Conditional Random Fileds ali CRF), drugi pa markovski skriti model (ang. Hidden Markov Model). S pomočjo CRF so Gardner et al. razvil HIDE, ki prav tako uporablja NER model. Uporaba globokega učenja in izboljšave, ki jih doprinese, so predstavili v delu [7].

Portugalski raziskovalci so v svojem delu [10] predstavili sistem, ki je podoben sistemu v mojem delu. Pri tem za predstavitev besed niso uporabili vložitev besed temveč morfološke značilke, ki so jih pridobili s pomočjo STRING orodja [11]. V delu uporabljam vektorske vložitve jezikovnega modela BERT, ki ne upoštevajo le morfologije besed temveč tudi kontekst, v katerem so uporabljene. V objavi razvijajo sistem za avtomatsko anonimizacijo portugalskih prostih besedil, moje delo pa obravnava anonimizacijo slovenskih besedil. Razviti sistem se od portugalskega razlikuje tudi po razpoznavanju besed, saj njihov model razpoznava imenske entitete (ang. Named Entity) in nato besede anonimizira na podlagi tega, v katero skupino entitet spadajo. Del sistema, predstavljenega v mojem delu, ki je zadolžen za klasifikacijo besed, pa bo identificiral le ali je besedo potrebno zakriti ali ne in ne v katero kategorijo entitet spada.

Poglavje 3

Metodologija

Sistem v delu je sestavljen iz treh komponent: predprocesiranje besedila, priprava vektorskih vložitev in razpoznavanje besed potrebnih anonimizacije. Sistem sprejme prosto besedilo, ga obdela in nato vrne anonimizirano besedilo. Dodatno lahko sistem vrne tudi tabelo vseh najdenih anonimiziranih entitet z njihovimi pozicijami v besedilu.

V delu uporabljam prosto dostopen jezikovni model BERT. Vnaprej naučen ni le model za angleščino temveč tudi večjezični model za okoli 100 jezikov, med katerimi je tudi slovenščina. V delu uporabljam vnaprej naučen večjezični model, dodatno priučen na korpusu sodnih besedil v slovenščini.

3.1 Podatkovne množice

Moj sistem za svoje delovanje potrebuje vnaprej naučene modele. Modele je potrebno vnaprej naučiti ali doučiti na primernih besedilih. Uporabil sem tri podatkovne množice oz. korpuse besedil.

Začetni večjezični model BERT [3] je bil vnaprej naučen na besedilih Wikipedije v 104 jezikih. Za dodatno priučenje modela sem uporabil besedila sodnih odločb, ki sem jih pridobil preko portala IUS-INFO [2] v velikosti 270 milijonov besed. S tem jezikovnim modelom sem pridobival vektorske vložitve.

Enota za razpoznavanje anonimiziranih entitet je v svojem jedru NER sistem, ki je sestavljen iz nevronske arhitekture, uteži nevronske mreže in parametrov učenja. Za primerjavo pomembnosti izbire pravih podatkov za nalogo anonimizacije sem pripravil dva modela, opisana v naslednjem razdelku.

3.1.1 Učne množice

Prvi model sem naučili na podatkih iz podatkovne množice ssj500k, ki vsebuje 500 tisoč slovenskih besed, označenih na nivoju sestavkov, stavkov in besed. Od teh jih je približno polovica označenih tudi z imenskimi entitetami kategorij oseba, organizacija, lokacija in drugo. S pomočjo teh imenskih entitet sem model naučil, katere besede so potrebne anonimizacije in katere ne.

Množica skupno zajema 504.950 besed, od katerih je 11.137 označenih kot ena izmed imenskih entitet. Označeni so štirje tipi imenskih entitet: ime osebe (4.436 ali 40%), organizacija (2.633 ali 24%), lokacija (2.445 ali 22%) ali drugo (1.623 ali 16%). V tip drugo so šteta imena ali modeli predmetov, dogodki, filmi in imena različnih ekip (npr. Nokia 7270, Life of Brian...).

Drugi model sem naučil na sodnih odločbah Vrhovnega sodišča Republike Slovenije, ki sem jih pridobil iz portala IUS-INFO [2]. Vse sodne odločbe so pred objavo na portalu anonimizirane na Vrhovnem sodišču, kjer za to skrbi Evidenčni oddelek. Anonimizirano prosto besedilo, zato moral najprej prečistiti in ročno označiti. Vsaka oznaka vsebuje tip anonimizirane besede, s pomočjo katerega sem opazoval, kako dobro sistem anonimizira posamezne tipe oznak. Omejil sem se na 6 tipov entitet: osebe, podjetja, naslove, občine, katastrske občine in vse ostale vrednosti, ki so se posamezno pojavile manj kot 5 krat. Med osebe spadajo vse omembe oseb z imenom, priimkom, vzdevkom ali drugim poimenovanjem. Med občine so spadale omembe krajev in občin, med naslove pa točnejše omembe lokacije.

Celotna učna množica drugega modela je sestavljena iz 100 dokumentov v času od osamosvojitve Slovenije leta 1991 do 2019 in zajema 197.876 besed,

od katerih mora biti anonimiziranih 5.310. V odločbah so se največkrat pojavljala imena oseb (3.684 ali 69,4%), sledijo imena podjetij (1.143 ali 21,5%), imena krajev in naslovi (431 ali 8,1%), ostalih označenih entitet, kot so številke parcel, opravilne številke odločb, pa je manj kot 1%.

Med obema učnima množicama je nekaj razlik. Množica ssj500k je večja od množice pripravljene iz sodnih odločb, vendar ima označenih veliko več besed, ki niso potrebne anonimizacije. Množica ne vsebuje znanja o anonimizaciji, temveč le o imenskih entitetah. Znanje o anonimizaciji odločb vsebuje množica, ki je bila pripravljena na anonimiziranih odločbah, saj so označene le tiste besede, ki so bile tudi zares anonimizirane. Model pripravljen iz druge množice bo zato pri enakih parametrih testiranja boljši pri anonimizaciji kot model pripravljen iz imenskih entitet v ssj500k.

Za boljšo razumljivost je v nadaljevanju prvi model imenovan NER model, drugi pa ANON model (zgrajen iz anonimiziranih sodnih odločb).

3.1.2 Testna množica

Za testiranje modela za anonimizacijo sem potreboval sodne odločbe, ki sem jih, tako kot učne podatke za pripravo jezikovnega modela in drugega modela za razpoznavo anonimiziranih besed, pridobil iz portala IUS-INFO [2].

Testno množico sem pripravil iz 100 sodnih odločb Vrhovnega sodišča Republike Slovenije. Vsako odločbo sem, kot pri učni množici, ročno prečistil in anonimizirane vrednosti opremil z dodatnimi oznakami. Vseh 100 dokumentov sem pretvoril v format tsv (ang. Tab Separated Values), kjer je prvi stolpec beseda, drugi pa oznaka "OTHER"ali "ANONYMIZE". Učna in testna množica nimata skupnih dokumentov.

Celotna testna množica je sestavljena iz 100 dokumentov v času od osamosvojitve Slovenije leta 1991 do 2019. Ti dokumenti skupaj zajemajo 149.924 besed, od katerih mora biti 5.819 anonimiziranih.

V odločbah testne množice so največkrat anonimizirana imena oseb (3.930 ali 67,54%), sledijo imena podjetij (1.347 ali 23,15%), imena krajev in naslovi (532 ali 9,1%), v nekaj primerih pa so bile anonimizirane tudi druge vrednosti,

kot so številke parcel in opravilne številke.

3.2 Predprocesiranje

Sistem na vhodu prejme prosto, neočiščeno besedilo, zato ga mora pred nadaljevanjem očistiti in razdeliti na stavke. Nevronska mreža v naslednjem koraku sistema na vhodu sprejme omejeno dolžino besedila. Zato mora ta del poskrbeti tudi za deljenje stavkov na primerno dolžino, da ne pride do izpuščanja besed.

3.3 Vektorske vložitve

Najbolj znane nevronske vektorske vložitve so zgrajene z algoritmom word2vec [12]. Zaradi svoje uporabnosti so se vložitve hitro razširile na skoraj celotno področje obdelave naravnega jezika. Z vložitvami besede predstavimo kot vektorje števil, na katerih lahko nato izvajamo matematične operacije. Najbolj znan primer je povezava med besedama kralj in kraljica:

Naj bo ϕ vektorska preslikava $W \to R^n$, kjer W predstavlja besedo in R^n vektorski prostor dimenzije n, takrat velja:

$$\phi(\text{"}kralj\text{"}) - \phi(\text{"}moški\text{"}) + \phi(\text{"}ženska\text{"}) \approx \phi(\text{"}kraljica\text{"})$$

Vektorske vložitve word2vec upoštevajo le statični semantični pomen besede, danes pa jezikovni modeli znajo upoštevati tudi kontekst, v katerem je beseda uporabljena. Enake besede, ki so uporabljene v drugačnem kontekstu, niso predstavljene z enakim vektorjem.

V tem delu za pridobivanje vložitev uporabljam jezikovni model BERT [6], ki upošteva kontekst, v katerem je beseda uporabljena. Avtorji to pokažejo s primerom besednih zvez "river bank"in "bank deposit". Beseda "bank"bi imela v obeh primerih enak besedni vektor, če model ne bi upošteval besed levo in desno (tj. konteksta).

Jezikovni model BERT, ki ga uporabljam v delu, je sestavljen iz 12 plasti kodirnikov največje dolžine 512, s skrito velikostjo 768 (ang. hidden size). Vsako polje za svoj žeton na izhodu vrne vektor dolžine enake skriti velikosti kodirnikov, ki ga nato preda v naslednjo plast. To pomeni, da lahko vektorske vložitve za vsak žeton dobimo na izhodu vsakega kodirnika. Pri tem so vložitve na izhodih kasnejših kodirnikov bolj uporabne kot tistih na začetku. V delu uporabljam vektorske vložitve predzadnje plasti.

Del sistema, ki je zadolžen za pridobivanje vektorskih vložitev, prejme seznam stavkov in vrne seznam dvojic, sestavljenih iz besede in njene vektorske vložitve. Jezikovni model BERT besede razdeli na tako imenovane žetone (ang. tokens), s katerimi naprej operira. Žeton je beseda ali del besede, pri čemer model, ko vrne rezultat, nakaže katere besede je razdelil (besedo zaznavati jezikovni model razdeli na žetone za, ##zna, ##vati). Ker pri nadaljnjem procesu ne želimo operirati z žetoni, temveč s prvotnimi besedami iz vhoda, moramo dobljene žetone združiti nazaj. Pri tem njihove vektorske vložitve seštejemo in vsako izmed 768 vrednosti vložitve delimo s številom vseh žetonom v besedi.

3.4 Razpoznavanje anonimiziranih enot

Za nalogo anonimizacije sem naučil preprost klasifikator, ki glede na podano vektorsko vložitev dimenzije 768 napove, ali mora biti beseda, ki jo vložitev predstavlja, zakrita ali ne. Klasifikator je zgrajen iz dveh LSTM (ang. Long short-term memory) plasti, ki jima sledi še plast z aktivacijsko funkcijo ReLU (angl. Rectified Linear Unit) in zadnja klasifikacijska plast z aktivacijsko funkcijo Softmax.

Za vsako učno množico opisano v razdelku 3.1.1 sem naučil svoj klasifikator razpoznavanja anonimiziranih enot. Za vse besede iz učne množice je bila pripravljena vektorska vložitev in primerna oznaka, 0 za razred "ne anonimiziraj"in 1 za razred "anonimiziraj". Klasifikatorja sta za učenje prejela vložitve in njihove oznake. Uporabljen je bil Adam optimizator [8], modela

pa sta bila učena dve epohi, pri čemer sta za to potrebovala približno 20 minut.

Poglavje 4

Ovrednotenje

Testiranje sistema za anonimizacijo je potekalo v dveh korakih. Prvi korak je bil priprava testne množice in dokumentov, ki sem jih v drugem koraku uporabil za testiranje modelov za anonimizacijo. V drugem koraku sem oba modela testiral pri različnih vrednostih meje razpoznavanja entitet potrebnih anonimizacije.

Pri testiranju na obeh modelih sem spreminjal mejo, pri kateri je sistem klasificiral besedo kot anonimizirano. V primeru meje X% je sistem anonimiziral vse besede, za katere je model z X% ali več trdil, da spadajo v skupino anonimiziranih entitet.

4.1 Metrike uspešnosti

Pri testiranju na testni množici sem napovedane primere razdelil v štiri skupine: primere, ki so bili pravilno označeni za anonimizirane in neanonimizirane, imenovane dejansko pozitivni in dejansko negativni primeri (ang. true positive TP in true negative TN) ter napačno pozitivne in napačno negativne primere (ang. false positive FP in false negative FN). Zaradi kritičnosti anonimizacije, so napačno negativni primeri veliko hujša napaka kot napačno pozitivni. Uspešnost sistema sem meril z napako klasifikacije na testni množici:

$$\text{Napaka klasifikacije} = \frac{\text{število napačnih napovedi}}{\text{število vseh napovedi}} = \frac{FP+FN}{TP+TN+FP+FN}$$

in napako pri anonimizaciji:

$$\text{Napaka anonimizacije} = \frac{\text{število besed, ki niso bile anonimizirane, vendar bi morale biti}}{\text{število besed, ki bi morale biti anonimizirane}} = \frac{FN}{TP+FN}$$

Model je uspešnejši od drugega, če ima pri enaki meji manjšo napako. Model je najboljši pri tisti meji, kjer sta napaki klasifikacije in anonimizacije najmanjši.

4.2 Rezultati

Testni sistem anonimizacije sem razdelil v tri procese. Prvi proces skrbi za pridobivanje vektorskih vložitev in je implementiran kot strežnik, ki na zahtevo vrača vektorske vložitve. Implementacija je bila že pripravljena kot bert-as-service [17], potrebne so bile le manjše spremembe in pravilni parametri, da je strežnik deloval s pravilnim prednaučenim modelom in vračal vložitve predzadnje plasti. Drugi proces je skrbel za zahteve po vložitvah in končno napoved. Tako kot prvi je tudi ta implementiran kot strežnik, ki sprejema zahteve iz aplikacij, posreduje zahtevo za pridobitev vložitve in nato glede na vložitev za vsako besedo napove izid. Vse besede skupaj z njihovimi izidi strežnik nato vrne aplikaciji v formatu json. Tretji proces skrbi za pravilno predobdelavo besedila, pošiljanje in prejemanje zahtev ter predstavitev oziroma uporabo rezultata anonimizacije. Proces je lahko pripravljen v katerem koli programskem jeziku ali orodju, s katerim je mogoče pošiljanje zahtev. Vključen je lahko v samostojno aplikacijo ali pa je okrog njega pripravljeno ogrodje, ki opravi nadaljnjo procesiranje rezultatov v obliki računanja statistike, preverjanja pravilnosti glede na testne podatke ali predstavitve rezultatov v uporabniškem vmesniku.

Kritičnost anonimizacije zahteva čim manjšo napako, kjer je napačna neanonimizacija slabša kot napačna anonimizacija. Pričel sem z mejo 10% in jo nato zmanjševal, dokler nista bili napaki klasifikacije in anonimizacije najmanjši.

4.2.1 Anonimizator naučen na ssj500k NER učni množici

Testiranje sem začel z modelom NER. Pri tem sem vseh 100 dokumentov testne množice očistil označb za anonimizacijo in besedila poslal skozi pripravljen sistem, ki je na izhodu ustvaril tsv (ang. Tab Separated Values) datoteko. V prvi stolpec je sistem zapisal besede in v drugi stolpec pa njihovo klasifikacijo. Generirane datoteke je nato primerjal z vnaprej pripravljenimi tsv datotekami s pravilnimi klasifikacijami testne množice.

Meja	TP	TF	FP	FN	Napaka klasifikacije	Napaka anonimizacije
10%	5.000	136.898	7.207	819	5,4%	14,1%
5%	5.211	135.122	8.983	608	6,4%	10,4%
3%	5.340	133.520	10.585	479	7,4%	8,2%
2%	5.426	132.060	12.045	393	8,3%	6,8%
1%	5.516	129.012	15.093	303	10,3%	5,2%
0,1%	5.699	115.312	28.793	120	19,3%	2,1%

Tabela 4.1: Rezultati testiranja NER modela pri različnih mejah.

Rezultati testiranja razvidni so v tabeli 4.1. Že pri 10% je bila napaka klasifikacije le 5,4% in napaka anonimizacije 14,1%, kar pomeni, da je bila anonimizacija v večini primerov uspešna in bi bila glede na delež uspešnosti kot pomoč ročni anonimizaciji že lahko uporabna. S testiranjem sem nadaljeval in z vsako iteracijo in nižanjem meje dobival manjšo napako anonimizacije in večjo napako klasifikacije, saj je sistem anonimiziral vse več besed, ki niso bile potrebne anonimizacije. Napaka klasifikacije in napaka anonimizacije sta bili obratno sorazmerni. Najboljše rezultate sem dobil med mejama 3% in 2%, kjer sta bili opazovani napaki zelo blizu. Kljub majhni učni množici so bili rezultati anonimizacije zelo dobri.

Uspešnost anonimiziranja različnih entitet pri mejah 3% in 2% je prikazana v tabeli 4.2. Iz nje je razvidno, da je sistem najbolje anonimiziral entitete, ki predstavljajo osebe, za tem pa občine in naslove. Uspešnost je mogoče pripisati dobri pokritosti teh entitet v učni množici. Slabše so bile

Meja		3%		2%	
Tip entitete	Skupno število	Napake	Uspešnost	Napake	Uspešnost
OSEBA	3.930	126	97%	98	98%
NASLOV	277	34	88%	29	90%
K.O.	167	39	77%	33	80%
PODJETJE	1.347	264	80%	217	84%
OBČINA	88	10	89%	10	89%
MISC	10	6	40%	6	40%

Tabela 4.2: Rezultati uspešnosti anonimizacije pri NER modelu na različnih tipih entitet.

razpoznane katastrske občine, za katere sem pričakoval, da bodo obravnavane podobno kot občine, vendar zaradi pomanjkanja primerov v učni množici in zaradi kontekstne razlike med omembo občine in katastrske občine temu ni bilo tako. Veliko slabša je uspešnost pri drugih ali "MISC"vrednostih, saj so v učni množici slabše zastopane, prav tako pa se kontekstno razlikujejo od "MISC"vrednosti v testni množici.

Zanimivost pri NER modelu je, da kot anonimizirano vrednost označuje tudi kratice in naslove zakonov. Razlog za to je verjetno v označbi učne množice ssj500k, ki ima kot imenske entitet označene tudi naslove zakonov, ti primeri so v množici označeni kot "MISC".

Glede ne rezultate lahko rečem, da je model delno uspešen. Zakrije večino entitet, ki morajo biti anonimizirane, vendar pri temu po nepotrebnem anonimzira preveč besed, ki jih ne bi smel. Trenutni sistem ne more biti samostojni avtomatski anonimizator, saj je potrebno vsak anonimiziran dokument še ročno pregledati in popraviti.

Vseh anonimiziranih entitet je v 100 odločbah 5.819, kar je v povprečju 58 besed na dokument, ki jih je potrebno, brez dodatnega orodja, ročno anonimizirati. V primeru 10% meje je model naredil 8.026 napak, od katerih je bilo 7.207 preveč in 819 premalo anonimiziranih besed. To pomeni, da

Meja	TP	TF	FP	FN	Napaka klasifikacije	Napaka anonimizacije
50%	3.456	142.898	1.207	2.363	2,4%	40,6%

Tabela 4.3: Rezultati anonimizacije z NER modelom pri meji 50%.

je v povprečju na dokument 72 preveč in 8 premalo anonimiziranih besed. Skupno je to 80 besed, ki jih potrebno ročno popraviti, kar je več kot 58 besed na dokument, ki bi jih morali anonimizirati brez avtomatskega sistema.

Rezultati anonimizacije pri meji 50% so vidni v tabeli 4.3. Pri manjši meji so rezultati anonimizacije sicer slabši, vendar je skupno manj napak. V tem primeru bi morali ročno v povprečju popraviti le 35 besed na dokument, kar je manj kot začetnih 58 besed ročne anonimizacije. Vendar nam tak sistem ne pomaga, saj bi želeli, da že sam pravilno anonimizira več besed in naredi manj nepotrebnih anonimizacij, ki jih je treba popravljati ročno.

4.2.2 Anonimizator naučen na učni množici označenih sodnih odločb

Testiranje sem ponovil še z ANON klasifikacijskim modelom na enakih mejah kot z NER modelom. Rezultati so predstavljeni v tabeli 4.4.

Meja	TP	TF	FP	FN	Napaka klasifikacije	Napaka anonimizacije
10%	5.689	143.518	587	130	0,48%	2,2%
5%	5.707	143.460	645	112	0,50%	1,9%
3%	5.715	143.416	689	104	0,53%	1,8%
2%	5.720	143.366	739	99	0,56%	1,7%
1%	5.727	143.289	816	92	0,61%	1,6%
0,1%	5.753	142.837	1.268	66	0,89%	1,1%

Tabela 4.4: Rezultati testiranja ANON modela za anonimizacijo pri različnih mejah.

Razvidne so velike izboljšave ANON modela v primerjavi z NER modelom, saj so tako napake klasifikacije kot napake anonimizacije precej manjše. Že pri meji 10% sta tako klasifikacijska napaka kot napaka anonimizacije manjši kot pri isti meji pri NER modelu. Enak trend se nadaljuje tudi pri nadaljnjih vrednostih. Iz tabele 4.4 je razvidno, da je model uspešen pri vseh vrednostih mej modela, pri katerih sem ga testiral. ANON model pravilno anonimizira skoraj vse primere in napačno anonimizira 10-krat manj besed kot NER model.

Meja		3%		2%	
Tip entitete	Skupno število	Napake	Uspešnost	Napake	Uspešnost
OSEBA	3.930	36	99,1%	33	99,2%
NASLOV	277	34	87,7%	33	88,1%
K.O.	167	3	98,2%	3	98,2%
PODJETJE	1.347	20	98,5%	19	98,6%
OBČINA	88	7	92,0%	7	92,0%
MISC	10	4	60,0%	4	60,0%

Tabela 4.5: Rezultati uspešnosti anonimizacije pri ANON modelu na različnih tipih entitet.

Uspešnost anonimiziranja različnih entitet pri mejah 3% in 2% je prikazana v tabeli 4.5. Tudi anonimizacija po entitetah je v vseh primerih boljša kot v NER modelu. Razen za tip entitete naslov in drugo sta imela modela pri obeh mejah uspešnost nad 90%, pri čemer je model z mejo 3% pravilno anonimiziral 87,7% naslovov. Najslabše so bile anonimizirane entitete tipa drugo, kjer je model pravilno označil le 2 entiteti več kot NER model. Razpoznavo teh entitet bi bilo potrebno pri ANON modelu izboljšati.

Glede na rezultate lahko ugotovim, da je model ANON, naučen že na samo 100 odločbah, zelo uspešen. Zakrije skoraj vse potrebne entitete in manj nepotrebnih besed kot NER model. Sistem ne more biti avtomatski anonimizator, saj še vedno dela napake, je pa veliko boljši od NER modela.

V testni množici je 5.819 anonimiziranih entitet, kar v 100 dokumentih pomeni povprečno 58 entitet na dokument. V primeru 10% meje na ANON modelu je sistem naredil 130 napak anonimizacije in 587 napak, kjer sistem entitete ni anonimiziral. Skupno je to 717 napak, kar v povprečju pomeni 7 napak na dokument, ki jih je potrebno ročno popraviti, kar je manj od 58, ki bi jih bilo potrebno ustvariti brez sistema za anonimizacijo. Prav tako je 7 ročnih popravkov manj od 80 napak, ki bi jih bilo potrebno popraviti pri uporabi NER modela.

Predlagani sistem z uporabo modela ANON je primeren za avtomatsko anonimizacije v primeru, da se vsak anonimiziran dokumnet po obdelavi pregleda in dopolni ter popravi napake. Predlagani sistem bi bil pri anonimizaciji v veliko pomoč, saj anonimizira večino potrebnih entitet. Sistem bi bilo mogoče še bistveno izboljšati z večjo učno množico.

4.3 Čas izvajanja

Zaradi obdelave velike števila dokumentov želimo, da je sistem avtomatske anonimizacije hiter. Pri tem sem meril hitrost obdelave enega dokumenta in čas celotnega testiranja pri poljubni meji.

Testna množica je sestavljena iz 100 dokumentov, obdelava enega je v povprečju trajala do 5 sekund. Čas obdelave je bil odvisen od dolžine besedila v dokumentu. Sistem za anonimizacijo Tacita naj bi za dokument v povprečju potreboval le sekundo, kar pomeni, da je sistem predstavljen v mojem delu počasnejši.

Učenje klasifikacijskih modelov je bilo kratko, za dve epohi sta oba modela potrebovala približno 20 minut. Čas dodatnega učenja je odvisen od velikosti učne množice.

Poglavje 5

Zaključek

V delu sem predstavil sistem avtomatske anonimizacije prostega besedila sodnih odločb. Sistem sem testiral na dveh modelih, prvega na učni množici ssj500k, drugega pa na množici ročno označenih sodnih odločb. Oba sem testiral na istih testnih dokumentih in primerjal njuno uspešnost. Modularna zasnova predlaganega modela omogoča zamenjavo delov sistema, zaradi česar je lahko sistem kadarkoli nadgrajen ali uporabljen tudi na dokumentih v drugih jezikih.

Sistem je bil pričakovano uspešnejši pri uporabi modela, naučenega na označenih sodnih odločbah Vrhovnega sodišča Republike Slovenije. Izbira te učne množice se izkaže za mnogo boljšo že pri majhnem številu učnih podatkov. Na besedilih sodnih odločb je bil naučen tudi sistem za anonimizacijo, Tacita. Rezultati mojega sistema, so kljub manjši učni množici presegli pričakovanja. Sistem je uspešno anonimiziral besede vseh tipov, ki jih je anonimiziral Evidenčni oddelek Vrhovnega sodišča, pri tem pa je bilo število napačno anonimiziranih in ne-anonimiziranih besed zelo malo. Predstavljeni sistem je bil uspešnejši od Tacite, vendar zaradi različnih učnih in testnih množic ne morem reči, da je zares boljši. Sklepam pa, da bi bil moj sistem, naučen na istih učnih množicah lahko boljši od sistema Tacita.

Predstavljeni sistem bi se v prihodnosti dalo izboljšati. Glavna točka izboljšav bi bila večja učna množica, ki bi jo pripravil s pomočjo Evidenčnega

oddelka Vrhovnega sodišča na enak način, kot so jo pripravili za sistem Tacita. Popolne avtomatske anonimizacije ne moremo pričakovati, bi pa z boljšo učno množico izboljšali klasifikacijski model in tako bolj precizno določali, katere besede morajo biti anonimizirane in katere ne.

Vektorskim vložitvam bi lahko dodali dodatne, lastne značilke, s katerimi bi izboljšali predstavitev besed in tako izboljšal njihovo klasifikacijo. Za izbiro značilk bi lahko uporabili osnovne značilke Tacite, ki v predstavitev besede vpeljejo znanje iz slovarjev imen, priimkov in krajev.

Čas obdelave enega dokumenta in učenja modelov bi lahko skrajšali z uporabo namenskih grafičnih procesnih enot in primernih knjižnic. Učenje in klasifikacija bi potekala mnogo hitreje kot trenutno na centralno procesni enoti.

S spremembami in izboljšavami bi predstavljeni sistem presegel sistem Tacita, tako v uspešnosti kot v hitrosti obdelave. Celoten sistem bi se na novih odločbah še naprej učil in izboljševal.

Literatura

[1] Anonimizator Tacita. http://www.lr-coordination.eu/sites/default/files/Slovenia/ELRCworkshop2018_kovacic_anonimizator_tacita_01.pdf. Accessed: 2020-01-19.

[2] Ius-info. https://www.iusinfo.si. Accessed: 2020-01-19.

[3] Multilingual BERT model, cased. https://storage.googleapis.com/bert_models/2018_11_23/multi_cased_L-12_H-768_A-12.zip. Accessed: 2020-01-19.

[4] John Aberdeen, John Burger, David Day, Lynette Hirschman, Patricia Robinson, and Marc Vilain. MITRE: description of the alembic system used for MUC-6. In *Proceedings of the 6th conference on Message understanding*, pages 141–155. Association for Computational Linguistics, 1995.

[5] Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606, 2017.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, 2019.

[7] James Gardner and Li Xiong. HIDE: an integrated system for health information de-identification. In *2008 21st IEEE International Symposium on Computer-Based Medical Systems*, pages 254–259. IEEE, 2008.

[8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015*, 2015.

[9] Zengjian Liu, Yangxin Chen, Buzhou Tang, Xiaolong Wang, Qingcai Chen, Haodi Li, Jingfeng Wang, Qiwen Deng, and Suisong Zhu. Automatic de-identification of electronic medical records using token-level and character-level conditional random fields. *Journal of biomedical informatics*, 58:S47–S52, 2015.

[10] Nuno Mamede, Jorge Baptista, and Francisco Dias. Automated anonymization of text documents. In *2016 IEEE Congress on Evolutionary Computation (CEC)*, pages 1287–1294. IEEE, 2016.

[11] Nuno Mamede, Jorge Baptista, Cláudio Diniz, and Vera Cabarrão. STRING: An hybrid statistical and rule-based natural language processing chain for Portuguese. 2012.

[12] Tomas Mikolov, Kai Chen, Gregory S Corrado, and Jeffrey A Dean. Computing numeric representations of words in a high-dimensional space, 2019. US Patent 10,241,997.

[13] Paul Ohm. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA l. Rev.*, 57:1701, 2009.

[14] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[15] Latanya Sweeney. Replacing personally-identifying information in medical records, the Scrub system. In *Proceedings of the AMIA annual fall symposium*, page 333. American Medical Informatics Association, 1996.

[16] Ozlem Uzuner, Peter Szolovits, and Isaac Kohane. i2b2 workshop on natural language processing challenges for clinical records. *Proc AMIA Symp*, 01 2006.

[17] Han Xiao. BERT-as-service. `https://github.com/hanxiao/bert-as-service`, 2018.