Prepoznavanje imena na slikama lekarskih izveštaja na srpskom jeziku u cilju zaštite ličnih podataka

Aldina Avdić, Ulfeta Marovac

***Apstrakt*— Savremeni način života neizostavno uključuje upotrebu računara i mobilnih telefona u svim svojim segmentima, pa i kada je u pitanju zdravstvo. Pored elektronskog zdravstva koje se sve više razvija, pogotovo od kada se svet suočio sa pandemijom korona virusa, sve više ljudi traži i savete o zdravlju na društvenim mrežama. Tom prilikom dodaju slike koje sadrže njihove zdravstvene rezultate, ne mareći o tome da na taj način ostavljaju i svoje lične podatke. U ovom radu data je metoda za prepoznavanje imena na slikama medicinskih izveštajima napisanih na srpskom jeziku u cilju njihove de-identifikacije. Ova metoda bazirana je na optičkom prepoznavanju karaktera, metodama obrade prirodnog jezika i pravilima i ima široku upotrebu, jer je de-identifikacija elektronskih medicinskih izveštaja neophodan korak za njihovu bilo kakvu dalju analizu.**

***Ključne reči*— de-identifikacija, elektronski medicinski izveštaji, prepoznavanje imenovanih entiteta, obrada prirodnog jezika, optičko prepoznavanje karaktera, zaštita privatnosti, srpski jezik.**

I. Uvod

Informacione i komunikacione tehnologije već duže vreme nalaze svoju primenu u zdravstvenom sistemu. E-zdravstvo (engl. *e-Health*) je termin nastao krajem dvadesetog veka i zasniva se na efikasnijem i kvalitetnijem pružanju zdravstvenih usluga uz pomoć savremenih tehnologija, mogućnostima za obezbeđivanje mobilnosti lekara i pacijenata uz integraciju sa postojećim sistemima [1].

Savremene tehnologije u zdravstvu omogućavaju kompletno elektronsko vođenje zdravstvene dokumentacije u svim segmentima rada (elektronski karton pacijenta), telemedicinu, telekonsultacije, upravljanje znanjem o zdravlju i mobilno i sveprisutno zdravstvo. Na ovaj način omogućava se lakše i uspešnije lečenje pacijenata i smanjenje administrativnih ograničenja prilikom pružanja medicinskih usluga.

Prednosti e-zdravstva su elektronsko praćenje i beleženje zdravstvenog stanja pacijenata, bolja dijagnostika, pristup medicinskim podacima bilo gde i bilo kada, kao i brz prenos informacija korisnicima putem telemedicine i Internet servisa.

Aldina Avdić – Departman za tehničko-tehnološke nauke, Državni Univerzitet u Novom Pazaru, Vuka Karadžića 9, 36300 Novi Pazar, Srbija (e-mail: apljaskovic@np.ac.rs).

Ulfeta Marovac – Departman za tehničko-tehnološke nauke, Državni Univerzitet u Novom Pazaru, Vuka Karadžića 9, 36300 Novi Pazar, Srbija (e-mail: umarovac@ np.ac.rs).

Elektronski medicinski zapis (engl. *EMR* ili *EHR* - *electronic medical (health) report*) je izveštaj pacijenta koji čuva podatke o zdravstvenom stanju, dijagnozama i terapijama. Oni se kreiraju pomoću medicinskih informacionih sistema i sadrže privatne podatke o pacijentu (ime, prezime, lični broj, datum rođenja, broj kartice, broj osiguranja, adresu itd.), beleške lekara, dijagnoze, laboratorijske izveštaje, terapije itd [2].

Motivacija za ovaj rad je pripremanje elektronskih izveštaja za dalju analizu korišćenjem metoda veštačke inteligencije, mašinskog učenja i metoda obrade prirodnog jezika. U kakvom god da su obliku elektronski medicinski izveštaji (tekst ili slika) prvi korak ka njihovoj etičkoj obradi jeste uklanjanje ličnih podataka (ovo je uređeno regulativom – engl. *General Data Protection Regulation - GDPR*). Za potrebe rada prikupljen je skup podataka, i kreirana i primenjena metoda za prepoznavanje imena i prezimena na srpskom jeziku sa slika medicinskih izveštaja bazirana na optičkom prepoznavanju slova (engl. *Optical Caracter Recognition - OCR*) i metodama obrade prirodnog jezika (engl. *Natural Language Processing - NLP*) i prepoznavanja imenovanih entiteta (engl. *Named Entity Recognition – NER*). Stoga će nadalje ovi pojmovi biti detaljno objašnjeni.

Uredba Evropske unije o zaštiti podataka o ličnosti GDPR definiše propise o zaštiti podataka ličnosti u Evropskoj uniji, ali se ona odnosi i na kompanije sa sedištem u zemljama izvan Evropske unije, ukoliko one obrađuju podatke o ličnosti rezidenata Evropske unije. Po članu 3 GDPR regulative potrebno je da se uskladi zaštita osnovnih prava fizičkih lica u vezi sa aktivnostima obrade ličnih podataka. U Srbiji je na snazi Zakon o zaštiti podataka o ličnosti iz 2018. godine koji je usklađen sa GDPR [3]. Ona obezbeđuje pojedincima veću kontrolu nad ličnim podacima i obavezuje kompanije koje prikupljaju i analiziraju lične podatke da promene svoj model poslovanja u skladu s njom. Pod ličnim podacima podrazumeva se bilo koja kombinacija ličnih činjenica koja tačno određuje jednog pojedinca, to su, između ostalog, ime i prezime, JMBG, podaci o lokaciji, fizički, fiziološki, genetski, mentalni, ekonomski, socijalni, kulturni ili bilo koji drugi faktori.

Osnovni zadatak OCR softvera je da digitalne slike, na kojima se nalaze skenirani ili fotografisani tekstovi sa štampača, pisaćih mašina, knjiga, novina, časopisa ili poslovne dokumentacije, pretvori u promenljivi digitalni tekstualni oblik, tako što će sa rasterskih slika, prepoznati

 ISBN 978-86-7466-930-3

slova, reči i čitave tekstove.[4].

Procesiranje prirodnog jezika je grana veštačke inteligencije koja omogućava sistemu da razume značenje podataka na ljudskom jeziku. Njen krajnji cilj je čitanje, dešifrovanje, razumevanje i nalaženje smislenosti u prirodnom jeziku. Većina NLP tehnika oslanja se na mašinsko učenje da bi se zaključilo o značenju podataka na prirodnim jezicima. Prepoznavanje imenovanih entiteta je grana NLP-a kojom se u tekstu označavaju sintagme u nestrukturiranom tekstu u unapred definisane kategorije kao što su imena osoba, organizacije, lokacije, medicinski kodovi, vremenski izrazi itd [5] [6].

Ovaj rad je organizovan na sledeći način. U drugom poglavlju dat je pregled radova iz oblasti. U trećem poglavlju opisani su korišćeni podaci i metoda za prepoznavanje imena na slikama medicinskih izveštajima napisanih na srpskom jeziku u cilju njihove de-identifikacije. Zatim sledi opis korišćenih tehnologija, dobijeni rezultati i njihova diskusija. Na kraju je dat zaključak i pravci daljeg istraživanja.

II. Stanje u oblasti

Povezana istraživanja na ovu temu opisuju metode za normalizaciju i izdvajanje znanja iz medicinske evidencije koje nisu direktno povezane sa primenom na srpskom jeziku. Većina istraživanja odnosi se na korpuse engleskog govornog područja i leksičke resurse koji su javno dostupni. Za bugarski jezik postoje rezultati izdvajanja informacija iz velikog korpusa nestrukturiranih podataka i njihovog pribavljanja u strukturiranom obliku [7].

Normalizacija je prvi korak koji se koristi u klasifikaciji i obeležavanju medicinskih termina [8]. Obrada medicinskog izveštaja sastoji se od nekoliko koraka kao što su prečišćavanje podataka, integracija, transformacija, redukcija i na kraju zaštita podataka. Izvlačenjem kliničkih odnosa iz medicinskih izveštaja mogu se identifikovati odnosi između referenci na lekove i njihovih atributa. Pregled medicinskih informacionih sistema pokazuje da 60% komercijalnih sistema koristi metode zasnovane na pravilima, dok se u naučnim istraživanjima više koriste metode mašinskog učenja [9]. Najpopularniji sistemi za obeležavanje medicinskog teksta su CTAKES i CLAMP sistemi [9] [10]. Autori su se u radu bavili identifikacijom medicinskih pojmova u tekstovima koje su napisali pacijenti, koristeći kraudsoursing [11] .

Na jezicima bivše Jugoslavije nisu pronađeni radovi koji bi se bavili procesom slobodnog teksta u medicinskim izveštajima i javno dostupnim leksičkim izvorima u medicinskom domenu. „Wordnet za biomedicinske nauke“ za srpski jezik sadrži skupove sinonimnih reči ili tačnije različite delove govora (engl. PoS, parts of speech) sa novim konceptom za šest ontoloških kategorija (genetika, virus, bakterije, ćelije, naučna polja i mikroorganizam) [12].

Klasifikacija u medicinskim izveštajima na srpskom jeziku opisana je u radu [13]. De-identifikacija ličnih podataka u srpskom jeziku opisana je u radu [14].

Primena OCR u zdravstvu opisana je u radu [15].

III. Podaci i metodologija

Za potrebe ovog istraživanja sakupljena je 71 slika sa društvene mreže Facebook, koja sadrži medicinski izveštaje koje sadrži lične podatke. Ove izveštaje su okačili korisnici zatvorene grupe Insulinska rezistencija. Neki učesnici u grupi svesni su izlaganja privatnih podataka i pre postavljanja slike svoje ime oboje ili iseku sliku tako da lični podaci nisu dostupni, ali budući da su ovo samo slike od januara 2022. godine, to govori koliko je jednostavno naći elektronske izveštaje koji sadrže lične podatke na internetu na sličnim grupama.

Metoda za prepoznavanje imena sadrži sledeće korake:

- na sve slike je prvo primenjen OCR softver otvorenog koda *Tessaract* [17], čime se izdvaja tekst sa slike (Sl. 1),

```
from PIL import Image
import pytesseract
import numpy as np
import spacy
from spacy import displacy
import translators as ts
import os
import re

directory = 'C:/Users/aldin/Desktop/Podaci/Insulinska'
NER = spacy.load("en_core_web_sm")

for filename in os.listdir(directory):
    f = os.path.join(directory, filename)
    if os.path.isfile(f):
        filename = f
        print(f)
        img1 = np.array(Image.open(filename))
        pytesseract.pytesseract.tesseract_cmd =
        r'C:\Users\aldin\AppData\Local\Tesseract-OCR\tesseract.exe'
        text = pytesseract.image_to_string(img1)
```

Sl. 1 Izdvajanje teksta iz slika

- zatim su primenjena pravila i regularni izrazi za obeležavanje imena i prezimena (Sl. 2). Pravila su definisana od strane autora, a odnose se na to da je u liniji koja se čita veća verovatnoća da se nađe ime i prezime, ako se u toj liniji nalaze reči „prezime“, „pacijent“, „ime i prezime“ i sl. Takođe, ako je na medicinskom izveštaju neka sintagma napisana velikim slovima, ili prvim početnim slovima sa blanko znakovima između, i to treba uzeti u obzir kao moguće ime i prezime pacijenta. Do ovih pravila došlo se analizom sadržaja prikupljenog skupa podataka.

```
        lines = text.split('\n')
        for line in lines:
            if "prezime" in line:
                print(line)
            if "pacijent" in line:
                print(line)
            rl = re.findall('^[a-zA-Z]+(?:[\s.]+[a-zA-Z]+)*$', line)
```

Sl. 2 Primeri pravila

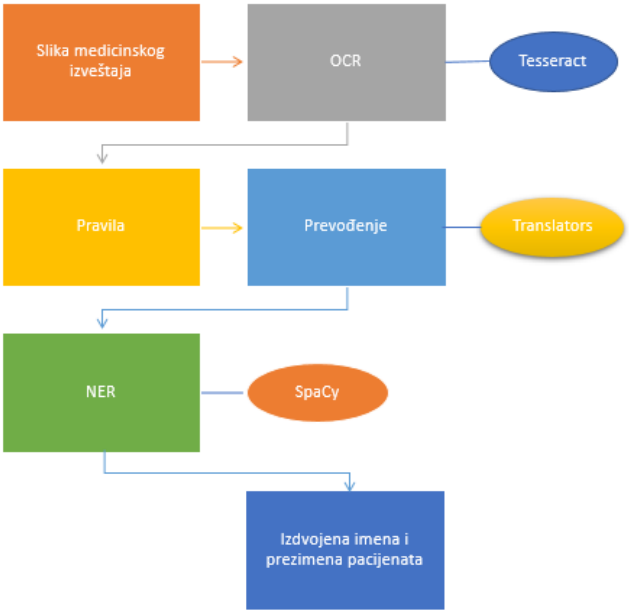
- u koliko je prethodni korak neuspešan, sledi prevođenje teksta na engleski korišćenjem *Python* paketa *Translators* [18],
- zatim se koristi *Spacy* softver za NER na engleskom za prepoznavanje vlastitih imena [19] (Sl. 3).

 ISBN 978-86-7466-930-3

U program su uključene biblioteke *Image* za rad sa slikama, *PyTesseract* za OCR, *NumPy* za jednostavniju obradu i programiranje, zatim *SpaCy* za prepoznavanje imenskih entiteta, *Translators* za prevođenje reči, *os* za rad sa fajlovima i *re* za rad sa regularnim izrazima.

```
        if rl:
            textt = ts.google(rl[0], from_language='sr',
            to_language='en')
            textl= NER(textt)
            for word in textl.ents:
                if word.label_=="PERSON":
                    print(word.text)
```

Sl. 3 Primeri prevođenja i primene Spacy alata za prepoznavanje imenovanih entiteta



Sl. 4. Koraci u prepoznavanju ličnih imena u slikama medicinskih izveštaja na srpskom jeziku

Ova metoda ne zahteva prethodno obeležavanje niti treniranje modela metodama mašinskog učenja (Sl. 4). To su koraci koji su pravci daljeg istraživanja, uz povećanje skupa podataka kako bi se prvobitni rezultati popravili. Uklanjanje ovih podataka na izvornoj slici nije rađeno, ali jeste na tekstovima EHR-ova koji se čuvaju u bazi podataka.

IV. Rezultati i diskusija

U tabeli 1 dati su rezultati primene metode za izdvajanje ličnih imena prezimena. Rezultati uključuju samo one podatke gde je izvučeno samo ime i prezime pacijenta.

Pod preciznošću se smatra odnos broja slika sa izdvojenim ličnim podacima i ukupnog broja slika, a pod odzivom odnos broja tačno izdvojenih imena i prezimena i ukupnog broja slika.

Ono što utiče na rezultate jeste i kvalitet slika i sama preciznost OCR alata, pa je kod jednog broja izveštaja izdvajanje teksta bilo nemoguće. Zato je potrebno uključiti i dodatna predprocesiranja slike. Rezultati pokazuju da je ovo dobra osnova, ali da bi se oni popravili potrebno je razmotriti korišćenje dodatnih resursa za srpski jezik, poput rečnika imena. Takođe, rezultati se mogu popraviti i korišćenjem modela mašinskog učenja i proširenjem trening skupa kao i proširenjem skupa pravila.

TABELA I
Rezultati primene metode

Broj slika	71
Broj slika sa izdvojenim ličnim podacima	41
Broj tačno izdvojenih imena i prezimena	22
Preciznost	58%
Odziv	54%

V. Zaključak

Pacijenti često traže mišljenja o svojim zdravstvenim rezultatima na socijalnim mrežama, u specijalizovanim grupama korisnika sa istom dijagnozom. Tada često postavljaju slike koje sadrže privatne podatke. Alat za de-identifikaciju elektronskih medicinskih izveštaja imao bi široku primenu, jer pored toga što bi se korisniku skrenula pažnja da postavlja dokument sa osetljivim podacima, čak bi se mogli koristiti podaci iz medicinskih informacionih sistema za dalju obradu, analizu i dobijanje znanja. U radu je dat jedan način za izdvajanje imena i prezimena pacijenta zasnovan na OCR i NER. U daljem radu radiće se na dopunjavanju resursa, pravila i uključivanja modela mašinskog učenja, proširenja trening skupa, uz predprocesiranje slika, kako bi se postigli što bolji rezultati. Takođe, jedan od pravaca daljeg istraživanja jeste uklanjanje (zamagljivanje) pronađenih ličnih podataka na procesiranoj slici medicinskog izveštaja.

Zahvalnica

Ovaj rad je delimično finansiran od strane Ministarstva prosvete, nauke i tehnološkog razvoja Republike Srbije u okviru projekata III44007.

Literatura

[1] A. R. Avdić, U. M. Marovac and D. S. Janković, “Smart Health Services for Epidemic Control,” in 2020 55th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST), 2020.

[2] S. Meystre and M. E. f. t. I. Y. S. o. P. Records, "Electronic Patient Records: Some answers to the data representation and reuse challenges: Findings from the section on Patient Records," Yearbook of Medical Informatics, vol. 16, no. 01, p. 47–48, 2007.

[3] M. Luca, E. Lievevrouw and I. V. Hoyweghen, "Fit for purpose? The GDPR and the governance of European digital health," Policy studies, vol. 41, no. 5, pp. 447-467, 2020.

[4] J. Memon, M. Sami, R. A. Khan and M. Uddin, "Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR)," IEEE Access, vol. 8, pp. 142642-142668, 2020).

[5] G. G. Chowdhury, “Natural language processing,” Annual review of information science and technology, vol. 37, no. 1, pp. 51-89, 2003.

[6] D. J. Hand and N. M. Adams, Data mining, Wiley StatsRef: Statistics Reference Online, 2014, pp. 1-7.

 ISBN 978-86-7466-930-3

[7] I. Nikolova, D. Tcharaktchiev, S. Boytcheva, Z. Angelov and G. Angelova, "Applying language technologies on healthcare patient records for better treatment of Bulgarian diabetic patients," in International Conference on Artificial Intelligence: Methodology, Systems, and Applications, Springer, Cham, 2014, September.
[8] A. R. Avdić, U. M. Marovac and D. S. Janković, "Normalization of Health Records in the Serbian Language with the Aim of Smart Health Services Realization," Facta Universitatis, Series: Mathematics and Informatics, pp. 825-841, 2020.
[9] A. M. Milenkovic, P. J. Rajkovic, T. N. Stankovic and D. S. Jankovic, "Application of medical information system MEDIS.NET in professional learning," in 19thTelecommunications Forum (TELFOR) Proceedings of Papers, Belgrade, 2011.
[10] V. Garla, V. L. Re III, Z. Dorey-Stein, F. Kidwai, M. Scotch, J. Womack, A. Justice and C. Brandt, "The Yale cTAKES extensions for document classification: architecture and application," Journal of the American Medical Informatics Association, vol. 18, no. 5, pp. 614-620, 2011.
[11] E. Soysal, J. Wang, M. Jiang, Y. Wu, S. Pakhomov, H. Liu and H. Xu, "CLAMP - a toolkit for eficiently building customized clinical natural language processing pipelines," Journal of the American Medical Informatics Association, vol. 25, no. 3, pp. 331-336, 2018.
[12] D. L. MacLean and J. Heer, "Identifying medical terms in patient-authored text: a crowdsourcing-based approach," Journal of the American Medical Informatics Association, vol. 20, no. 6, pp. 1120-1127, 2013.
[13] N. e. a. Ivković-Berček, "Kooperativan rad na dogradnji Srpskog wordneta.".
[14] A. Avdić, U. Marovac and D. Janković, "Automated labeling of terms in medical reports in Serbian," Turkish Journal of Electrical Engineering & Computer Sciences, vol. 28, no. 6, pp. 3285-3303, 2020.
[15] B. Šandrih, C. Krstev and R. Stanković, "Development and evaluation of three named entity recognition systems for serbian-the case of personal names," in In Proceedings of the International Conference on Recent Advances in Natural Language , 2019, September.
[16] R. Mittal and A. Garg, "Text extraction using OCR: a systematic review. In 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)," 2020, July.
[17] R. Smith, "An overview of the Tesseract OCR engine," in In Ninth international conference on document analysis and recognition (ICDAR 2007), 2007, September.
[18] P. Translators, "https://pypi.org/project/translators/," [Online]. [Accessed May 2022].
[19] X. Schmitt, S. Kubler, J. Robert, M. Papadakis and Y. LeTraon, "A replicable comparison study of NER software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate," in Sixth International Conference on Social Networks Analysis, Management and and Security (SNAMS), 2019, October. *Title of Standard*, Standard number, date.

ABSTRACT

The modern way of life inevitably includes the use of ICT in all its segments, even in healthcare. In addition to e-health, whose development is growing, especially since the world faced the coronary virus pandemic, most people are also looking for health advice on social networks. On that occasion, they upload pictures that contain their health results, with their personal data. This paper presents a method for recognition of personal names in images of medical reports written in Serbian to de-identify them. This method is based on optical character recognition, natural language processing methods and rules and has wide application, as de-identification of electronic medical reports is a necessary step for their further analysis.

Recognition of names on images of medical reports in Serbian to protect personal data

Aldina Avdić, Ulfeta Marovac

 ISBN 978-86-7466-930-3