УНИВЕРЗИТЕТ У НОВОМ САДУ
ФАКУЛТЕТ ТЕХНИЧКИХ НАУКА У НОВОМ САДУ

Аутоматско издвајање именованих ентитета из медицинских докумената на српском језику

ДОКТОРСКА ДИСЕРТАЦИЈА

Ментор:
проф. др Александар Ковачевић

Кандидат:
Александар Каплар

Нови Сад, 2024. године

УНИВЕРЗИТЕТ У НОВОМ САДУ **ОБРАЗАЦ – 5а**
ФАКУЛТЕТ ТЕХНИЧКИХ НАУКА

КЉУЧНА ДОКУМЕНТАЦИЈСКА ИНФОРМАЦИЈА[1]

Врста рада:	Докторска дисертација
Име и презиме аутора:	Александар Каплар
Ментор (титула, име, презиме, звање, институција):	др Александар Ковачевић, редовни професор, Факултет техничких наука, Универзитет у Новом Саду
Наслов рада:	Аутоматско издвајање именованих ентитета из медицинских докумената на српском језику
Језик и писмо рада:	Српски (латиница)
Физички опис рада:	Унети број: Страница____127_____ Поглавља_____8_____ Референци___128_____ Табела_____37_____ Слика_____29_____ Графикона____0_____ Прилога_____2_____
Научна област:	Електротехничко и рачунарско инжењерство
Ужа научна област (научна дисциплина):	Примењене рачунарске науке и информатика
Кључне речи / предметна одредница:	дубоко учење, препознавање именовани ентитета, обрада природног језика, аутоматско издвајање семантике, медицински документи
Апстракт на језику рада:	Мноштво корисних клиничких информација у електронским здравственим картонима је и даље у неструктурираној форми попут извештаја и анамнеза.

[1] Аутор докторске дисертације потписао је и приложио следеће Обрасце:
5б – Изјава о ауторству;
5в – Изјава о истоветности штампане и електронске верзије докторског рада и дозвола за објављивање личних података;
5г – Изјава о коришћењу.
Ове Изјаве се чувају у институцији у штампаном и електронском облику и не кориче се са радом.

	Како би се омогућила употреба наведених информација потребно их је претворити у структурирану форму, односно потребно је извршити екстраховање знања из неструктурираних медицинских докумената. У оквиру ове дисертације, развијен је систем за аутоматско препознавање именованих ентитета из медицинских докумената на српском језику. Развијени систем користи савремене методе дубоког учења, односно састоји се од ансамбла условних случајних поља, рекурентних неуронски мрежа и више-језичких трансформера. Прототип система постигао је Ф1 меру од 0.899, што је упоредиво са нивоом сагласности анотатора. Прототип система развијен у оквиру овог истраживања омогућава екстраховање знања из медицинских докумената на српском језику. Омогућена је употреба именованих ентитета као основу за даља истраживања, која до сада нису била могућа јер су зависила од лексичких ресурса који за српски језик не постоје у адекватној мери. Поред тога што резултати предложеног система представљају међукорак за комплексније системе, они се могу директно користити од стране лекара за увид у дијагнозе, тестове и терапије великог броја пацијената за широк временски период што би било неизводљиво са подацима у неструктурираној форми. Као подршка у истраживању, могу да омогуће лекарима да лакше и брже врше опсервацијске студије.
Датум прихватања теме од стране надлежног већа:	26.01.2023.
Датум одбране: (Попуњава накнадно институција)	
Чланови комисије: (титула, име, презиме, звање, институција)	Председник: др Драган Ивановић, редовни професор, Факултет техничких наука, Нови Сад Члан: др Младен Николић, ванредни професор, Математички факултет, Београд Члан: др Милан Рапаић, редовни професор, Факултет техничких наука, Нови Сад Члан: др Јелена Сливка, ванредни професор, Факултет техничких наука, Нови Сад Ментор: др Александар Ковачевић, редовни професор, Факултет техничких наука, Нови Сад
Напомена:	

UNIVERSITY OF NOVI SAD
FACULTY OF TECHNICAL SCIENCES

KEY WORD DOCUMENTATION[2]

Document type:	Doctoral dissertation
Author:	Aleksandar Kaplar
Supervisor (title, first name, last name, position, institution)	PhD, Aleksandar Kovačević, full professor, Faculty of Technical Sciences, University of Novi Sad
Thesis title in English:	Automatic named entity recognition from medical documents in the Serbian language
Language and script:	Serbian (latin)
Physical description:	Number of: Pages_____127_____ Chapters_____8_____ References____128_____ Tables_____37_____ Illustrations___29_____ Graphs_____ Appendices___2_____
Scientific field:	Electrical and computer engineering
Scientific subfield (scientific discipline):	Applied Computer Science and Informatics
Subject, Key words:	deep learning, named entity recognition, natural language processing, automatic extraction of semantics, medical documents
Abstract in English:	The majority of useful clinical information contained in electronic health records remains in unstructured form such as medical histories and discharge summaries. To enable the use of aforementioned information it needs to be converted into a structured form, that is, knowledge

[2] The author of the doctoral dissertation has signed the following Statements:
5б – Statement on the authorship,
5в – Statement that the printed and e-version of the doctoral dissertation are identical and authorization to use personal data,
5г – Copyright statement.
The paper and e-versions of Statements are held at the institution and are not included into the printed thesis.

	extraction from medical documents needs to be performed first. Within this dissertation, a system for the automatic recognition of named entities from medical documents in the Serbian language has been developed. The developed system utilizes modern deep learning methods, consisting of an ensemble of conditional random fields, recurrent neural networks, and multilingual transformers. The prototype system achieved an F1 score of 0.899, which is comparable to the inter-annotator agreement. The prototype developed in this research enables the extraction of knowledge from medical documents written in the Serbian language. It allows the use of named entities as a basis for further research, which was not previously possible due to the lack of adequate lexical resources for the Serbian language. In addition to the results of the proposed system representing an intermediate step towards more complex systems, they can be directly used by physicians to gain insight into diagnoses, tests, and therapies for a large number of patients over a broad time period, which would be unfeasible with data in unstructured form. As a support in research, they can enable physicians to more easily and quickly conduct observational studies.
Date of endorsement by the scientific board:	26.01.2023.
Date of defence: (Filled in by the institution)	
Thesis defence board: (title, first name, last name, position, institution)	Chair: PhD, Dragan Ivanović, Full Professor, Faculty of Technical Sciences, Novi Sad Member: PhD, Mladen Nikolić, Associate Professor, Faculty of Mathematics, Belgrade Member: PhD, Milan Rapaić, Full Professor, Faculty of Technical Sciences, Novi Sad Member: PhD, Jelena Slivka, Associate Professor, Faculty of Technical Sciences, Novi Sad Mentor: PhD, Aleksandar Kovačević, Full Professor, Faculty of Technical Sciences, Novi Sad
Note:	

Želim da izrazim svoju iskrenu zahvalnost svom mentoru, prof. dr Aleksandru Kovačeviću, na neizmernoj podršci i nesebičnoj pomoći tokom celog procesa istraživanja.

Zahvaljujem se dr Milanu Stošoviću, čija je stručna podrška bila ključna za uspeh ovog istraživanja.

Posebnu zahvalnost dugujem svojoj supruzi i porodici na razumevanju i neprekidnoj podršci.

Ovu tezu posvećujem svojoj ćerki Anastasiji.

Indeks slika

Slika 1. Primer CRF modela za prepoznavanje imenovanih entiteta preuzeto iz (Koller & Friedman, 2009)..... 9

Slika 2 Osnovne razlike u arhitekturi mreža: (a) primer jednog feed forward neurona, (b) primer jenog rekuretnog neurona..... 16

Slika 3. Neuron sa sigmoidnom aktivacionom funkcijom..... 17

Slika 4. Rekurentna neuronska mreža razvijena po ulazima u mrežu 17

Slika 5. Memorijske ćelije LSTM mreže (Olah, 2015) 18

Slika 6. Bidirekciona LSTM CRF mreža (Z. Huang et al., 2015) 20

Slika 7. Arhitekture Word2Vec modela..... 21

Slika 8. Jednostavan primer enkoder-dekoder arhitekture..... 23

Slika 9. Arhitektura transformera (Vaswani et al., 2017). 25

Slika 10. Skalirani skalarni proizvod vektora (Vaswani et al., 2017)..... 26

Slika 11. Mehanizam višestruke pažnje (Vaswani et al., 2017) 27

Slika 12. Arhitektura GPT transformera (Radford et al., 2018) 29

Slika 13. Razlika između BERT i GPT arhitekture za pre-treniranje (Devlin et al., 2018) 29

Slika 14. Arhitektura BERT modela (Devlin et al., 2018)..... 30

Slika 15. Bagging ansambl 33

Slika 16. Boosting ansambl..... 33

Slika 17. Stacking ansambl 34

Slika 18. Tačnost, preciznost i odziv 37

Slika 19. Anotaciona šema..... 56

Slika 20. Primeri anotiranih rečenica..... 57

Slika 21. Proces obuke anotatora 59

Slika 22. Raspodela podataka po skupovima..... 62

Slika 23. Odnos broja primeraka po klasi sa prosečnim brojem tokena po klasi 63

Slika 24. Arhitektura sistema..... 66

Slika 25. Primer pretprocesiranja rečenice. 67

Slika 26. Arhitektura LM-LSTM-CRF mreže 70

Slika 27. Pristupi za korišćene transformer modela: a) pre-treniranje jezičkog modelovanja, b) nastavljanje zadatka pre-trerniranja jezičkog modelovanja postojećih više-jezičkih modela, c) korišćenje pre-treniranih više-jezičkih modela 73
Slika 28. Osnovna podela grešaka sistema. 89
Slika 29. Podela lažno pozitivnih grešaka. 92

Indeks tabela

Tabela 1. Primer predikcija za tri klase 38

Tabela 2. Primer računanja Kappa mere 39

Tabela 3. Smernice za tumačenje *kappa* vrednosti 40

Tabela 4. Primer IOB2 formata 41

Tabela 5. Preciznost, Odziv i F1 mera anotiranih klasa 58

Tabela 6. *Cohen's Kappa* i Tačnost za atribute EVENT i TIMEX3 klasa 60

Tabela 7. Broj entiteta po klasama za obučavajući i test skup 61

Tabela 8. Broj tokena po klasi 62

Tabela 9. Atributi CRF algoritma sa primerom za token "slabost" u rečenici "Za bubrežnu slabost zna od 3.2021. godine" 68

Tabela 10. Hiperparametri LM-LSTM-CRF mreže 71

Tabela 11. Spisak korišćenih modela i korpusa za pre-treniranje 76

Tabela 12. Spisak modela po ansamblu 77

Tabela 13. Preciznost, odziv i F1 mera za tačno poklapanje entiteta. Modeli obeleženi sa $^{(PT)}$ su pretrenirani na skupu od 17000 neanotiranih dokumenata. 80

Tabela 14. Preciznost, odziv i F1 mera za modele na nivou tokena 81

Tabela 15. Performanse Ensemble Best model za tačno poklapanje entiteta klasa 82

Tabela 16. Performanse Ensemble Best model na nivou tokena 82

Tabela 17. F1 mere savremenih modela za medicinski NER nad reprezentativnim skupovima podataka. 86

Tabela 18. Primeri lažno negativnih grešaka. Skraćenice: C – kontekst greške, T – očekivane klase, P – klase dodeljene od strane modela, TREA - TREATMENT, PROB – PROBLEM, OCCU – OCCURRENCE, CLIN – CLINICAL_DEPT 90

Tabela A.1. Rezultati za tačno poklapanje entiteta CRF modela 113

Tabela A.2. Rezultati za tačno poklapanje entiteta LM-LSTM-CRF modela 114

Tabela A.3. Rezultati za tačno poklapanje entiteta T-RoBERTa modela 114

Tabela A.4. Rezultati za tačno poklapanje entiteta BERT Multilingual Cased modela 115

Tabela A.5. Rezultati za tačno poklapanje entiteta BERT Multilingual Uncased modela 115

Tabela A.6. Rezultati za tačno poklapanje entiteta XLM RoBERTa Base modela 116

Tabela A.7. Rezultati za tačno poklapanje entiteta XLM RoBERTa Large modela 116

Tabela A.8. Rezultati za tačno poklapanje entiteta PT – BERT Multilingual Cased modela 117

Tabela A.9. Rezultati za tačno poklapanje entiteta PT – XLM RoBERTa Base modela 117

Tabela B.1. Rezultati na nivou tokena za CRF model 118

Tabela B.2. Rezultati na nivou tokena za LM-LSTM-CRF model 119

Tabela B.3. Rezultati na nivou tokena za T-RoBERTa model 120

Tabela B.4. Rezultati na nivou tokena za BERT Multilingual Uncased model 121

Tabela B.5. Rezultati na nivou tokena za BERT Multilingual Cased model 122

Tabela B.6. Rezultati na nivou tokena za XML RoBERTa Base model 123

Tabela B.7. Rezultati na nivou tokena za XML RoBERTa Large model 124

Tabela B.8. Rezultati na nivou tokena za PT - BERT Multilingual Cased model 125

Tabela B.9. Rezultati na nivou tokena za PT – XML RoBERTa Base model 126

Tabela B.10. Rezultati na nivou tokena za Ensamble Best model 127

Sadržaj

Indeks slika ix
Indeks tabela xi
Rezime xv
Abstract xvii
1. Uvod 1
1.1. Značaj obrade teksta u medicinskim zdravstvenim kartonima 1
1.2. Značaj prepoznavanja imenovanih entiteta u medicinskim dokumentima 2
1.3. Opis cilja, metodologije, rezultata i zaključaka istraživanja 4
2. Teorijske osnove 9
2.1. Uslovna slučajna poljima – Conditional Random Fields 9
2.1.1. Zaključivanje 13
2.2. Rekurentne neuronske mreže - RNN 16
2.3. Vektori reči 20
2.4. Transformer 22
2.4.1. Enkoder-dekoder arhitektura 22
2.4.2. Mehanizam pažnje 23
2.4.3. Arhitektura transformera 24
2.4.4. Varijacije transformer arhitekture 28
2.5. Modeli zasnovani na ansamblima 31
2.6. Metode evaluacije modela 35
2.6.1. Metode evaluacije anotacija 38
2.7. Obeležavanje imenovanih entiteta 40
3. Pregled aktuelnog stanja u oblasti 43
3.1. Prepoznavanje imenovanih entiteta 43
3.2. Prepoznavanje imenovanih entiteta u medicinskim dokumentima 45
3.3. Prepoznavanje imenovanih entiteta u medicinskim dokumentima na srpskom jeziku 47
4. Korpus zlatnog standarda 51
4.1. Anotaciona šema 52

4.2. Proces anotacije 57
4.3. Karakteristike korpusa 60
5. Model sistema za prepoznavanje imenovanih entiteta 65
5.1. Arhitektura sistema 65
5.2. Pretprocesiranje korpusa 66
5.3. Obučavanje modela mašinskog učenja 67
5.3.1. Podela korpusa za obučavanje 67
5.3.2. Uslovna slučajna polja (CRF) 67
5.3.3. Rekurentna neuronska mreža sa dugotrajnom kratkoročnom memorijom (LSTM) 69
5.3.4. Modeli zasnovani na transformerima 72
5.3.5. Modeli zasnovani na ansamblu 76
6. Eksperimentalni rezultati 79
6.1. Eksperimentalna postavka za evaluaciju modela 79
6.2. Preciznost, odziv i F1 mera na nivou modela 80
7. Diskusija 85
7.1. Prednosti i mane modela dubokog učenja 87
7.2. Uticaj pre-treniranja na modele dubokog učenja 88
7.3. Analiza grešaka 89
7.4. Ograničenja 93
8. Zaključak 97
Literatura 99
Biografija 111
Prilozi 113
A. Tabele rezultata za tačno poklapanje entieta po klasama 113
B. Tabele rezultata na nivou tokena 118

Rezime

Osnovni nosioci informacija o pacijentima, u kliničkom kontekstu, su zdravstveni kartoni. U poslednje dve decenije došlo je do značajnog prelaska sa papirnih zdravstvenih kartona na elektronske zdravstvene kartone. U nedavnom periodu dostupnost elektronskih zdravstvenih kartona značajno je povećana, čime je omogućena njihova primena za vršenje različitih medicinskih istraživanja i pronalaženje postupaka za poboljšanje medicinskih usluga. Jedan od ograničavajućih faktora za upotrebu elektronskih zdravstvenih kartona u istraživanjima jeste činjenica da je mnoštvo korisnih kliničkih informacija i dalje u nestrukturiranom obliku, poput anamneza, izveštaja i otpusnih listi. Nestrukturirane dokumente potrebno je pretvoriti u strukturiranu formu kako bi se omogućila njihova upotreba, odnosno potrebno je izvršiti ekstrahovanje znanja iz medicinskih dokumenata.

Obrada prirodnog jezika je multidisciplinarna oblast računarskih nauka i lingvistike, koja se bavi problemom razumevanja i korišćenja prirodnog jezika od strane računarskih sistema. Jedan od osnovnih zadataka obrade prirodnog jezika, prilikom ekstrahovanja znanja, jeste prepoznavanje imenovanih entiteta, čiji je cilj identifikacija imenovanih entiteta (poput imena osoba, organizacija, geografskih lokacija, vremena, itd.) u nestrukturiranim tekstovima. U kliničkom kontekstu, pod imenovanim entitetima se mogu smatrati klinički događaji (simptomi, nazivi lekova, medicinski zahvati, itd.), vremenski podaci (datumi, učestalosti, fraze koje označavaju trajanje simptoma ili terapija, itd.) i vrednosti (numerički podaci poput rezultata laboratorijskih nalaza, doze lekova, itd.). Imenovani entiteti su osnova za mnoge druge zadatke obrade prirodnog jezika a mogu se i direktno koristi na primer za pretragu dokumenata koji pominju simptom „mučnina“.

Problemu prepoznavanja imenovanih entiteta, na engleskom jeziku, u opštem domenu kao i u medicinskom posvećen je relativno veliki broj istraživanja. Dok je relativno mali broj istraživanja posvećen prepoznavanju imenovanih entiteta na srpskom jeziku u opštem domenu, gde je većina istraživanja zasnovana na domenski-specifičnim rešenjima zasnovanih na rečnicima i pravilima. Primenljivost savremenih fleksibilnih rešenja, koja su inicijalno razvijena za engleski jezike, nisu razmatrana.

Cilj disertacije je razvoj sistema za automatsko prepoznavanje imenovanih entiteta u medicinskim dokumentima napisanim na srpskom jeziku. Za

ostvarivanje navedenog cilja prikupljen je korpus medicinskih dokumenata sa Klinike za nefrologiju Univerzitetskog kliničkog centra Srbije. Prikupljeni dokumenti su iskorišćeni za obučavanje savremenih modela dubokog učenja, koji su u literaturi pokazali izvanredne rezultate nad korpusima na engleskom jeziku.

Razvijeni prototip sistema kombinuje više različitih algoritama mašinskog učenja i dubokog učenja. Naime, sistem se sastoji od ansambla uslovnih slučajnih polja, rekurentnih neuronskih mreža i savremenih više-jezičkih transformera. Prototip sistema je ostvario rezultata za F1 meru od 0,899 što je na nivou saglasnosti anotatora, koji su anotirali korpus, čija je F1 mera bila 0,904.

Prototip sistema razvijen u okviru ovog istraživanja omogućava ekstrahovanje znanja iz medicinskih dokumenata na srpskom jeziku. Omogućena je upotreba imenovanih entiteta kao osnovu za dalja istraživanja, koja do sada nisu bila moguća jer su zavisila od leksičkih resursa koji za srpski jezik ne postoje u adekvatnoj meri. Pored toga što rezultati predloženog sistema predstavljaju međukorak za kompleksnije sisteme, oni se mogu direktno koristiti od strane lekara za uvid u dijagnoze, testove i terapije velikog broja pacijenata za širok vremenski period što bi bilo neizvodljivo sa podacima u nestrukturiranoj formi. Kao podrška u istraživanju, mogu da omoguće lekarima da lakše i brže vrše opservacijske studije.

Abstract

The primary carriers of patient information, in a clinical context, are electronic health records. In the past two decades, there has been a significant shift from paper-based health records to electronic health records. Recently, the availability of electronic health records has greatly increased, enabling their use for various medical research and the discovery of procedures to improve medical services. One of the limiting factors for the use of electronic health records, in research, is that the majority of useful clinical data remains in unstructured forms such as medical histories, reports, and discharge summaries. Unstructured documents need to be converted into a structured form in order for them to be used, that is, knowledge extraction from medical documents needs to be performed first.

Natural language processing is a multidisciplinary field of computer science and linguistics that deals with the problem of enabling computer systems to understand and use natural language. One of the basic tasks of natural language processing, when extracting knowledge, is the recognition of named entities, which aims to identify named entities (such as names of people, organizations, geographic locations, times, etc.) in unstructured texts. In a clinical context, named entities can include clinical events (symptoms, names of drugs, medical procedures, etc.), time data (dates, frequencies, phrases indicating the duration of symptoms or therapies, etc.), and values (numerical data such as laboratory results, drug dosages, etc.). Named entities are the basis for many other natural language processing tasks and can be directly used, for example, to search for documents mentioning the symptom „nausea“.

The problem of named entity recognition in the English language, both in general and medical domains, has been the subject of a relatively large number of studies. However, a relatively limited number of studies have been conducted on named entity recognition for the Serbian language, where most of the research, in general domain, was focused on domain-specific solutions which use rule-based and dictionary-based approaches. The applicability of modern flexible solutions, initially developed for English language, has not been considered.

The primary goal of this dissertation is the development of a system for automatic recognition of named entities in medical documents written in the Serbian language. To achieve the aforementioned goal, a corpus of medical documents from the Nephrology Clinic of the University Clinical Center of Serbia was collected. The collected documents were used to

train modern deep learning models, which have shown outstanding results on English-language corpora in the literature.
The developed prototype of the system combines several different machine learning and deep learning algorithms. Namely, the system consists of an ensemble of conditional random fields, recurrent neural networks, and modern multilingual transformers. The prototype of the system achieved an F1 score of 0.899, which is at the level of inter-anotator agreement, which was 0.904.

The prototype developed in this research enables the extraction of knowledge from medical documents written in the Serbian language. It allows the use of named entities as a basis for further research, which was not previously possible due to the lack of adequate lexical resources for the Serbian language. In addition to the results of the proposed system representing an intermediate step towards more complex systems, they can be directly used by physicians to gain insight into diagnoses, tests, and therapies for a large number of patients over a broad time period, which would be unfeasible with data in unstructured form. As a support in research, they can enable physicians to more easily and quickly conduct observational studies.

1. Uvod

1.1. Značaj obrade teksta u medicinskim zdravstvenim kartonima

Iako su najraniji primeri dokumentovanja zdravstvenih podataka pacijenata pronađeni na papirusima i natpisima iz antičkog Egipta, redovno i sistematsko dokumentovanje zdravstvenih podataka se vrši tek od početka 20. veka. Zdravstveni podaci su obično bili pisani na papiru, koji su bili organizovani u fasciklama, tj. zdravstvenim kartonima, po tipu podataka, i pri čemu je samo jedna kopija kartona postojala (Evans, 2016).

Komitet zadužen za poboljšanje zdravstvenih kartona, američkog Instituta za medicinu (*Institute of Medicine)*, je 1991. godine objavio izveštaj sa nazivom *„The Computer-Based Record: An Essential Technology for Health Care“*. U tom izveštaju komitet se zalaže za brz razvoj i implementaciju elektronskih zdravstvenih kartona (eng. *electronic health records*, EHR), jer su identifikovali jedinstven potencijal EHR-a za poboljšanje kvaliteta nege pacijenta, i za istovremeno redukovanje troškova kroz kontinualno poboljšanje kvaliteta nege (Ornstein et al., 1992).

U istom izveštaju, komitet je naveo da papirni zdravstveni kartoni imaju nedostatke u pogledu količine sadržaja, formata, tačnosti, i pristupačnosti podataka za određivanje efektivnosti zdravstvenih usluga i ishoda. Identifikovano je da je oblast medicine informaciono intenzivna, gde se od 35% do 50% vremena utroši na informacione i komunikacione aktivnosti. Procenili su da je oko 70% informacionih potreba lekara neispunjeno prilikom susreta sa pacijentom. Komitet navodi da su navedeni problemi rešivi uz pomoć EHR i informacionih sistema koji bi ih koristili (Ornstein et al., 1992).

Prvi EHR su razvijeni i korišćeni od strane akademskih medicinskih institucija. Opšte korišćenje EHR je sporo napredovalo zbog inicijalne visoke cene prelaska na sistem koji zahteva računare, grešaka prilikom unosa podataka, i inicijalno slabe prihvaćenosti od strane lekara. Kao rezultat toga, prvi sistemi bili su implementirani kao dopuna, a ne zamena, za papirne zdravstvene kartone (Evans, 2016).

Moderni EHR generalno sadrže demografske, administrativne, dijagnostičke, terapijske, i kliničke podatke dobijene tokom rutinskog pružanja zdravstvenih usluga. EHR su omogućili vršenje velikog broja opservacijskih kliničkih istraživanja, epidemioloških studija, studija za evaluaciju primene i bezbednosti lekova, a mogu se primeniti i za studije izvodljivosti i selekciju pacijenata za klinička istraživanja (Cowie et al., 2017).

U protekloj deceniji dostupnost EHR je značajno povećana (W. Lee et al., 2018a; Nayel & Shashirekha, 2017; Peng et al., 2019; S. Wu et al., 2020). Mnoštvo informacija koje se nalaze u EHR pored navedene mogućnosti poboljšanja kvaliteta nege pacijenata mogu identifikovati potencijalne probleme, podržati i olakšati medicinska istraživanja, i smanjiti troškove nege pacijenata (K. Huang et al., 2019; Kovačević et al., 2013; W. Lee et al., 2018a; Y. Wu et al., 2017). Jedna od glavnih prepreka za pristup informacijama u EHR je to što je do 80 procenata važnih kliničkih podataka u nestrukturiranom obliku, poput teksta u anamnezama i otpusnim listama. Kako bi pristupili navedenim informacijama u kliničkim tekstovima, odnosno transformisali u strukturiranu formu pogodnu za upotrebu u računarskim sistemima, istraživači su uložili dosta truda na razvoj metoda za obradu prirodnog jezika (eng. *natural langauge processing*, NLP) (Keretna et al., 2015; S. Wu et al., 2020).

1.2. Značaj prepoznavanja imenovanih entiteta u medicinskim dokumentima

Jedan od najčešćih i veoma bitnih zadataka NLP-a je prepoznavanje imenovanih entiteta (eng. *named entity recognition*, NER), čiji cilj je da se ekstrahuju imenovani entiteti i klasifikuju u unapred određene kategorije poput imena osoba, organizacija, geografskih lokacija, datuma, vremena i sl. Imenovani entiteti su ujedno i osnovni semantički elementi koji su nosioci određenog tipa značenja teksta (Palshikar, 2013).

Zadatak prepoznavanja imenovanih entiteta se istakao kao ključan korak pretprocesiranja tekstova pre primene mnogih drugih NLP zadataka koji zavise od imenovanih entiteta (poput ko-referenciranja i ekstrakcije relacija) (Nayel & Shashirekha, 2017; Si et al., 2019), kao i zadataka kojima pospešuje performanse (poput indeksiranja naučnih članaka, razumevanja relacije između delova teksta i sl.), prvenstveno zbog činjenice da su imenovani entiteti obično usko povezani sa relevantnim informacijama u tekstu (Kozareva et al., 2007).

U medicinskom domenu, cilj NER sistema je identifikacija klinički relevantnih termina u nestrukturiranim medicinskim tekstovima, koje su napisali lekari u kliničkim uslovima, i da se klasifikuju u predefinisane kategorije poput dijagnoza, oboljenja, lekova (de Oliveira et al., 2021; Xu et al., 2018). Na primer, u rečenici „Zbog progresije hronične bubrežne insuficijencije indikovano kreiranje AVF.“, „progresije hronične bubrežne insuficijencije“ pripada kategoriji problema (bolesti) a „kreiranje AVF“ pripada kategorije tretmana.

Identifikacija i razumevanje značenja imenovanih entiteta nije trivijalan zadatak, na primer, razlika između gena pacova (Nf2) i ljudskog gena (NF2), uključujući protein koji taj gen proizvodi i rezultujućeg oboljenja, je razlika između malog i velikog slova (Goulart et al., 2011). Za razliku od drugih domena, medicinski tekstovi u sebi sadrže veliki broj imenovanih entiteta, gde jedan imenovan entitet može imati više različitih oblika, a jedna skraćenica može predstavljati više različitih entiteta u zavisnosti od konteksta u kojem je korišćena (Xu et al., 2018).

Značaj NER se ogleda i u njegovoj primeni u mnogim medicinskim sistemima poput kliničkih sistema za podršku prilikom odlučivanja. Na primer, sistemi za detekciju grešaka prilikom doziranja lekova potrebno je da izvrše integraciju više sistema uključujući i elektronske zdravstvene kartone (Ibáñez-Garcia et al., 2019). Kako bi se pristupilo podacima o primenjenim lekovima i njihovim dozama, iz nestrukturiranih delova elektronskih zdravstvenih kartona, potrebno je primeniti NER.

NER sistemi se mogu koristiti direktno, od strane lekara, za uvid u dijagnoze, testove i terapije velikog broja pacijenta za širok vremenski period što bi bilo neizvodljivo ručnom obradom, odnosno čitanjem dokumenata. Pri čemu mogu da predstavljaju važan izvor podataka lekarima prilikom pronalaženja odgovarajućih pacijenata za opservacijske studije.

U toku razvoja metoda za NLP razne vrste metoda su bile prilagođene i razvijene za klinički NER. U ranim fazama istraživanja kliničkog NER-a istaknuti sistemi, poput MedLEE-a (Friedman et al., 1994) i MetaMap-a (Aronson, 2001), su koristili metode bazirane na rečnicima i pravilima za prepoznavanje imenovanih entiteta (Dehghan et al., 2013). Kasniji razvoji metoda mašinskog učenja su uticali na razvoj metoda kliničkog NER-a. Na primer, 18 od 20 timova koji su učestvovali 2010. i 2012. godine na

i2b2 (Informatics for Integrating Biology and the Bedside) NER zadatku (Sun et al., 2013b; Uzuner et al., 2011) koristili su metodu mašinskog učenja baziranu na uslovnim slučajnim poljima (*Conditional Random Fields*, CRF) (Dehghan et al., 2013). Nedavno, najsavremeniji rezultati su postignuti sa metodama dubokog učenja poput rekurentnih neuronskih mreža sa dugotrajnom kratkoročnom memorijom (*long short-term memory*, LSTM) i modelima baziranim na transformerima (Habibi et al., 2017; Hu & Ma, 2023; Y. Wu et al., 2017).

Većina istraživanja i razvoja metoda kliničkog NLP-a fokusirana je na engleski jezik, dok njihova primenljivost na druge jezik je često nedovoljno istražena (Akhtyamova et al., 2020; W. Lee et al., 2018a).

U pregledu literature metoda dubokog učenja, sprovedenog 2019. godine (S. Wu et al., 2020), istaknuto je da se u manje od 9% studija koriste podaci koji nisu na engleskom ili kineskom jeziku. Srpski jezik je i dalje nedovoljno istražen u domenu kliničkog NER-a i NLP-a. Nekoliko faktora inhibira proces istraživanja, gde su najbitniji: nedostatak adekvatnih leksičkih resursa za srpski jezik (Avdic et al., 2020; Marovac et al., 2023), nepostojanje javno dostupnih korpusa za klinički NER, kompleksnost srpskog jezika sa sedam gramatičkih padeža i tri gramatička roda.

1.3. Opis cilja, metodologije, rezultata i zaključaka istraživanja

Osnovni cilj ovog istraživanja je razvoj sistema za automatsko prepoznavanje imenovanih entiteta koji omogućava direktno korišćenje informacija iz nestrukturiranih delova medicinskih dokumenata, kao i dalja istraživanja iz oblasti medicinskog NLP-a. Dosadašnja istraživanja u domenu medicinskih dokumenata na srpskom jeziku su se fokusirala na razvoju domenski-specifičnih[3] rešenja zasnovanih na rečnicima i pravilima, dok primenljivost savremenih fleksibilnih rešenja, koja su inicijalno razvijena za engleski jezik, nisu razmatrana. Na osnovnu navedenog cilja istraživanja može se definisati polazna pretpostavka (hipoteza):

- **Hipoteza**: Moguće je primeniti savremene tehnike, zasnovane na mašinskom učenju, za automatsko prepoznavanje imenovanih entiteta iz medicinskih dokumenata na srpskom jeziku.

[3] Domenski-specifična rešenja su rešenja koja su usko vezana za jedan domen, ili oblasti, i obično nisu šire primenljiva na druge domene.

Prema polaznoj hipotezi i navedenom cilju moguće je identifikovati više potciljeva bitnih za postizanje cilja istraživanja:

- Prvi potcilj se odnosni na prikupljanje medicinskih dokumenata. Prikupljanje korpusa medicinskih dokumenata je izvršeno uz eksplicitno odobrenje etičke komisije Univerzitetskog kliničkog centra Srbije, uz strogo praćenje protokola za očuvanje privatnosti pacijenata.
- Nakon prikupljanja korpusa medicinskih dokumenata potrebno je, uz konsultovanje medicinskih stručnjaka, identifikovati bitne kategorije imenovanih entiteta i ručno obeležavanje istih kako bi se formirao skup obeleženih medicinskih dokumenata koji će se iskoristiti za obučavanje i evaluaciju sistema za automatsko prepoznavanje imenovanih entiteta.
- Treći potcilj obuhvata utvrđivanje efikasnosti primene i nedostataka savremenih metoda za automatsko prepoznavanje imenovanih entiteta nad obeleženim skupom medicinskih dokumenata na srpskom jeziku.
- Kreiranje specifikacije i dizajna modela prepoznavanje imenovanih entiteta u medicinskim dokumentima na srpskom, na osnovu identifikovanih prednosti i nedostataka savremenih metoda, jeziku je četvrti potcilj.
- Peti potcilj se odnosi na evaluacija razvijenog modela i analizu njegovih grešaka.

Očekivani rezultat navedenog istraživanja obuhvata:

- formiranje anotiranog korpusa medicinskih dokumenata na srpskom jeziku,
- kreiranje modela za prepoznavanje imenovanih entiteta iz medicinskih dokumenata na srpskom jeziku,
- implementacija prototipa sistema za prepoznavanje imenovanih entiteta iz medicinskih dokumenata na srpskom jeziku.

Kao što je navedeno na početku ovog poglavlja, kako bi omogućili korišćenje informacija iz nestrukturiranog teksta, poput teksta u elektronskim zdravstvenim kartonima, potrebno ih je pretvoriti u struktuiranu formu, odnosno potrebno je izvršiti ekstrahovanje znanja iz medicinskih dokumenata. Podaci u struktuiranoj formi se dalje mogu koristiti za pretrage, sumarizacije, podršku prilikom odlučivanja, statističku analizu, i otkrivanje novih informacija poput ishoda terapija i procene rizika (Jensen et al., 2012; Meystre et al., 2008).

Prvi korak u procesu ekstrahovanja znanja je prepoznavanje imenovanih entiteta. Prototip sistema, koji je cilj ovog istraživanja, omogućava ekstrahovanje znanja iz medicinskih dokumenata na srpskom jeziku. Primena savremenih metoda dubokog učenja predstavljaju značajan korak za dalji razvoj medicinskog NER za srpski jezik, tako što omogućavaju istraživanja koja su do sada nisu bila moguća zbog zavisnosti od leksičkih resursa koji za srpski jezik ne postoje u adekvatnoj meri. Pored toga što rezultati predloženog sistema predstavljaju međukorak za kompleksnije sisteme, oni se mogu direktno koristiti od strane lekara za uvid u dijagnoze, testove i terapije velikog broja pacijenta za širok vremenski period što bi bilo neizvodljivo sa podacima u nestrukturiranoj formi. Kao podrška u istraživanju, mogu da omoguće lekarima da lakše i brže vrše opservacijske studije.

Za ostvarivanje navedenog cilja prikupljen je korpus medicinskih dokumenata sa Klinike za nefrologiju Univerzitetskog kliničkog centra Srbije, nad kojima je primenjen algoritam deidentifikacije[4]. Nakon deidentifikacije izvršena je inicijalna analiza dokumenata koja je pokazala da izveštaji pacijenata koji su upućeni na Kliniku za nefrologiju radi dijalize, od strane drugih klinika Univerzitetskog Kliničkog centra Srbije, sadrže mnoštvo medicinskih termina koje se pojavljuju samo jednom u korpusu. Zbog navedenog problema, izabran je podskup medicinskih izveštaja pacijenata koji boluju od hronične i akutne bubrežne insuficijencije kao osnova za dalje istraživanje, odnosno anotiranje[5] dokumenata. Nad anotiranim dokumentima izvršeno je pretprocesiranje u vidu transliteracije sa ćirilice na latinicu i uklanjanje dijakritika. Pretprocesirani dokumenti su iskorišćeni za obučavanje modela mašinskog učenja, odnosno korišćeni su algoritmi uslovnih slučajnih polja, rekurentnih neuronskih mreža i transformeri. Kao najefektivnije rešenje se pokazao ansambl obučenih modela koji je formiran sa strategijom većinskog glasanja. Konačni rezultati najboljeg modela su

[4] Deidentifikacija je postupak koji se koristi za sprečavanje identifikacije identiteta osobe u dokumentu. U kontekstu medicinskih dokumenata, deidentifikacija služi za prepoznavanje i uklanjanje (ili maskiranje) ličnih podataka pacijenta (npr. ime i prezime, jedinstveni matični broj, lični broj osiguranika, mesto stanovanja, datum rođenja i sl.) iz dokumenta kako bi se ti dokumenti mogli koristi u istraživanjima bez narušavanja privatnosti pacijenta.

[5] Anotiranje u navedenom kontekstu označava obeležavanje reči u dokumentima sa odgovarajućom kategorijom imenovanih entiteta.

ostvarili rezultat za F1 meru od 0,899 što je na nivou saglasnosti anotatora koji su anotirali korpus čija je F1 mera bila 0,904.

Struktura ove disertacija je organizovana u devet poglavlja. U ovom, prvom, poglavlju su opisana uvodna razmatranja. Poglavlje se sastoji od tri odeljka. U prvom odeljku opisan je značaj obrade teksta u medicinskim zdravstvenim kartonima. Fokus drugog odeljka je opis značaja prepoznavanja imenovanih entiteta u medicinskim dokumentima. Nakon toga, u trećem odeljku, naveden je cilj istraživanja i polazna hipoteza. Navedeni su potciljevi istraživanja i očekivani rezultati.

Teorijske osnove iz oblasti ove doktorske disertacije date su u drugom poglavlju. Drugo poglavlje je organizovano u pet odeljaka. Prva četiri odeljka opisuju modele mašinskog učenja korišćene za formiranje sistema za prepoznavanje imenovanih entiteta. Prvi odeljak opisuje uslovna slučajna polja koji pripada grupi klasičnih modela mašinskog učenja. Drugi i treći odeljak opisuju rekurentne neuronske mreže i transformere, tim redom, koji pripadaju grupi modela dubokog učenja. Način kombinovanja više modela, odnosno modeli zasnovani na ansamblu, su dati u četvrtom odeljku. Opis tehnika za evaluaciju modela je dat u petom odeljku.

Pregled aktuelnog stanja iz oblasti ove disertacije dat je u trećem poglavlju, koje je organizovano u tri odeljka. Pregled prepoznavanja imenovanih entiteta u opštem domenu je dat u prvom odeljku. Fokus drugog odeljka je pregled stanja za prepoznavanje imenovanih entiteta u medicinskom domenu. U trećem odeljku je fokus na prepoznavanje imenovanih entiteta u medicinskom domenu na srpskom jeziku.

Fokus četvrtog poglavlja je opis načina formiranja korišćenog korpusa medicinskih dokumenata. U prvom odeljku je opisana anotaciona šema, skup klasa imenovanih entiteta, korišćena prilikom procesa anotiranja podataka. Proces anotacije i rezultati slaganja anotatora su dati u drugim odeljku. Opis konačnih karakteristika anotiranog korpusa je dat u trećem odeljku.

Model sistema za prepoznavanje imenovanih entiteta u medicinskim dokumentima je dat u petom poglavlju. Peto poglavlje se sastoji iz tri odeljka. Prvi odeljak prestavlja opštu arhitekturu razvijenog sistema za prepoznavanje imenovanih entiteta. Drugi odeljak opisuje prvi korak koji izvršava sistem, odnosno postupak pretprocesiranja medicinskih

dokumenata iz korpusa. U trećem odeljku je opisan postupak obučavanja i formiranja konačnih modela iskorišćenih za formiranje sistema.

Eksperimentalna postavka, evaluacija performansi individualnih modela i modela sistema, kao i prikaz rezultata najboljih modela je dat u šestom poglavlju u tri odeljka.

Sedmo poglavlje sadrži diskusiju rezultata i analizu grešaka konačno formiranog sistema za prepoznavanje imenovanih entiteta. U sedmom poglavlju je navedena i diskusija o ograničenjima ovog istraživanja.

Konačni zaključak ove disertacije sa fokusom na ostvarene rezultate i dalje pravce istraživanja i razvoja naveden je u osmom poglavlju.

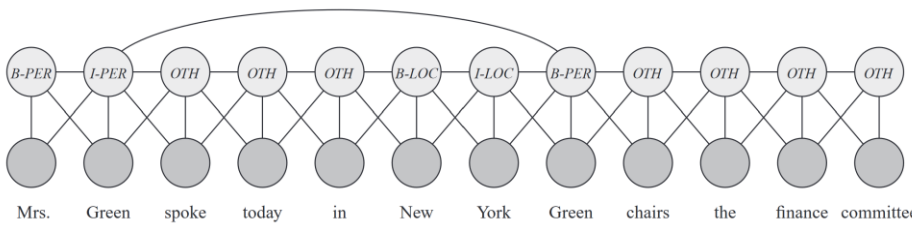
Nakon zaključka navedena je literatura na koju se oslanja istraživanje iz ove disertacije.

2. Teorijske osnove

U ovom poglavlju dat je kratak opis korišćenih algoritama. Prvo su navedeni algoritmi iskorišćeni za formiranje modela za prepoznavanje imenovanih entiteta. Nakon navedenih algoritama opisane su metrike koje se koriste za evaluaciju performansi modela.

2.1. Uslovna slučajna poljima – Conditional Random Fields

Uslovna slučajna polja (Conditional Random Fields, CRF) predstavljaju probabilistički model za segmentiranje i labeliranje sekvencijalnih podataka (Lafferty et al., 2001). CRF se može posmatrati kao Markovljeva mreža gde se neusmerena grafovska reprezentacija i parametrizacija koristi da predstavi uslovnu verovatnoću (Koller & Friedman, 2009). Primer CRF modela dat je na slici 1.



Slika 1. Primer CRF modela za prepoznavanje imenovanih entiteta preuzeto iz (Koller & Friedman, 2009).

Pre formalne definicije CRF modela u nastavku su sažeto navedene potrebne teorijske osnovne.

Neka je $\boldsymbol{V}$ skup slučajnih promenljivih, a $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ disjunktni podskupovi skupa $\boldsymbol{V}$. Neka je P zajednička raspodela verovatnoće nad $\boldsymbol{V}$. Za $\boldsymbol{X}$ kažemo da je uslovno nezavistan od $\boldsymbol{Y}$ na osnovu $\boldsymbol{Z}$ u raspodeli P akko važi $P(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{Z}) = P(\boldsymbol{X} \mid \boldsymbol{Z})P(\boldsymbol{Y} \mid \boldsymbol{Z})$. Odnosno, ako P zadovoljava $(\boldsymbol{X} = \boldsymbol{x} \perp \boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{Z} = \boldsymbol{z})$ za sve vrednosti $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$. (Aleksandar Kovačević, 2011; Koller & Friedman, 2009).

Raspodele verovatnoće za koje važe navedene uslovne nezavisnosti mogu se kompaktno i efektivno predstaviti uz pomoć grafovskih modela (Aleksandar Kovačević, 2011; Sutton et al., 2012). Zajednička raspodela verovatnoće nad skupom promenljivih može predstaviti kao proizvod

lokalnih funkcija koje zavise od mnogo manjeg podskupa promenljivih (Sutton et al., 2012).

Graf se može definisati kao struktura podataka $G = (V, E)$ koja se sastoji od skupa čvorova (V) i skupa grana (E). Par čvorova V_i, V_j iz skupa čvorova mogu biti spojene sa *usmerenom granom* ili sa *neusmerenom granom*. Grafovi mogu biti *usmereni* ukoliko su čvorovi spojeni isključivo sa usmerenim granama, *neusmereni* ukoliko su čvorovi spojeni isključivo sa neusmerenim granama, ili *delimično usmereni* ukoliko su čvorovi spojeni sa usmerenim i neusmerenim granama (Koller & Friedman, 2009). Parametrizacija neusmerenih grafova se vrši uz pomoć faktora.

Definicija 1: Neka je $\boldsymbol{V}$ skup slučajnih promenljivih. Faktor ψ se definiše kao funkcija koja preslikava vrednost iz $\boldsymbol{V}$ u $\mathbb{R}$, $\psi: Val(\boldsymbol{V}) \rightarrow \mathbb{R}$.

Neka je dat podskup $\boldsymbol{D} = a \subset \boldsymbol{V}$, neusmerni grafovski model se može definisati kao skup raspodela koje mogu biti zapisane u faktorisanom obliku

$$P(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{Z} \prod_{a \in \boldsymbol{D}} \psi_a(\boldsymbol{x}_a, \boldsymbol{y}_a), \tag{1}$$

gde je Z normalizaciona konstanta data u obliku

$$Z = \sum_{\boldsymbol{x},\boldsymbol{y}} \prod_{a \in \boldsymbol{D}} \psi_a(\boldsymbol{x}_a, \boldsymbol{y}_a). \tag{2}$$

Funkcija $\psi_a(\boldsymbol{x}_a, \boldsymbol{y}_a)$ se obično posmatra u formi

$$\psi_a(\boldsymbol{x}_a, \boldsymbol{y}_a) = exp\left\{\sum_k \theta_{ak} f_{ak}(\boldsymbol{x}_a, \boldsymbol{y}_a)\right\}, \tag{3}$$

gde je parametar θ_a vektor realnih brojeva, a f_{ak} funkcija nad atributima (Sutton et al., 2012).

Neusmereni graf G se može smatrati Markovljevom mrežom ukoliko zadovoljava osobinu Markova. Odnosno, za slučajnu promenljivu (čvor) $S \in \boldsymbol{V}$, neka je $N(S)$ skup suseda promenljive S u grafu G. Za raspodelu verovatnoće P predstavljenu pomoću neusmerenog grafa G za svaku slučajnu promenljivu važi:

$$P(S \mid \boldsymbol{V} - S) = P(S \mid N(S)). \quad (4)$$

Raspodela P_Ψ, gde je $\Psi = \{\psi_1(\boldsymbol{D}_1), \dots, \psi_k(\boldsymbol{D}_k)\}$ faktoriše Mrakovu mrežu G ako svaki $\boldsymbol{D}_k$ $(k = 1, \dots, K)$ je potpuno povezani podgraf, faktori koji parametrišu Markovu mrežu često se nazivaju potencijali klika (Koller & Friedman, 2009). Odnosno, ukoliko raspodela P_Ψ se faktoriše po G, onda P_Ψ zadovoljava osobinu Markova nad grafom G.

Za definisanje CRF modela možemo da posmatramo nepoznat proces koji generiše vrednosti koje se mogu pregledati, npr. automatsko određivanje vrste reči u rečenici. Neka nam je poznata struktura navedenog procesa, gde su poznati skup $\boldsymbol{X}$ koji predstavlja skup promenjivih sa poznatim vrednostima i skup $\boldsymbol{Y}$ koji predstavlja skup promenljivih koje predstavljaju željene kategorije (Aleksandar Kovačević, 2011). Modelovanje navedenog procesa se može izvršiti uz pomoć neusmerenog grafa, gde se umesto zajedničke raspodele $P(\boldsymbol{Y}, \boldsymbol{X})$ vrši reprezentacija uslovne raspodele $P(\boldsymbol{Y} \mid \boldsymbol{X})$. U kontekstu Markovljevih mreža, ova raspodela se naziva *uslovna slučajna polja (*eng. *conditional random fields)* (Koller & Friedman, 2009; Sutton et al., 2012).

Za uspešno formiranje navedenog modela procesa potrebno je realizovati *zadatak učenja* i *zadatak zaključivanja*. Za zadatak učenja, ako je dat skup vrednosti $\boldsymbol{X}$ i skup kategorija $\boldsymbol{Y}$, potrebno je pronaći funkciju nad atributima tako da se optimizuje neki predefinisani kriterijum. Zaključivanje kao zadatak treba da pronađe najverovatniji skup kategorija $\boldsymbol{y}^*$ za datu vrednost $\boldsymbol{x}$, odnosno potrebno je izračunati $\boldsymbol{y}^* = argmax_y P(\boldsymbol{y} \mid \boldsymbol{x})$ (Aleksandar Kovačević, 2011).

Za određivanje kategorija podataka u obliku sekvence često se koristi linearni CRF (eng. *linear chain CRF*), odnosno model u kome čvorovi iz skupa kategorija $\boldsymbol{Y}$ formiraju linearni lanac (Aleksandar Kovačević, 2011). U kontekstu prepoznavanja imenovanih entiteta, linearni CRF se obično sastoji od dva faktora za svaku reč: prvi faktor se koristi za predstavljanje zavisnosti između trenutne i prethodne kategorije reči, drugi faktor koji predstavlja zavisnosti između trenutne kategorije i konteksta reči (odnosno atributa).

Definicija 2: CRF je neusmerni graf G čiji čvorovi su iz $\boldsymbol{X} \cup \boldsymbol{Y}$, sa skupom faktora $\psi_1(\boldsymbol{D}_1), \dots, \psi_k(\boldsymbol{D}_k)$ tako da svaki $\boldsymbol{D}_i \not\subseteq \boldsymbol{X}$. Graf predstavlja uslovnu raspodelu:

$$P(\boldsymbol{Y} \mid \boldsymbol{X}) = \frac{1}{Z(\boldsymbol{X})} \hat{P}(\boldsymbol{Y}, \boldsymbol{X}) \quad (5)$$

$$\hat{P}(\boldsymbol{Y}, \boldsymbol{X}) = \prod_{i=1}^{k} \psi_i(\boldsymbol{D}_i) \quad (6)$$

$$Z(\boldsymbol{X}) = \sum_{\boldsymbol{Y}} \hat{P}(\boldsymbol{Y}, \boldsymbol{X}) \quad (7)$$

Ukoliko posmatramo CRF nad $\boldsymbol{Y} = \{Y_1, \dots, Y_k\}$ i $\boldsymbol{X} = \{X_1, \dots, X_k\}$ sa granama $Y_i - Y_{i-1}$ i sa granama $Y_i - X_i$, onda iz gore navedene definicije raspodela ima sledeći oblik:

$$P(\boldsymbol{Y} \mid \boldsymbol{X}) = \frac{1}{Z(\boldsymbol{X})} \hat{P}(\boldsymbol{Y}, \boldsymbol{X}) \quad (8)$$

$$\hat{P}(\boldsymbol{Y}, \boldsymbol{X}) = \prod_{i=2}^{k} \psi_i(Y_i, Y_{i-1}) \prod_{i=1}^{k} \psi_i(Y_i, X_i) \quad (9)$$

$$Z(\boldsymbol{X}) = \sum_{\boldsymbol{Y}} \hat{P}(\boldsymbol{Y}, \boldsymbol{X}) \quad (10)$$

Feleksibilnost CRF se ogleda u tome što se ne reprezentuje raspodela nad promenljivima iz $\boldsymbol{X}$, što omogućava modelu da koristi značajan skup posmatranih promenljivih čije zavisnosti mogu biti kompleksne (Koller & Friedman, 2009).

Gore navedena formula se može zapisati u kompaktnijem obliku uvođenjem funkcija nad atributima. Svaka od funkcija nad atributima je u obliku $f_k(y_t, y_{t-1}, x_t)$, gde je potrebno da postoji atribut u formi $f_{ij}(y, y', x) = \mathbf{1}\{y = i\}\mathbf{1}\{y' = j\}$ za svaku granu između čvorova skupa $\boldsymbol{Y}$ i da postoji atribut $f_{io}(y, y', x) = \mathbf{1}\{y = i\}\mathbf{1}\{x = o\}$ za svaku granu između $\boldsymbol{Y}$ i $\boldsymbol{X}$ (Sutton et al., 2012). Iz navedenih formula može da se definiše linearni CRF:

Definicija 3: Neka su $\boldsymbol{Y}$ i $\boldsymbol{X}$ nasumični vektori, neka je $\theta = \{\theta_k\} \in \mathbb{R}^K$ vektor parametara, i $\{f_k(y, y', \boldsymbol{x}_t)\}_{k=1}^K$ skup funkcija nad atributima. Onda je linerani CRF raspodela $P(\boldsymbol{Y} \mid \boldsymbol{X})$ data u formi:

$$P(\boldsymbol{Y} \mid \boldsymbol{X}) = \frac{1}{Z(\boldsymbol{X})} \prod_{t=1}^{T} exp\left\{\sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, \boldsymbol{x}_t)\right\}, \quad (11)$$

gde je $Z(\boldsymbol{X})$ normalizaciona funkcija data u formi:

$$Z(\boldsymbol{X}) = \sum_{y} \prod_{t=1}^{T} exp\left\{\sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, \boldsymbol{x}_t)\right\} \quad (12)$$

(Sutton et al., 2012).

2.1.1. Zaključivanje

Za zadatak zaključivanja s CRF modelom koriste se algoritmi skrivenih Markovljevih mreža. Za određivanje marginalne raspodele, odnosno za određivanje $Z(\boldsymbol{X})$ prilikom obučavanja, koristi se algoritam napred-nazad (eng. *forward-backward*) i Viterbijev algoritam za određivanje najverovatnijeg niza kategorija.
Algoritam napred-nazad vrši određivanje marginalne raspodele $P(\boldsymbol{X}) = \sum_{\boldsymbol{Y}} P(\boldsymbol{X}, \boldsymbol{Y})$ upotrebom zakona distributivnosti:

$$\begin{aligned} P(\boldsymbol{X}) &= \sum_{\boldsymbol{Y}} \prod_{t=1}^{T} \psi_t\,(y_t, y_{t-1}, x_t) \\ &= \sum_{y_T} \sum_{y_{T-1}} \psi_T(y_T, y_{T-1}, x_T) \sum_{y_{T-2}} \psi_{T-1}(y_{T-1}, y_{T-2}, x_{T-1}) \ldots \end{aligned} \quad (13)$$

Iz navedenog mogu se definisati alfa varijable (α_t) za prolaz napred deo algoritma:

$$\begin{aligned} \alpha_t(j) &\stackrel{\text{def}}{=} P(\boldsymbol{x}_{1\ldots t}, y_t = j) \\ &= \sum_{\boldsymbol{y}_{1\ldots t-1}} \psi_t(j, y_{t-1}, x_t) \prod_{t'=1}^{t-1} \psi_{t'}(y_{t'}, y_{t'-1}, x_{t'}) \end{aligned} \quad (14)$$

gde suma nad $\boldsymbol{y}_{1\ldots t-1}$ obuhvata sve moguće vrednosti slučajnim promenljivim $y_1, y_2, \ldots, y_{t-1}$ (Sutton et al., 2012). Alfa vrednosti mogu se izračunati uz pomoć rekurzije:

$$\alpha_t(j) = \sum_{i \in S} \psi_t(j, i, x_t)\alpha_{t-1}(i), \tag{15}$$

gde je inicijalna vrednost $\alpha_1(j) = \psi_1(j, y_0, x_1)$ (Sutton et al., 2012). Analogno alfa varijablama definišu se beta varijable (β_t) za prolaz nazad deo algoritma:

$$\begin{aligned} \beta_t(i) &\overset{\text{def}}{=} P(\boldsymbol{x}_{t+1\ldots T}, y_t = i) \\ &= \sum_{\boldsymbol{y}_{t+1\ldots T}} \prod_{t'=t+1}^{T} \psi_{t'}(y_{t'}, y_{t'-1}, x_{t'}), \end{aligned} \tag{16}$$

sa rekuzijom:

$$\beta_t(i) = \sum_{j \in S} \psi_{t+1}(j, i, x_{t+1})\beta_{t+1}(j), \tag{17}$$

i inicijalizacijom $\beta_T(i) = 1$(Sutton et al., 2012).

Na osnovu alfa i beta varijabli verovatnoća $P(\boldsymbol{X})$ se može izračunati kao $P(\boldsymbol{X}) = \sum_{y_T} \alpha_T(y_T)$ i $P(\boldsymbol{X}) = \beta_0(y_0) \overset{\text{def}}{=} \sum_{y_1} \psi_1(y_1, y_0, x_1)\,\beta_1(y_1)$ (Sutton et al., 2012). Kombinovanjem alfa i beta rekurzije dobija se:

$$P(y_{t-1}, y_t | \boldsymbol{x}) \propto \alpha_{t-1}(y_{t-1})\psi_t(y_t, y_{t-1}, x_t)\beta_t(y_t). \tag{18}$$

Za računanje najverovatnijeg niza kategorija $\boldsymbol{y}^* = argmax_y P(\boldsymbol{y} \mid \boldsymbol{x})$ koristi se Viterbijeva rekurzija koja sumiranje u gore navedenoj marginalnoj raspodeli menja sa maksimizacijom:

$$\delta_t(j) = \max_{i \in S} \psi_t(j, i, x_t)\delta_{t-1}(i), \tag{19}$$

$$max_y P(\boldsymbol{y} \mid \boldsymbol{x}) = \delta_T(j) \tag{20}$$

Upotrebom definicije CRF dobija se:

$$P(y_{t-1}, y_t|\boldsymbol{x}) \propto \alpha_{t-1}(y_{t-1}) exp\left\{\sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, x_t)\right\} \psi_t \beta_t(y_t) \quad (21)$$

i

$$\delta_t(j) = \max_{i \in S} exp\left\{\sum_{k=1}^{K} \theta_k f_k(j, i, x_t)\right\} \delta_{t-1}(i) \quad (22)$$

2.1.1.1.1. Obučavanje CRF modela

Obučavanje CRF modela predstavlja postupak određivanja parametara $\theta = \{\theta_k\}$ na osnovu datih funkcija osobina i anotiranog skupa podataka (obučavajuće skupa) (Aleksandar Kovačević, 2011; Sutton et al., 2012). Određivanja parametara CRF modela prilikom postupka obučavanja se vrši tako da vrednost parametara ima maksimalnu verodostojnost (eng. *maximum likelihood*) nad obučavajućim skupom.
Određivanje parametara se obično vrši tako da vrednost uslovne logaritamske verodostojnosti bude maksimalna na obučavajućem skupu:

$$l(\theta) = \sum_{i=1}^{N} \log p\left(\boldsymbol{y}^{(i)}|\boldsymbol{x}^{(i)}\right), \quad (23)$$

gde je N broj primera u obučavajućem skupu.
Ubacivanjem definicije CRF u jednačinu iznad dobija se:

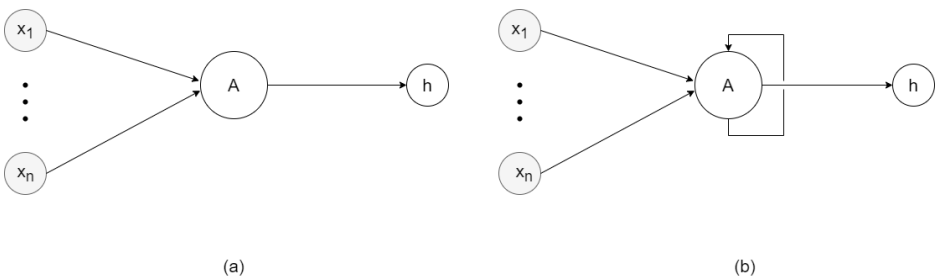
$$l(\theta) = \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{k=1}^{K} \theta_k f_k\left(y_t^{(i)}, y_{t-1}^{(i)}, x_t^{(i)}\right) - \sum_{t=1}^{N} \log Z(x^{(i)}). \quad (24)$$

Određivanje vrednosti parametara θ analitičkim putem, tako što se parcijalni izvod izraza (iznad) izjednači sa nulom, je vrlo teško. Tipično korišćeni pristupi za određivanje parametara θ su iterativni postupci ili tehnike zasnovane na gradijentu (Aleksandar Kovačević, 2011).

Za implementaciju CRF modela u okviru ove disertacije iskorišćen je CRF-Suite[6] iz *scikit-learn* biblioteke (Pedregosa et al., 2011). Navedena biblioteka za određivanje vrednosti parametara θ prilikom obučavanja koristi L-BFGS algoritam (D. C. Liu & Nocedal, 1989).

2.2. Rekurentne neuronske mreže - RNN

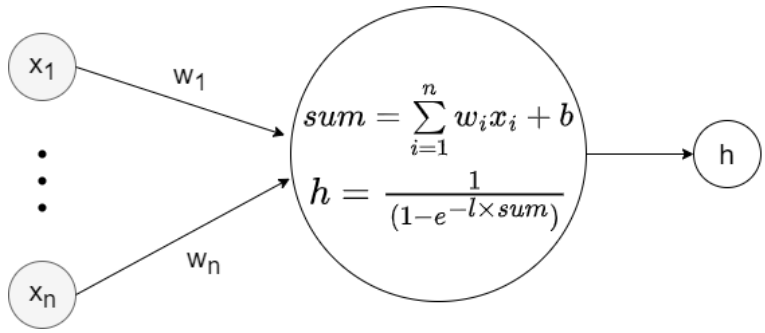
U toku razvoja metoda mašinskog učenja razvijene su različite arhitektura veštačkih neuronskih mreža za različite tipove zadataka. Kada je reč o zadatku klasifikacije i regresije uglavnom se koriste neuronske mreže sa propagacijom unapred (eng. *feed-forward*), za obradu slike najčešće se koriste konvolutivne neuronske mreže, dok se za obradu sekvencijalnih podataka se obično koriste rekurentne neuronske mreže. Razlike između standardnih neuronskih mreža sa propagacijom unapred i rekurentnih neuronskih mreža je u tome što neuroni u rekurentnom sloju mreže, uz pomoć rekuretne veze (Slika 2.b.), koriste sopstvene izlaze kao ulaznu vrednost pri računanju izlaza za sledeću iteraciju.



Slika 2 Osnovne razlike u arhitekturi mreža: (a) primer jednog feed forward neurona, (b) primer jenog rekuretnog neurona

Osnovni proračun izlaza (aktivacije) neurona u mreži sa propagacijom unapred (Slika 3) se računa kao zbir slobodnog koeficijenta (eng. *bias*) i sume proizvoda ulaza i njihovih težina koji se prosleđuje aktivacionoj funkciji. Aktivaciona funkcija ima ključnu ulogu u uvođenju nelinearnosti u neuronsku mrežu. U odsustvu aktivacione funkcije, mreža bi se sastojala isključivo od linearnih kombinacija ulaza, što bi rezultiralo celokupnom linearnošću mreže u odnosu na ulaze. Primena aktivacione funkcije omogućava mreži da postane nelinearna u odnosu na ulaze, što je suštinsko svojstvo za njenu sposobnost da modeluje složene, nelinearne odnose između ulaznih podataka i izlaznih rezultata.

[6] https://sklearn-crfsuite.readthedocs.io/en/latest/

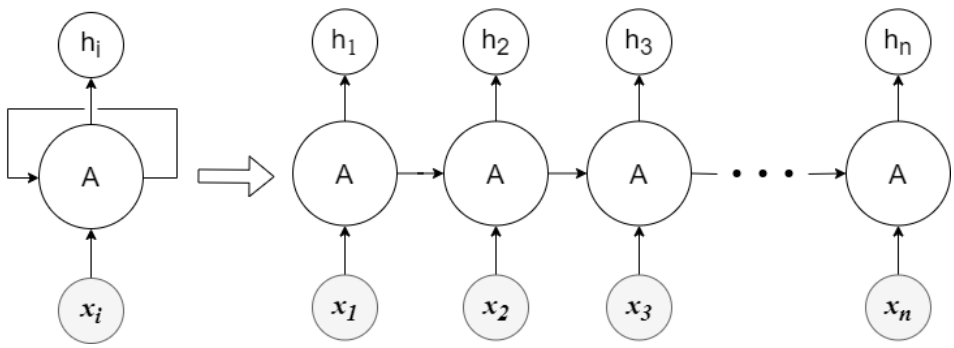


Slika 3. Neuron sa sigmoidnom aktivacionom funkcijom

Dodavanje rekurentne veze neuronu, prikazanom na slici 3, proširuje računanje zbira sa proizvodom izlaza (aktivacije) neurona u prethodnom koraku ($h^{(t-1)}$) sa njegovom težinom (c_j):

$$sum^{(t)} = \sum_{i=1}^{n} w_i x_i^{(t)} + b + \sum_{j=1}^{r} c_j h^{(t-1)} \qquad (25)$$

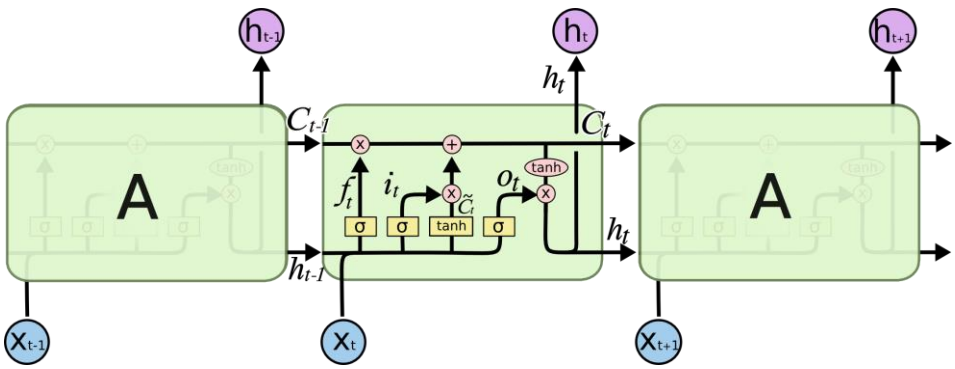
Rekurentna veza omogućava da se sačuvaju informacije o prethodnim aktivacijama (kratkotrajna memorija) kako bi se izračunala vrednost tekuće aktivacije. Sama rekurentna veza se može posmatrati kao sloj neurona gde rekurentne veze su veze koje sekvencijalno povezuju ulazne vektore (Slika 4). Navedena kratkotrajna memorija je korisna u mnogim zadacima poput predikcija reči na osnovu unetih slova (poput predlaganja reči i korekcija na pametnim telefonima), procesiranju govornog jezika, predikcija sledeće reči u rečenici i sl.



Slika 4. Rekurentna neuronska mreža razvijena po ulazima u mrežu

Jedan od osnovnih problema sa rekurentnim mrežama se javlja prilikom njihovog obučavanja. Neuronske mreže se obučavaju tako što se vrednosti težina podešavaju propagacijom greške iz izlaznog sloja tako što se računa gradijent greške na osnovu aktivacione funkcije. Ukoliko je broj slojeva u neuronskoj mreži veliki, može doći do problema nestajanja gradijenta (eng. *vanishing gradient*) što prouzrokuje da se težine u početnim slojevima ne menjaju. Kod propagacije greške u rekurentnim vezama, u zavisnosti od inicijalnih vrednosti težina, može doći do eksponencijalnog rasta ili nestajanja gradijenta (Hochreiter & Schmidhuber, 1997). Kako bi rešili navedeni problem, autori (Hochreiter & Schmidhuber, 1997), predložili su rekurentnu arhitekturu sa dugotrajnom kratkoročnom memorijom (eng. *long short-term memory*, LSTM).

Autori LSTM arhitekture uvode pojam memorijske ćelije i koncept kapija. Problem nestajanja gradijenta rešavaju uz pomoć konstantne propagacije greške kroz rekurentnu jedinicu (eng. *constant error carousel*). Ilustracija memorijske ćelije LSTM mreže, preuzeta od (Olah, 2015), dat je na slici 5.



Slika 5. Memorijske ćelije LSTM mreže (Olah, 2015)

Za razliku od standardne rekurentne neuronske mreže gde se jedinica sastoji od jednog neurona, LSTM memorijska jedinica se sastoji od četiri (žuti pravougaonici na slici 5). Osnova memorijske ćelije je njeno stanje (eng. *constant error carousel*) obeleženo sa C_t. Na stanje memorijske ćelije utiču kapije, prva kapija f_t služi da izbaci nerelevantne informacije iz prethodnog stanja (C_{t-1}), navedena kapija se obično naziva kapija zaboravljanja (*forget gate*). Navedena kapija koristi izlaz ćelije u prethodnom koraku (h_{t-1}) i trenutni ulaz (x_t) koji se prosleđuje kroz sigmoidalnu aktivacioni funkciju:

$$f_t = \sigma\left(W_f^{(x)} x_t + W_f^{(h)} h_{t-1} + b_f\right). \quad (26)$$

Za svaku vrednost u C_{t-1}, f_t generiše vrednost između 0 i 1 kao bi se vrednosti iz stanja ćelije bile izbačene ili sačuvane. Nakon prve kapije generišu se vrednosti sa kojima će se ažurirati trenutno stanje ćelije. Ažuriranje se vrši iz dva koraka, ulazne kapije i_t (input gate) koja određuje koja vrednosti u vektoru stanja ćelije želimo da ažuriramo i sloja za generisanje novih vrednosti stanja ćelije $\widetilde{C_t}$:

$$i_t = \sigma\left(W_i^{(x)} x_t + W_i^{(h)} h_{t-1} + b_i\right), \quad (27)$$

$$\widetilde{C_t} = tanh\left(W_C^{(x)} x_t + W_C^{(h)} h_{t-1} + b_C\right). \quad (28)$$

Novo stanje ćelije C_t se računa po formuli:

$$C_t = f_t \times C_{t-1} + i_t \times \widetilde{C_t}. \quad (29)$$

Izlaz iz memorijske ćelije se računa uz pomoć izlazne kapije o_t (output gate) koji se množi sa novim stanjem ćelije koje je propušteno kroz tangentnu aktivacionu funkciju:

$$o_t = \sigma\left(W_o^{(x)} x_t + W_o^{(h)} h_{t-1} + b_o\right), \quad (30)$$

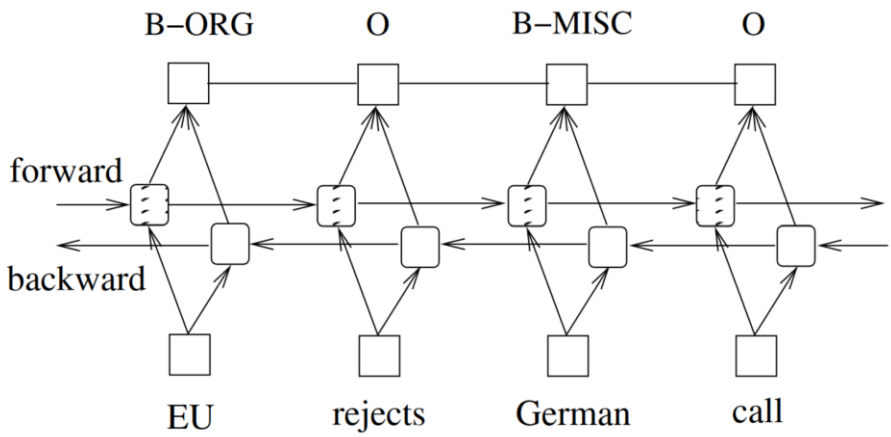
$$h_t = o_t \times \tanh(C_t). \quad (31)$$

Navedena arhitektura memorijske ćelije omogućava formiranje rekurentnih neuronskih mreža sa sposobnošću pamćenja dugih sekvenci (do 1000 koraka) bez gubljenja sposobnosti za kratkoročnu memoriju, odnosno pamćenje kratkih sekvenci (Hochreiter & Schmidhuber, 1997).

Određivanje težinskih parametara LSTM mreže (obučavanje) se najčešće vrši upotrebom BPTT (eng. *Backpropagation through time*) algoritma (Williams & Zipser, 1995) ili RTRL (eng. *Real-Time Recurrent Learning*) algoritma (Robinson & Fallside, 1987; Williams & Zipser, 1989).

U ovom radu iskorišćena je bidirekciona LSTM mreža sa CRF slojem kao jedan od modela za prepoznavanje imenovanih entiteta. Model bidirekcione LSTM CRF mreže predložen od strane autora (Z. Huang et al., 2015) je dat na slici 6.

Bidirekciona LSTM CRF mreža se sastoji od dve LSTM mreže imenovane *forward* i *backward* na slici 6, kao i CRF mreže koja je na slici prikazana kao veza između izlaza. Kako bi se poboljšala tačnost prilikom prepoznavanja imenovanih entiteta, bidirekciona LSTM CRF mreža koristi informacije o prošlim ulazima u mrežu putem *forward* LSTM mreže, koristi informacije o budućim ulazima na osnovu *backward* LSTM mreže, kao i informacija o tagovima na nivou rečenice na osnovu CRF mreže. Algoritam predložene od strane (Z. Huang et al., 2015) za ulaznu rečenicu prvo izračunava izlaze iz LSTM mreža. Izlaz iz LSTM mreža koristi se kao ulazi u CRF, gde se primenom napred-nazad algoritma (opisanog u poglavlju 2.1.1) dobijaju izlazi iz CRF mreže. Dobijeni izlazi koriste se za primenu tehnika zasnovanih na gradijentu za podešavanje parametara LSTM i CRF mreža.



Slika 6. Bidirekciona LSTM CRF mreža (Z. Huang et al., 2015)

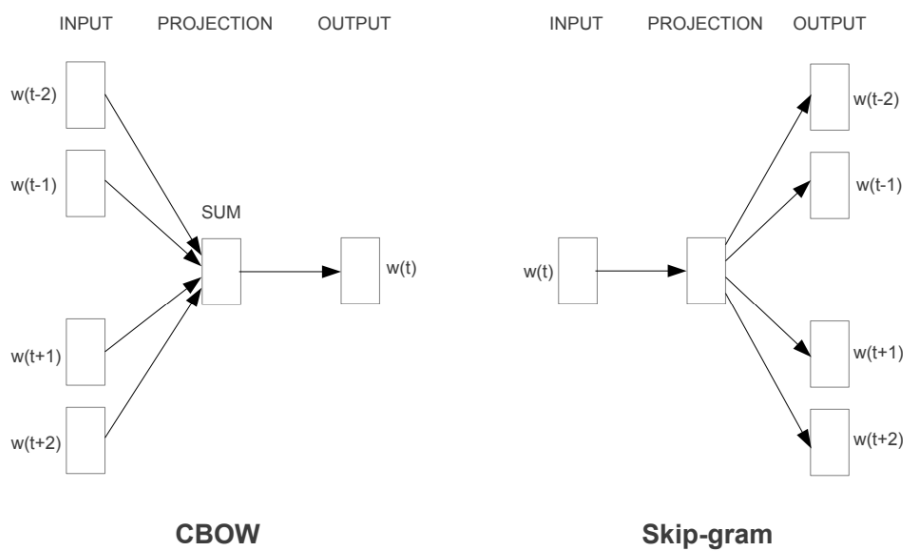
2.3. Vektori reči

Veliki broj algoritama mašinskog i dubokog učenja, poput neuronskih mreža, kao ulazne podatke očekuju numeričke vrednosti. Kako bi omogućili obradu teksta, uz pomoć algoritama koji zahtevaju numeričke

vrednosti, reči obrađivanog teksta se pretvaraju u vektore numeričkih vrednosti.

Jedan od najosnovnijih NLP pristupa za pretvaranje reči u vektore je uz pomoć binarnog kodiranja (eng. *one-hot encoding*). Binarno kodiranje reprezentuje reči sa vektorom čija je dužina jednaka dužini rečnika, gde na svim pozicijama tog vektora se nalazi vrednost nula, osim na poziciji posmatrane reči gde se nalazi vrednost jedan. Mana ovog pristupa je što ne obuhvata semantičko značenje reči, te nije moguće utvrditi sličnost sa vektorima sličnih reči (Nikola Nikolić, 2021). Druga mana ovog pristupa se ogleda u veličini vektora, na primer ukoliko se rečnik sastoji od 50 000 reči svaka reč će biti predstavljena vektorom iste dužine gde su vrednosti većinski nule.

Pristup koji se pokazao kao pristup sa boljim performansama je embeding (eng. *word embedding*) pristup. Vektori reči kod embeding pristupa su vektori realnih brojeva u višedimenzionom prostoru. Osnovna ideja embeding pristupa je da se prilikom preslikavanja reči na vektor koriste semantičke veze između reči na osnovu njihovih konteksta kako bi slične reči međusobno bile blizu u vektorskom prostoru (Jurafsky & Martin, 2020). Jedna od prednosti embeding pristupa, pored modelovanja semantičkih veza reči, u odnosu na vektor jedinice su manje dimenzije embeding vektora. Zbog manjeg broja dimenzija potrebno je obučiti manji broj parametara u krajnjim arhitekturama za rešavanje NLP zadataka, što može da dovede do boljih performansi.



Slika 7. Arhitekture Word2Vec modela.

Generisanje embeding vektora moguće je obučiti prilikom obučavanja modela, kao poseban ulazni sloj modela, ili obučiti kao nezavistan model za generisanja embeding vektora. Jedan od načina kako da se generišu embeding vektori, kao nezavistan model, uz pomoć neuronskih mreža je *Word2Vec* model predstavljen od strane autora (Mikolov et al., 2013). U navedenom radu autori su predstavili dve arhitekture za generisanje embedinga: *CBOW* model i *Skip-gram* model (Slika 7). CBOW model (eng. *Continuous Bag of Words* – CBOW) koristi kontekst, odnosno reči u neposredno okolini trenutne reči, kako bi izvršio predikciju trenutne reči. Za razliku od CBOW modela, *Skip-gram* model vrši predikciju konteksta na osnovu trenutne reči.

Prilikom obučavanja vektora reprezentacije reči, odnosno embedinga, obično je potreba velika količina teksta kako bi navedeni modeli mogli da se obuče. Obučavanje se vrši na nenadgledani način, odnosno pomoću samo učenjea (eng. *self-supervised*) jer svi podaci o kontekstu reči koji su potrebni modelu za obučavanje se nalaze u samim tekstovima sa kojima se model obučava. Nakon obučavanja, obučeni modeli se mogu iskoristiti u drugim arhitekturama kao sloj koji nije potrebno obučavati.

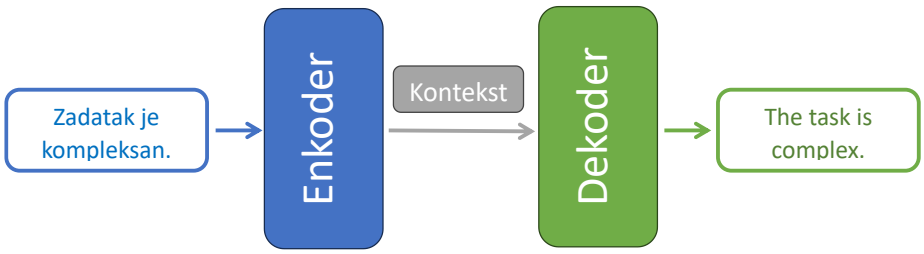
2.4. Transformer

Transformer arhitektura razvijena je od strane autora (Vaswani et al., 2017) za rešavanje problema modelovanja sekvence na sekvencu (eng. *sequence-to-sequence modeling)*, gde je njihov prvobitan cilj bio da poboljšaju performanse modela za mašinsko prevođenje teksta. Prethodno dominantni modeli za modelovanje sekvence na sekvencu su bili zasnovani na rekurentnim neuronskim mrežama i konvolutivnim neuronskim mrežama koje su organizovane u enkoder-dekoder arhitekturu (Vaswani et al., 2017).

2.4.1. Enkoder-dekoder arhitektura

Enkoder-dekoder arhitektura sastoji se iz dve mreže: enkoder i dekoder. Enkoder funkcioniše tako što se ulazna sekvenca, npr. rečenica na srpskom jeziku, propušta kroz enkoder koji generiše vektorsku reprezentaciju te rečenice (kontekst). Dekoder kao ulaz dobija kontekst, odnosno izlaz iz enkodera, i dekodira u izlaznu sekvencu, npr. rečenicu na engleskom jeziku. Bitno je napomenuti da se dužina sekvenci može razlikovati kao na slici 8.

Transformer arhitektura menja kompleksne neuronske mreže u enkoderu i dekoderu mehanizmom pažnje.



Slika 8. Jednostavan primer enkoder-dekoder arhitekture.

2.4.2. Mehanizam pažnje

Mehanizam pažnje je inicijalno razvijen za zadatak prevođenja teksta, koji se sa probabilističke perspektive može posmatrati kao pronalaženje tražene rečenice $\boldsymbol{y}$ koja maksimizuje uslovnu verovatnoću $\boldsymbol{y}$ na osnovu date izvorne rečenice $\boldsymbol{x}$, odnosno $argmax_y P(\boldsymbol{y} \mid \boldsymbol{x})$.

Inicijalni pristupi za rešavanje zadatak prevođenja teksta, upotrebom neuronskih mreža, su koristili RNN organizovane u enkoder-dekoder arhitekture. Tako organizovane neuronske mreže postizale su rezultate slične rezultatima do tada najboljih tradicionalnih rešenja za prevođenje teksta.

Jedan od nedostataka identifikovan kod inicijalnih pristupa se ogleda u upotrebi kontekstnog vektora fiksne dužine, u koji je potrebno enkodirati sve relevantne informacije potrebne za prevođenje izvorne rečenice. Fiksna dužina kontekstnog vektora je ograničila performanse prilikom prevođenja dužih rečenica, pogotovo kada se te rečenice bile duže od rečenica iz obučavajućeg skupa na kome je obučena arhitektura. Autori (Bahdanau et al., 2014) su predloži mehanizam pažnje u kojoj se rečenica enkodira u niz kontekstnih vektora, dok se prilikom dekodiranja adaptivno bira podskup vektora relevantan za trenutnu poziciju prilikom prevođenja. U predloženom modelu definišu uslovnu verovatnoću:

$$P(y_i|y_1, \dots, y_{i-1}, \boldsymbol{x}) = g(y_{i-1}, s_i, c_i), \quad (32)$$

gde je g nelinearna funkcija koja daje verovatnoću y_i, a s_i je skriveno stanje RNN mreže u i-tom koraku dato formulom $s_i = f(s_{i-1}, y_{i-1}, c_i)$. U formuli iznad verovatnoća svake reči y_i zavisi od različitog kontekstnog vektora c_i. Razlika od inicijalnog pristupa, u kome enkoder generiše jedan kontekstni vektor c na osnovu skrivenih stanja $c = q(\{h_1, \dots, h_{T_x}\})$, je u tome što se c_i računa kao težinska suma nad skrivenim težinama h_i:

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j, \quad (33)$$

gde se težina α_{ij} za svako h_j računa kao

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}, \quad (34)$$

pri čemu je

$$e_{ij} = a(s_{i-1}, h_j) \quad (35)$$

model poravnjanja (mehanizam pažnje) koji ocenjuje koliko dobro se ulazi oko pozicije j i izlazi oko pozicije i slažu. Ocena je bazirana na osnovu skrivenog stanja RNN s_{i-1} i skrivenog stanja h_j ulazne rečenice. U jednačini (35), a predstavlja neuronsku mrežu sa propagacijom unapred koja se obučava sa svim ostalim delovima predložene arhitekture.

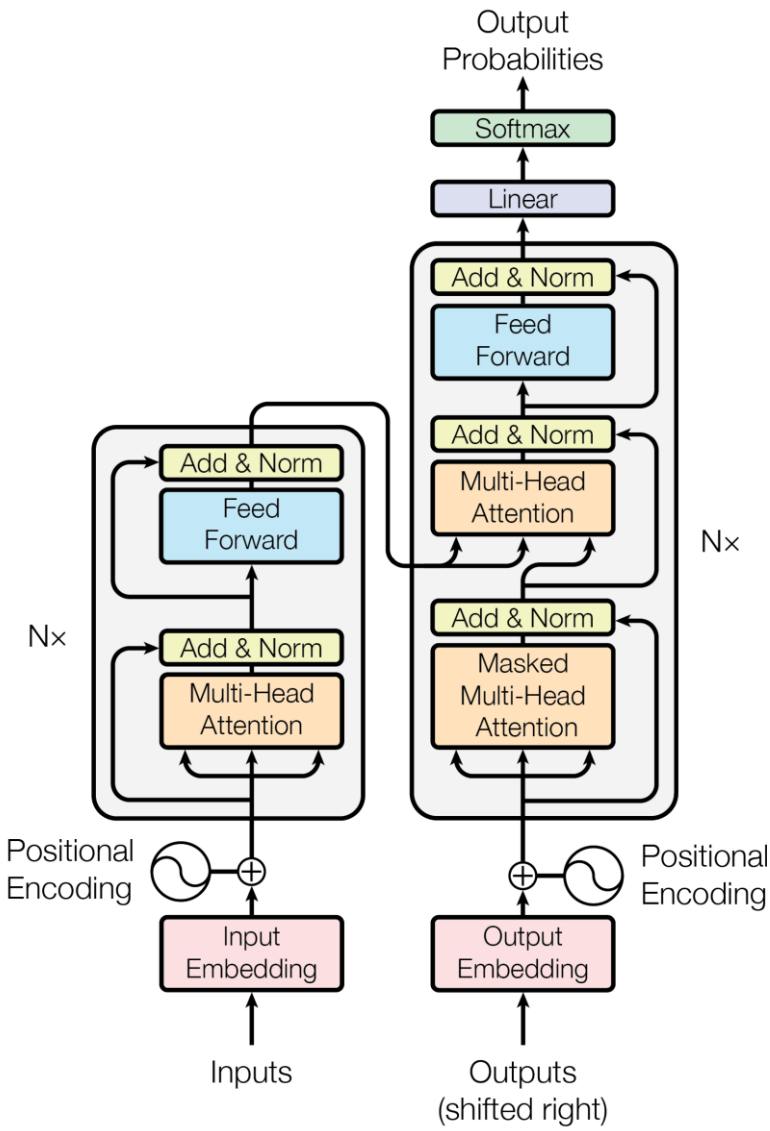
2.4.3. Arhitektura transformera

Arhitektura transformera, predstavljena u (Vaswani et al., 2017), data je na slici 9.

Sloj enkodera u prikazanoj arhitekturi sastoji se od dva podsloja. Prvi podsloj je mehanizam pažnje, a drugi podsloj je neuronska mreža sa propagacijom unapred. Oko svakog podsloja iskorišćenja je rezidualna konekcija (He et al., 2016) koja prosleđuje ulazne podatke u podsloj komponeti za normalizaciju sloja (Ba et al., 2016). Odnosno izlaz iz svakog podsloja (y) u arhitekturi je dat kao $y = NormSloj(x + PodSloj(x))$, gde x predstavlja ulazni vektor u podsloj.

Sloj dekodera se sastoji od tri podsloja. Na ulazu je podsloj sa mehanizmom pažnje kojem je, uz pomoć maskiranja, onemogućeno da se prilikom dekodiranja određene pozicije obraća pažnja na naredne pozicije. Odnosno, predikcija reči na poziciji *i* može da zavisi samo od dekodiranih

reči čija je pozicija manja od *i*. Na izlazu je podsloj neuronske mreže sa propagacijom unapred, a između ulaznog i izlaznog podsloja se nalazi podsloj sa mehanizmom pažnje koji za ulaze koristi izlaz iz enkoder komponente. Kao i u sloju dekodera sloju, između svakog podsloja su rezidualne konekcije i normalizacija sloja.



Slika 9. Arhitektura transformera (Vaswani et al., 2017).

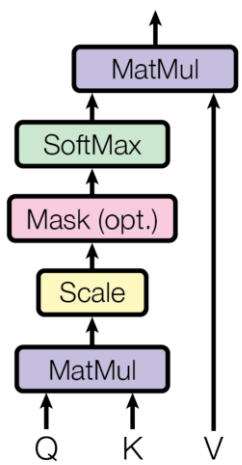
Enkoder i dekoder komponente transformera su formirane tako što su im slojevi ponovljeni *N* puta.

Autori (Vaswani et al., 2017) funkciju pažnje definišu kao preslikavanja upita (eng. *query*) i skupa ključ-vrednost (eng. *key-value*) parova na izlaz (eng. *output*). Gde su upit, ključevi, vrednosti, i izlaz vektori. Izlaz se računa kao težinska suma vrednosti, gde se težina koja se dodeljuje vrednostima računa pomoću funkcije kompatibilnosti upita sa odgovarajućim ključem.

Za funkciju pažnje uvode skalirani skalarni proizvod vektora (eng. *Scaled Dot-Prodcut Attention*, Slika 10) definisanu kao:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \qquad (36)$$

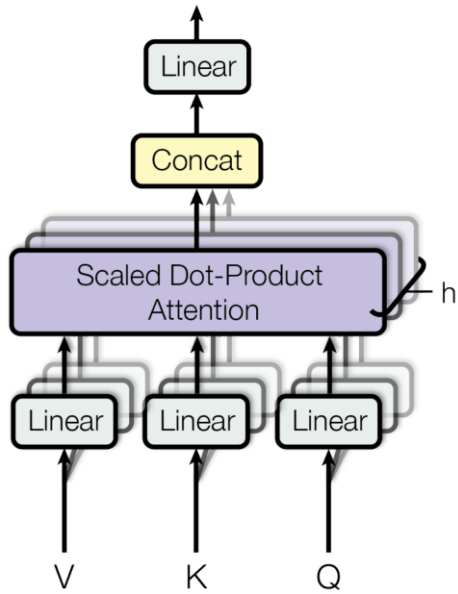
gde je Q matrica skupa upita za koji se računa pažnja, K matrica ključeva, V matrica vrednosti, d_k predstavlja dimenziju vektora upita i ključeva. Navedena funkcija pažnje je znatno brža od funkcije definisane od strane (Bahdanau et al., 2014), koja koristi neuronsku mrežu sa propagacijom unapred, a ujedno i rešava problem opadanja performansi standardnog skalarnog proizvoda vektora kada je vrednost parametra d_k velika.



Slika 10. Skalirani skalarni proizvod vektora (Vaswani et al., 2017)

Umesto da primenjuju jednu funkciju pažnje gde su dimenzije ključeva, vrednosti, i upita iste, (Vaswani et al., 2017) uvode mehanizam višestruke pažnje (eng. *Multi-Head Attention*) prikazan na slici 11. Mehanizam višestruke pažnje linearno projektuje upite, ključeve i vrednosti h puta sa različitim, naučenim, linearnim projekcijama na d_k dimenziju za upite i

ključeve, i na d_v dimenziju za vrednosti. Na projektovane verzije upita, ključeva i vrednosti primenjuju navedenu funkciju pažnje u paraleli čiji je rezultat vektor dimenzije d_v. Dobijeni vektori se konkateniraju nakon čega se opet vrši linearna projekcija, nakon čega se dobija izlazni vektor.



Slika 11. Mehanizam višestruke pažnje (Vaswani et al., 2017)

Formalno definisano višestruka pažnja se dobija po formuli:

$$MultiHead(Q,K,V) = Concat(head_1,\dots,head_h)W^O$$
$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (37)$$

Gde su linearne projekcije matrice $W_i^Q \in \mathbb{R}^{d_{model}\times d_k}$, $W_i^K \in \mathbb{R}^{d_{model}\times d_k}$, $W_i^V \in \mathbb{R}^{d_{model}\times d_v}$ i $W^O \in \mathbb{R}^{hd_v\times d_{model}}$.

Ulazni i izlazni tokeni se konvertuju u vektore dimenzije d_{model} uz pomoć embeding komponente. Kako bi model koristio podatke o poziciji reči u sekvenci, uzimajući u obzir da u sebi ne sadrži rekurentne slojeve, dodata je komponenta za enkodiranje pozicije (eng. *Positional encodings*) nad ulaznim vektorima. Enkodiranje poziciji se vrši tako što se ulazni vektori saberu sa vektorima koji su generisani uz pomoć sinusne i kosinusne funkcije:

$$PE_{(pos,2i)} = \sin\left(pos/10000^{2i/d_{model}}\right)$$
$$PE_{(pos,2i+1)} = \cos\left(pos/10000^{2i/d_{model}}\right) \quad (38)$$

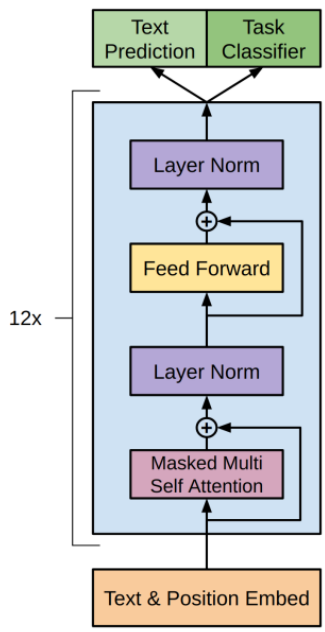
gde pos predstavlja poziciju, a i predstavlja dimenziju. Navedena funkcija je izabrana uz pretpostavku da će model moći lako da nauči da pruža pažnju na osnovu relativnih pozicija zbog činjenice da za bilo koji fiksni ofset k, PE_{pos+k} se može prestaviti kao linearna funkcija PE_{pos}.

Obučavanje transformer modela se vrši upotrebom tehnika za stohastičko određivanje gradijenta, gde su autori iskoristili Adam algoritam (Kingma & Ba, 2014) za optimizaciju.

2.4.4. Varijacije transformer arhitekture

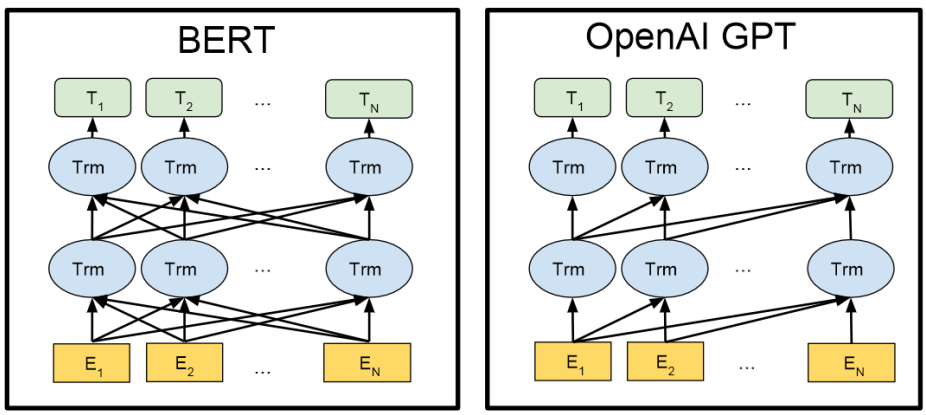
Osnovne varijace transformer arhitekture su transformer arhitektura zasnovana na dekoder sloju i transformer zanovana na enkoder sloju.

Transformer arhitektura zasnovana na dekoder sloju, predstavljena od strane autora (P. J. Liu et al., 2018), koristi samo dekoder slojeve originalne transformer arhitekture. Izbacivanje enkoder sloja autori su postigli bolje performanse na dugim ulaznim sekvencama za zadatak generisanja teksta u odnosu na RNN arhitekture i originalne transformer arhitekture. Transformer arhitektura zasnovana na dekoder sloju je auto-regresivna, odnosno prilikom generisanja sekvence koristi prethodno generisane tokene za generisanje sledećeg tokena u sekvenci.



Slika 12. Arhitektura GPT transformera (Radford et al., 2018)

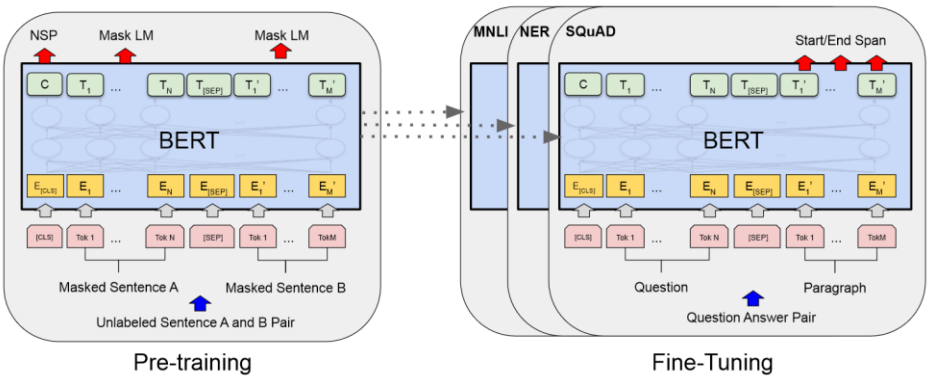
Autori (Radford et al., 2018), iskoristili su transformer arhitekturu zasnovnu na dekoder sloju (Slika 12) u kombinaciji sa generativnim pre-treniranjem modela (GPT). Predloženi model se prvo pre-trenira na velikim količinama teksta, na nenadgledan način (eng. *unsupervised pre-training*), za zadatak jezičkog modelovanja (eng. *language modeling*), odnosno za predviđanje sledeće reči na osnovu postojeće sekvence reči. Navedeno jezičko modelovanje omogućava modelu da nauči duboko razumevanje jezičke sintakse, semantike i konteksta. Nakon generativnog pre-treniranja, arhitektura može lako da se do-obuči na nadgledan način (eng. *supervised fine-tunig*) za specifične NLP zadatke, gde se obučavaju samo poslednji slojevi arhitekture (Slika 12. *Text Prediction* i *Text Classifier*).



Slika 13. Razlika između BERT i GPT arhitekture za pre-treniranje (Devlin et al., 2018)

Transformer arhitektura zasnovana na enkoder sloju je iskorišćena od strane autora (Devlin et al., 2018) za formiranje modela bidirekcione enkoder reprezentacije za tranformere (eng. *Bidirectional Encoder Representations from Transformers*, BERT). Kao i GPT model i BERT model se prvo pre-trenira na velikim količinama teksta, nakon čega se do-obučava za specifične NLP zadatke. Osnovna razlika između GPT i BERT arhitekture (Slika 13) je u tome što GPT koristi takozvanu sleva-nadesno arhitekturu, a BERT koristi bidirekcionu arhitekturu. U levo-ka-desno arhitekturu mehanizam pažnje može samo da koristi kontekst tokena koji se nalaze sa njegove leve strane u sekvenci, odnosno samo prethodne tokene, dok u bidirekcionoj arhitekturi mehanizam pažnje koristi kontekst sa leve i desne strane tokena. Dodatno, duboke bidirekcione reprezentacije

teksta u BERT modelu su postignute tako što se u postupku pre-treniranja koristi zadatak maskiranog jezičkog modelovanja (eng. *masked language model*, MLM). MLM nasumično maskira ulazne tokene a zadatak modela je da na osnovu konteksta odredi koji su originalni tokeni koji su bili maskirani. Dodatno, uz MLM, BERT koristi i zadatak predviđanja sledeće rečenice (eng. *next sentence prediction*). Osim izlaznih slojeva, BERT arhitektura (Slika 14) je ista za zadatak pre-treniranja i do-obučavanja.



Slika 14. Arhitektura BERT modela (Devlin et al., 2018)

Bidirekciona priroda BERT modela omogućava mu da bolje razume kompleksne zavisnosti prilikom procesiranja rečenice. Dodatno mnogi NLP zadaci, poput odgovaranja na pitanja (eng. *question answering*), zavise od razumevanja zavisnosti između dve rečenice što je omogućeno BERT modelu da modeluje na osnovu zadatka predviđanja sledeće rečenice. Navedene prednosti su omogućile BERT modelu da nadmaši performanse do tada najboljih modela na jedanaest različitih NLP zadataka.

Pored varijacije u arhitekturi modeli se razlikuju i po tokenizatoru koju koriste prilikom obrade ulaznih reči (tokena) u model. BERT model koristi *WordPiece* tokenizator (Y. Wu et al., 2016), dok GPT model koristi *Byte-Pair Encoding* (BPE) tokenizator (Sennrich et al., 2016).

Oba tokenizatora su obučena sa sličnim ciljem, odnosno počinju sa početnim rečnikom (koji je obično sačinjen od azbučnih slova i određenih specijalnih karaktera, npr. interpunkcijskih znakova) i cilj im je da nauče pravila spajanja, reči iz rečnika, tako da se proširi rečnik do željene veličine sa najčešćim pod-rečima (pod-tokenima). Inicijalno sve reči bi bile predstavljene kao kombinacija slova u reči, tako da bi reč „nefrotoksičan“ inicijalno bila predstavljena kao kombinacija od 13 pod-

tokena. Obučavanjem, tokenizator će proširivati rečnik sa najčešćim kombinacijama i njih dodavati u rečnik. Tako da u krajnje obučenom tokenizator reč „nefrotoksičan“ može da predstavljena sa npr. tri pod-tokena „nefro“, „toski“ i „čan“.

Osnova razlika između BPE i *WordPiece* tokenizatora je što BPE spaja najčešći par pod-tokena kao i njega dodaje u rečnik kao novi pod-token, a *WordPiece* spaja par sa najboljim rangom. Rangiranje para pod-tokena, u *WordPiece* tokenizatoru, se vrši tako što se računa odnos frekvencija para pod-tokena podeljene sa proizvodom frekvencije pojedinačnih pod-tokena. Rangiranjem na osnovu frekvencija *WordPiece* tokenizator daje veći prioritet prilikom spajanja para pod-tokena gde su individualni pod-tokeni ređi u korpusu nad kojim se obučava.

2.5. Modeli zasnovani na ansamblima

Modeli zasnovani na ansamblima predstavljaju metode koje kombinuju više baznih modela kako bi se napravio konačan model koji daje preciznije rezultate. Ideja iza sistema baziranih na ansamblu se obično posmatra kao implementacija koncepta mudrosti mase (eng. *wisdom of the crowd*), gde kombinovanjem više različitih mišljenja dolazimo do odluke koja je bolja od konsultovanje jedne individue (Polikar, 2006; Sagi & Rokach, 2018).

Neki od razloga zbog kojih modeli zasnovani na ansamblu često rezultuju sa poboljšanjem performansi krajnjeg modela su (Ganaie et al., 2022; Sagi & Rokach, 2018):

- Sprečavanje preobučavanja (eng. *overfitting aviodance*): Sa malom količinom obučavajućih podataka obučavajući algoritam je sklon pronalaženju hipoteza sa kojima su predikcije uvek tačne nad obučavajućim skupom dok za neviđene[7] podatke pravi jako loše predikcije. Kombinovanjem modela sa različitim hipotezama smanjuje se verovatnoća da je izabran pogrešan klasifikator.
- Izbegavanje lokalnog optimuma: Jedan model može da se zaglavi u lokalnom optimumu, kombinovanjem više modela smanjuje se rizik lokalnog optimuma.

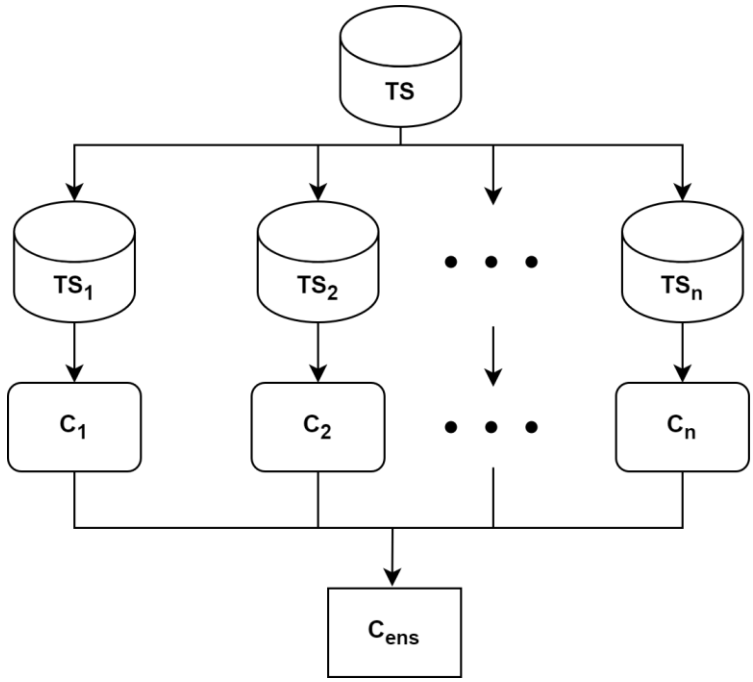
[7] Obučavanje modela se obično vrši uz pomoć obučavajućeg skupa i test skupa. Model u toku obučavanje koristi podatke iz obučavajućeg skupa, dok krajnja validacija modela se vrši nad podacima iz test skupa. Odnosno, validacija se vrši nad podacima koje model nije imao priliku da „vidi“ prilikom procesa obučavanja.

- Reprezentacija: Optimalna hipoteza može biti van opsega hipoteza individualnih modela. Kombinovanje hipoteza od više različitih modela može da rezultuje sa optimalnom hipotezom.

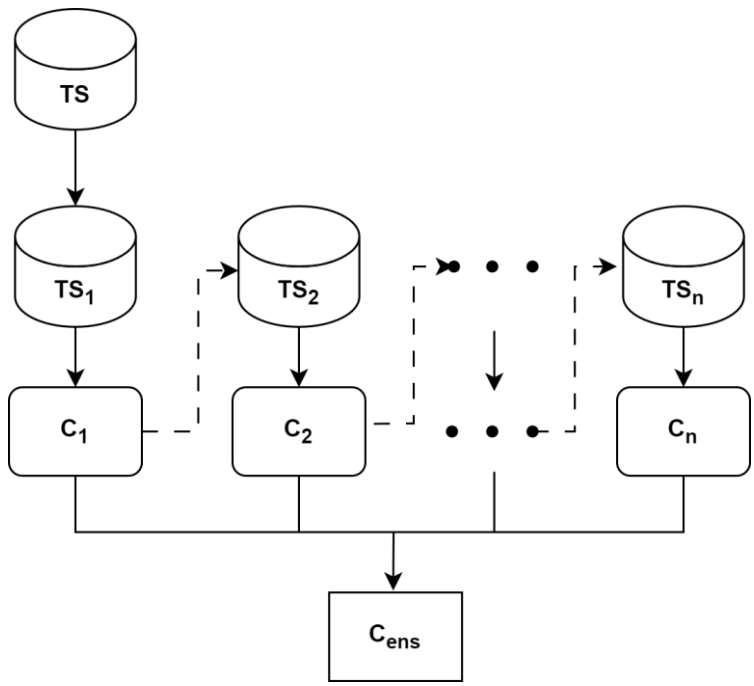
Postoje razne strategije za kombinovanje više baznih modela u ansambl od kojih su najpoznatije: prosta agregacija (eng. *bagging*), pojačavanje (*boosting*), slaganje (*stacking*), i agregacija obučenih modela (*after the fact ensamble*).

*S*trategija proste agregacije (Breiman, 1996) generiše više verzija jednog tipa klasifikatora, koje kombinuje u anasambl (Slika 15). Različiti klasifikatori se dobijaju tako što se obučavaju na različitim grupama obučavajućih podskupova. Grupe obučavajućih podskupova (eng. *bags*) se generišu tako što se podaci iz obučavajućeg skupa uzorkuju na nasumičan način sa ponavljanjem. Kako bi se obezbedilo dovoljno primera podataka, po obučavanom klasifikatoru, svaka grupa obučavajućeg podskupa sadrži isti broj podataka kao i obučavajući skup. Konačno obučeni klasifikatori se na kraju kombinuju običnom primenom strategije većinskog glasanja, odnosno konačna klasa je klasa za koju je glasao najveći broj klasifikatora.

Kao i strategija proste agregacije, strategija pojačavanja od slabog klasifikatora, klasifikator čije je preciznost malo bolja od nasumičnog pogađanja, generiše model koji ima dobru sposobnost generalizacije (Slika 16). Strategija pojačavanja generiše ansambl klasifikatora tako što svaki prethodni klasifikator utiče na obučavanje sledećeg klasifikatora u konačnom modelu. Odnosno, prilikom obučavanja klasifikatora generiše se obučavajući podskup u kome su prioritetni primeri oni na kojima je prethodni klasifikator pravio greške. Određivanje prioriteta primera se vrši tako što su svim primerima u obučavajućem skupu dodeljeni težinski faktori, koji inicijalno imaju istu težinu. Pri svakoj iteraciji se gleda gde je trenutni klasifikator grešio i tim primerima se povećavaju težinski faktori, dok se na primerima koji su tačno klasifikovani smanjuju težinski faktori. Pored težinskih faktora primera, svakom klasifikatoru se takođe dodeljuje težinski faktor na osnovu njihovih performansi. Konačno kombinovanje klasifikatora se vrši tako što se kombinuju rezultati individualnih klasifikatora na osnovu njihovih težinskih vrednosti.

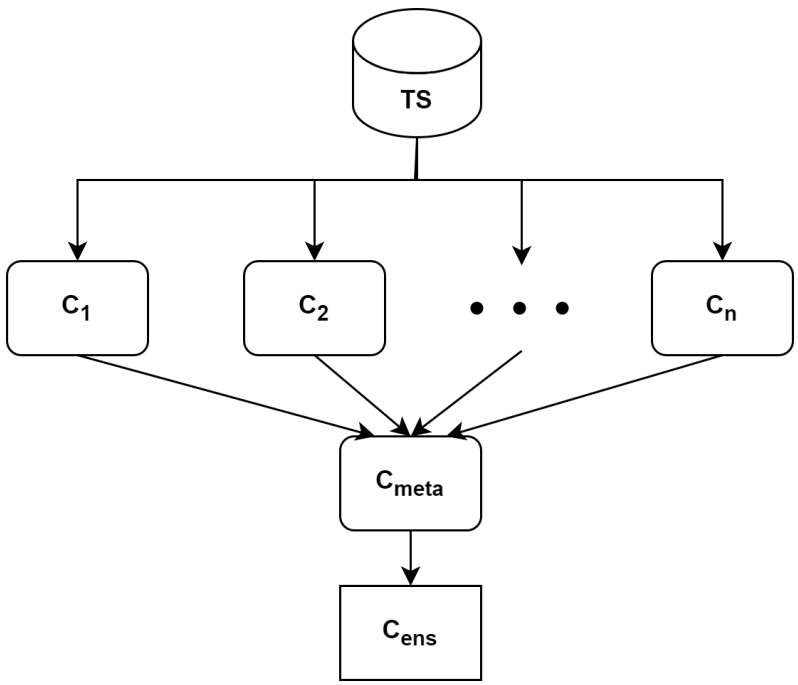


Slika 15. Ansambl proste agregacije



Slika 16. Ansambl pojačavanja

Strategija slaganja (Wolpert, 1992) osnovne klasifikatore kombinuje uz pomoć meta-klasifikatora (Slika 17). Ideja iza strategije slaganja je da se kombinuju osnovni klasifikatori sa novim klasifikatorom koji će imati sposobnost da prepozna problematične tipove klasa iz obučavajućeg skupa za svaki od osnovnih klasifikatora koje kombinuje. Odnosno, izlazi iz osnovnih klasifikatora predstavljaju ulaze u takozvani meta-klasifikator, čiji je zadatak da nauči veze između izlaza klasifikatora i klasa koje oni tačno klasifikuju. Obučavanje klasifikatora u okviru strategiju slaganja se obično vrši sa k-strukom unakrsnom validacijom[8], gde se obučavajući skup podeli na k jednakih podskupova i svaki osnovni klasifikator se obučava sa $k-1$ različitih podskupova.



Slika 17. Ansambl slaganja

Agregacija obučenih modela (eng. *after the fact ensamble*) se može posmatrati kao najosnovnija tehnika za formiranje ansambl modela, gde

[8] Unakrsna validacija je tehnika za proveru generalizacionih sposobnosti statističkih modela nad neviđenim podacima. Ukoliko je obučavajući skup previše mali da se podeli na obučavajući i test skup, ili ukoliko je potrebno mali obučavajući skup podeliti na obučavajući i validacioni može se iskoristi k-unakrsna validacija. K-struka unakrsna validacija (eng. *K-fold cross-validation*) deli skup (obučavajući) podataka na *k* jednakih podskupova. Model se obučava sa *k*-1 podskupova a testira (ili evaluira) na jednom izostavljenom podskupu.

su individualni modeli obučeni nezavisno jedni od drugih. Njihova kombinacija može da se vrši uz pomoć meta-klasifikatora ili neke od strategije za kombinaciju kao što je strategija većinskog glasanja.

2.6. Metode evaluacije modela

U okviru oblasti mašinskog učenja razvijen je veliki broj različitih algoritama za rešavanje različitih tipova problema. Kada je reč o nadgledanom učenju (eng. *supervised learning*) algoritam se obučava uz pomoć obeleženog skupa podataka, odnosno poznati su nam parovi ulaznih vrednosti i očekivanih izlaza iz modela (predikcija). Pre postupka obučavanja obeleženi skup podataka se deli na obučavajući skup (eng. *training set*), test skup (eng. *test set*) i ukoliko postoji dovoljno raspoloživih podataka na validacioni skup (eng. *validation set*). Obučavajući skup se koristi u procesu obučavanja, kako bi se parametri modela podesili na vrednosti koji omogućavaju da model ima dobre prediktivne sposobnosti za rešavanje problema za koji se obučava. Validacioni skup se koristi za praćenje performansi modela tokom njegovog obučavanja, kao i za podešavanja hiperparametara (parametri modela koji se ne podešavaju u toku obučavanja, npr. broj skrivenih slojeva u neuronskoj mreži) modela. Test skup je skup podataka koji se koristi za konačnu evaluaciju modela. Podaci koji se nalaze u test skupu se obično odvajaju iz skupa podataka pre samog postupka izbora algoritma i obučavanja modela. Kako bi konačna evaluacija bila objektivna test skup ne sme da se koristi u procesu izbora algoritma, obučavanja i podešavanja hiperparametara.

Evaluacija klasifikacionih modela se vrši uz pomoć test skupa, gde se porede stvarni i očekivani izlazi iz modela. Ukoliko je reč o višeklasnoj klasifikaciji, rezultate klasifikacije određene klase možemo posmatrati kao binarnu klasifikaciju gde su elementi izabrane klase posmatrani kao pozitivni primeri, a elementi ostalih klasa klasifikovani kao negativni primeri. Tako dobijene rezultate klasifikacije možemo svrstati u četiri kategorije: tačno pozitivno (eng. *true positive*, tp), lažno pozitivno (eng. *false positive*, lp), tačno negativno (eng. *true negative*, tn) i lažno negativno (eng. *false negative*, ln). Najosnovnija mera za evaluaciju tačnosti klasifikacionog modela je tačnost (eng. *accuracy*). Tačnost se dobija kao procenat tačnih predikcija od svih načinjenih predikcija, odnosno:

$$Tačnost = \frac{tp + tn}{tp + lp + tn + ln} \quad (39)$$

Jedan od osnovnih problema sa merom tačnosti je što ne prezentuje prave performanse kada u skupu podataka klase nisu ravnomerno zastupljene. Na primer, ukoliko se radi o binarnoj klasifikaciji kada želimo primere iz obučavajućeg skupa da klasifikujmo u dve klase *pozitivan* ili *negativan*, npr. određivanje da li je pacijent oboleo od hronične bubrežne insuficijencije. Ukoliko u test skupu imamo ukupno 50 pacijenta koji su oboleli od bolesti a 950 koji nisu, ako napišemo naivni klasifikator koji svakom primeru dodeljuje klasu *negativan* njegova tačnost će biti 95%. Tačnost od 95% u kome model nije tačno klasifikovao nijednu vrednost klase *pozitivan* ne predstavlja realističnu procenu kvaliteta modela. Da bi se dobila realističnija procena kvaliteta modela, pogotovo kada skup podataka nije ravnomerno raspodeljen po klasama, što je veoma čest slučaj prilikom prepoznavanja imenovanih entiteta, preporučuje se upotreba preciznosti (eng. *precision*) i odziva (eng. *recall*) (Slika 18).

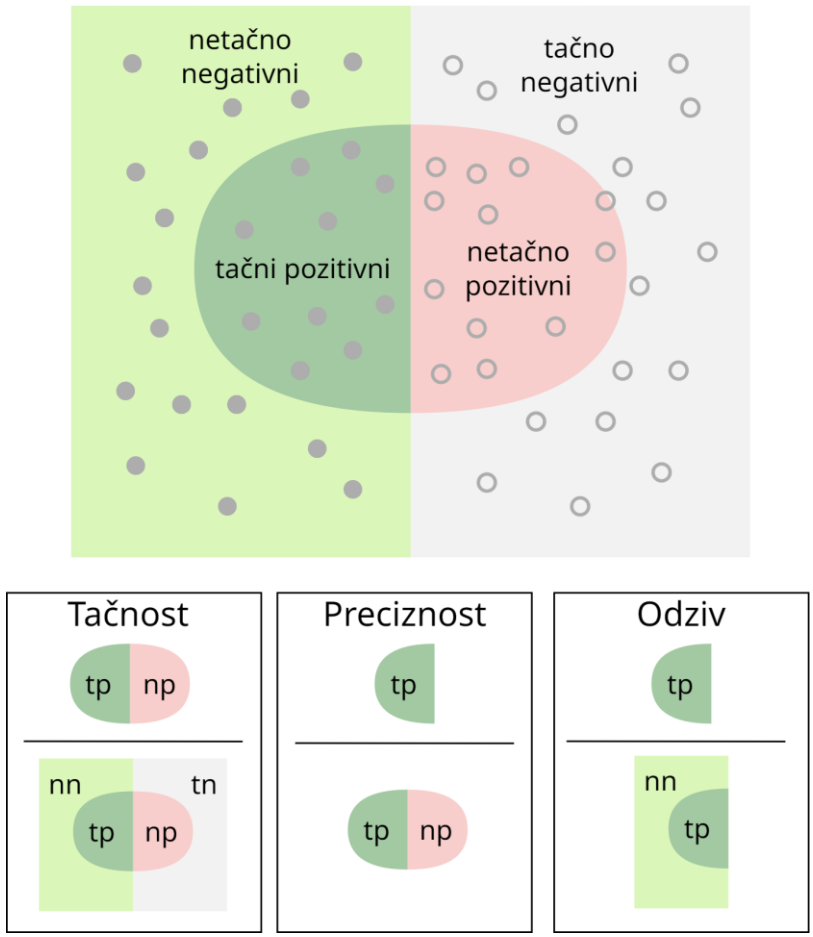
Mera preciznost je definisana kao procenat tačno pozitivnih primera klasa od ukupnog broja primera klasifikovanih kao pozitivni:

$$Preciznost = \frac{tp}{tp + np} \quad (40)$$

Odziv je mera koja je definisana kao procenat tačno klasifikovanih pozitivnih primera klasa od svih pozitivnih primera u skupu podataka:

$$Odziv = \frac{tp}{tp + nn} \quad (41)$$

U primeru iznad, kada klasifikator sve klasifikuje kao *negativne* primere, i preciznost i odziv će imati vrednost 0/0. Ukoliko se izmeni primer da se dobije realističniji scenario tako da 30 od 50 pozitivnih primera je tačno klasifikovano, a 900 od 950 negativnih primera je klasifikovano kao negativno. U takvom scenariju dobijamo 93% kao vrednosti za tačnost, dok za preciznost dobijamo vrednost 37,5% a za odziv dobijamo 60%. Iz navedenog se vidi da preciznost i odziv mnogo bolje predstavljaju klasifikacionu sposobnost iskorišćenog klasifikatora.



Slika 18. Tačnost, preciznost i odziv

Postoje različiti načini za kombinaciju preciznosti i odziva u jednu vrednost kako bi se različiti klasifikacioni mogli lakše uporediti, od kojih je najzastupljenija F-mera. F mera je definisana kao:

$$F_{mera} = (1 + \beta^2) \cdot \frac{preciznost \cdot odziv}{(\beta^2 \cdot preciznost) + odziv}, \qquad (42)$$

gde je β težinski parametar koji određuje koliko će odziv biti važniji od preciznosti. U praksi se najčešće koristi takozvana F1 mera, gde je vrednost parametra β jednaka broju 1, koja predstavlja harmonijsku sredinu preciznosti i odziva.

Kao što je napomenuto, kada se radi o višeklasnoj klasifikaciji računanje metrika preciznosti, odziva i F1 mere se vrši za svaku klasu pojedinačno.

Kako bi odredili objedinjene performanse modela potrebno je izračunati prosek metrika. U praksi se često koriste tri strategije za računanje proseka: makro prosek, ponderisani prosek i mikro prosek. U tabeli 1. je predstavljen primer predikcija sa tri klase (EVENT, TIMEX3. VALUE).

Klasa	**tp**	**lp**	**ln**	**Preciznost**	**Odziv**	**F1**
EVENT	5	3	3	0.63	0.63	0.63
TIMEX3	1	3	0	0.25	1.00	0.40
VALUE	8	0	8	1.00	0.5	0.67

Tabela 1. Primer predikcija za tri klase

Najjednostavnija strategija za računanje proseka je makro prosek, gde se vrednost, za svaku od metrika, dobijaju kao prosečna vrednost nad individualnim klasama. Na osnovu primera u tabeli iznad, za F1 meru makro prosek se dobija kao $((0.63 + 0.4 + 0.67)/3)$, odnosno 0.57.

Ponderisani prosek je prosek koji uzima u obzir proporciju stvarne zastupljenosti (tačno pozitivni + lažno negativni) svake od klasa u skupu podataka. U primeru datom u tabeli imamo 32% primera EVENT klase, 4% primera TIMEX3 klase i 64% primera VALUE klase. Ponderisani prosek za F1 meru se dobija kao $(0.63 \cdot 0.32 + 0.4 \cdot 0.04 + 0.67 \cdot 0.64)$, odnosno 0.65.

Mikro prosek računa globalan prosek tako što računa vrednosti za sve tačno pozitivne, lažno pozitivne i lažno negativne primere. U tabeli iznad imamo 14 tačno pozitivnih primera, 6 lažno pozitivnih primera i 11 lažno negativnih primera. Tako da mikro prosek za preciznost, odziv i F1 meru će biti respektivno 0.70, 0.56 i 0.62.

2.6.1. Metode evaluacije anotacija

Obučavanje algoritama mašinskog učenja, nadgledanim učenjem, zahteva postojanje obučavajućeg i test skupa podataka. Odnosno zahteva prikupljanje ulaznih podataka i njihovo obeležavanje (anotiranje) odgovarajućim klasama za klasifikacione probleme (ili konkretnim numeričkim vrednostima za probleme regresije). Prikupljanje i anotiranje podataka je proces u kome ljudi, obično stručnjaci iz oblasti problema, obeležavaju prikupljene podatke odgovarajućim klasama. Kvalitet anotacija (saglasnost anotatora, *inter-annotator agreement*, IAA), između ostalog, zavisi od preciznosti anotacinih uputstava i obučenosti anotatora.

Postoje različite mere za ocenu kvaliteta anotacija poput mere *Cohen's Kappa* (Cohen, 1960; Sim & Wright, 2005) za ocenu saglasnosti dva anotatora, mera *Fleiss Kappa* (Fleiss, 1971; Fleiss et al., 2013) koja se koristi za ocenu saglasnosti više od dva autora, F1 mera koja može da se koristi za ocenu saglasnosti dva ili više anotatora. U okviru ove disertacije korišćena je F1 mera kao i mera *Cohen's Kappa* defiisana kao:

$$K = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}, \qquad (43)$$

gde je $\Pr(a)$ je procentualno slaganje anotatora (od ukupnog broja anotiranih primera koliko se tačno poklapa), a $\Pr(e)$ je očekivana saglasnost između anotatora ukoliko bi svaki od njih nasumično izabrao klasu za svaku anotaciju (Pustejovsky & Stubbs, 2012). Očekivana saglasnost anotatora se računa tako što se za svaku klasu prvo izračuna koliko je verovatno da oba anotatora nasumice izaberu tu klasu. Verovatnoća da anotatori nasumice izaberu istu klasu se računa kao proizvod procentualne vrednosti koliko često su anotatori anotirali tu klasu tokom procesa anotiranja. Konačna vrednost očekivane saglasnosti je suma prethodno izračunatih verovatnoća za sve klase.

Primer: Neka su anotatori A i B imali zadatak da obeleže 200 dokumenata sa klasama *pozitivan* i *negativan*, i neka su ih obeležili kao u tabeli ispod:

	$A_{pozitivan}$	$A_{negativan}$
$B_{pozitivan}$	104	35
$B_{negativan}$	11	50

Tabela 2. Primer računanja Kappa mere

Anotatori A i B su obeležili iste primere sa klasom *pozitivan* 104 puta i 50 puta sa klasom *negativan*, $\Pr(a) = \frac{104+50}{200} = 0{,}77$. Anotator A je obeležio klasu *pozitivan* 115 (104+11) puta dok je anotator B obeležio klasu *pozitivan* 139 (104 + 35) puta, te dobijmo nasumičnu verovatnoću klase *pozitivan* kao proizvod procentualnog korišćenja klase: $\frac{115}{200} \cdot \frac{139}{200} = 0.575 \cdot 0.695 \approx 0.40$. Na isti način dobijamo za klasu *negativan*: $\frac{85}{200} \cdot \frac{61}{200} = 0.425 \cdot 0.305 \approx 0.13$. Konačna vrednost za $\Pr(e)$ je 0.53 $(0.40 + 0.13)$. Čime dobijamo vrednost *kappe* $K = 0.51$. Opšte smernice za tumačenje *kappa* vrednosti, definisane od strane autora (Landis & Koch, 1977), dati su u tabeli 3.

k	Nivo slaganja
< 0	Loš
0.01-0.20	Nezadovoljavajući
0.21-0.40	Slab
0.41-0.60	Umeren
0.61-0.80	Znatan
0.81-1.00	Savršen

Tabela 3. Smernice za tumačenje *kappa* vrednosti

2.7. Obeležavanje imenovanih entiteta

U procesu anotiranja svaki imenovani entitet će biti obeležen jednom od klasa od interesa, gde jedan imenovani entitet može da se sastoji od više reči (tokena). Prilikom prepoznavanja imenovanih entiteta klasifikacioni modeli bi trebalo da se obuče da prepoznaju sve tokene koje pripadaju imenovanom entitetu. Iz tog razloga, potrebno je sve tokene u tekstu obeležiti na način koji jasno definiše početak i kraj tokena.

Postoje razni formati za obeležavanje imenovanih entiteta, od koji je jedan od najpoznatijih i najčešće korišćenih *IOB2* format (Ratnaparkhi, 1998). *IOB2* format je proširenje *IOB1* (*BIO*) formata (Ramshaw & Marcus, 1995), u kome se tokeni koji pripadaju imenovanom entitetu obeležavaju jednim od prefiksa:

- I (*inside*) – token pripada imenovanom entitetu,
- O (*outside)* – token ne pripada imenovanom entitetu,
- B (*begin*) – imenovani entitet počinje sa tokenom.

Razlika između *IOB1* i *IOB2* formata je u tome što token dobija prefiks B samo ukoliko je token pre njega pripadao istoj klasi, dok u *IOB2* formatu svi tokeni koji predstavljaju početak imenovanog entiteta se obeležavaju prefiksom B (Sang & Veenstra, 1999). Primer entiteta *IOB2* za rečenicu „Mihajlo Idvorski Pupin rođen je 9. oktobra 1854. u Idvoru.“ u kojoj prepoznajemo klase OSOBA, DATUM i NASELJE je data u tabeli 4.

Prilikom evaluacija klasifikacionih modela (sekcija 2.6) u medicinskom domenu se obično koriste dve strategije: poklapanje na nivou tokena i tačno poklapanje imenovanih entiteta.

Evaluacija na nivou tokena se vrši na nivou individualnih tokena koji pripadaju imenovanim entitetima. Tako da se tokeni smatraju jednakim, odnosno za klasu dobijamo tačno pozitivan rezultata, ukoliko je klasa (*IOB2* formatu) koja je dodeljena tokenu ista kao i anotirana klasa za taj token.

Token	**IOB2 format**
Mihajlo	B-OSOBA
Idvorski	I-OSOBA
Pupin	I-OSOBA
Rođen	O
Je	O
9	B-DATUM
.	I-DATUM
oktobra	I-DATUM
1854	I-DATUM
.	I-DATUM
U	O
Idvoru	B-NASELJE
.	O

Tabela 4. Primer IOB2 formata

Tačno poklapanje entiteta je jedna od četiri strategije evaluacije definisane u SemEval NER zadatku iz 2013. godine (Segura-Bedmar et al., 2013). Evaluacija tačnog poklapanja entiteta smatra da su dva entiteta jednaka ukoliko imaju identične granice entiteta i isti tip entiteta. Odnosno, da bi se obeleženi entitet računao kao tačno pozitivan, svi tokeni u entitetu moraju biti isti kao i tokeni u entitetu sa kojim se poredi. Ostale tri strategije su evaluacija tačnog poklapanja granice entiteta (bez obzira na dodeljen tip), delimično poklapanje entiteta bez obzira na tip, i delimično poklapanje gde entiteti moraju imati dodeljen i isti tip.

Evaluacija klasifikacionih modela pomoću pristupa za tačno poklapanje entiteta je mnoga stroža i teža za postići od evaluacije na nivou tokena. Kao primer poređenja ova dva pristupa može se iskoristi entitet DATUM koji je prikazan u tabeli 4 i koji se sastoji od 5 tokena. Ukoliko je klasifikator pogrešno klasifikovao jedan od tih pet tokena, za poklapanje na nivou tokena imaćemo 4 tačno pozitivna rezultata i jedan lažno negativan rezultat, dok za tačno poklapanje entiteta imaćemo samo jedan lažno negativan rezultat.

Poklapanje na nivou tokena može se koristi kao mera za opštu evaluaciju performansi modela. Na primer, ukoliko su za neku od klasa dobijeni dobri rezultate na nivou tokena, a loši rezultati za tačno poklapanje entiteta, to može biti indikacija da model ima problema samo sa određenim tipom tokena koji pripada entitetu. U tom slučaju, potrebno je pronaći dodatno rešenja u vidu izmene samog modela ili dodatni pretprocesiranjem ili postprocesiranjem. Sa druge strane ukoliko za određenu klasu su dobijeni loši rezultati i na nivou tokena onda je potrebno analizirati celokupnu klasu entiteta kako bi se uvideli razlozi za loše performanse nad tom klasom.

3. Pregled aktuelnog stanja u oblasti

U ovom poglavlju biće dat pregled postojećih pristupa za prepoznavanje imenovanih entiteta sa fokusom na pristupe u medicinskom domenu. Prvo će biti dat kratak istorijski pregled razvoja metoda za prepoznavanje imenovanih entiteta u opštem domenu. Nakon toga biće opisane metode za prepoznavanje imenovanih entiteta u medicinskom domenu, praćeno sa opisom metoda primenjivanih za prepoznavanje imenovanih entiteta na srpskom jeziku.

3.1. Prepoznavanje imenovanih entiteta

Osnovni i veoma značajan korak, za analiziranje nestruktuiranih tekstualnih podataka u elektronskim zdravstvenim kartonima, je prepoznavanje imenovanih entiteta. U sklopu konferencija za razumevanje poruka (eng. *Message Understanding Conferences*, MUC), organizovanih između 1987. i 1998. godine, razvijene su prve metode za NER. Termin „Imenovani entiteti“ prvi put je upotrebljen u sklopu šeste MUC konferencije (Grishman & Sundheim, 1996), koji je korišćen za naziv zadatka za prepoznavanje naziva ljudi, organizacija, geografskih lokacija kao i vremena, valuta i procenata u novinskim člancima (Goyal et al., 2018; J. Li et al., 2020).

Rani pristupi za prepoznavanje imenovanih entiteta su većinski koristili metode zasnovane na pravilima. NER sistemi zasnovani na pravilima se sastoje od skupa ručno definisanih semantičkih i sintaksnih pravila za prepoznavanje imenovanih entiteta, domenski specifičnih rečnika za identifikaciju klasa imenovanih entiteta i modula za izdvajanje koji procesira tekst upotrebom pravila i rečnika (Mohit, 2014). Poznati sistemi poput LaSIE-II (Humphreys et al., 1998), NetOwl (Krupka & Hausman, 1998) su prepoznavali imenovane entitet prvo pokušavajući da nađu poklapanje u rečnicima i onda primenom pravila, poput kontekstnih pravila gde određene reči aktiviraju pravilo (npr. „Gulf“ ili „Mountain“ za identifikaciju mesta). Rani sistemi su ostvarili relativno dobre performanse u vidu visoke preciznosti (eng. *precision*) ali i slabog odziva (eng. *recall*) obično zbog domenski-specifičnih pravila i nekompletnih rečnika (J. Li et al., 2020).

Iako sistemi zasnovani na pravilima mogu da ostvare dobre rezultate, kada imaju adekvatno definisana pravila i opširne rečnike, imali su ograničenu

upotrebu. Najveće mane koje su uticale na njihovu ograničenu upotrebu su njihova domenska specifičnost, gde se pravila i rečnici definišu za jedan domen te ih nije lako prilagoditi drugim domenima, i potreba da eksperti definišu i održavaju pravila što ih čini veoma skupim za razvoj. Zbog navedenih mana sistema baziranih na pravilima, dalja istraživanja su bila fokusirana na primenu mašinskog učenja za NER.

Pojava javno dostupnih korpusa podataka, kao i napredak u oblasti mašinskog učenja, značajno je doprinela razviju automatskog prepoznavanja imenovanih entiteta (Aleksandar Kovačević, 2011; Mohit, 2014).

Sistemi bazirani na metodama mašinskog učenja se mogu podeliti na metode zasnovane na nadgledanim učenjem i metode zasnovane na nenadgledanim učenjem. Za NER se najčešće koriste metode sa nadgledanim učenjem, koje se sastoje iz obučavajućeg skupa podataka, skupa atributa i algoritma za učenje. Obučavajući skup je kolekcija anotiranih dokumenata u kome su reči koji pripadaju imenovanim entitetima anotirane, odnosno obeležene sa odgovarajućom klasom entiteta. Atributi (eng. *feature*) predstavljaju podatke koji su značajni za predikciju (Marković, 2017), u kontekstu NLP-a to mogu biti osobine reči poput upotrebe velikih slova, vrste reči i sl. (J. Li et al., 2020). Algoritam za učenje koristi atribute i obeležene entitete kako bi prepoznao imenovane entitete (Goyal et al., 2018). Jedna od mana algoritama mašinskog učenja je to što izbor atributa u velikoj meri utiče na performanse algoritma. Tako da sam izbor atributa je značajan korak koji obično zahteva domenske eksperte koji će vršiti njihov izbor (J. Li et al., 2020).

Poslednjih godina algoritmi dubokog učenja, podskup algoritama mašinskog učenja u kome algoritam automatski uči reprezentacije za detekciju bitnih atributa, postaju dominantni za NER i ostvaruju vrhunske rezultate (J. Li et al., 2020). Dva tipa modela se ističu u kontekstu dubokog učenja, modeli bazirani na LSTM mrežama poput bidirekcione LSTM CRF mreže i noviji modeli bazirani na transformer arihtekturi poput bidirekcione enkoder reprezentacije za tranformere koji je postigao vrhunske rezultate za nekoliko različitih NLP zadataka, uključujući i NER zadatak. Modeli bazirani na transformer arhitekturi su efektivniji od LSTM modela kada su transformirani pre-trenirani na ogromnom skupu podataka, ali pokazuju znatno lošije rezultate za NER kada nisu pre-trenirani i kada je obučavajući skup ograničen (J. Li et al., 2020).

3.2. Prepoznavanje imenovanih entiteta u medicinskim dokumentima

Rana istraživanja u oblasti NER za tekstove iz medicinskog domena bila su bazirana na adaptaciji metoda razvijenih u sklopu MUC konferencija (Meystre et al., 2008). Pristupi razvijeni u okviru ranih istraživanja bili su zasnovani na rečnicima i ručno kreiranim pravilima od strane eksperata. Istaknuti primeri su sistemi MedLEE (Friedman et al., 1994) i MetaMap (Aronson, 2001). Iako su sistemi zasnovani na rečnicima i pravilima pružali zadovoljavajuće performanse, nekoliko inherentnih ograničenja onemogućilo je njihovo masovno korišćenje u praksi. U većini slučajeva rečnici su prikupljeni, a pravila razvijana na osnovu vrlo ograničenog korpusa medicinskih dokumenata. Najčešće je taj korpus bio iz jedne medicinske ustanove i vezan samo za jedan domen (npr. kardiologiju). To je rezultovalo tzv. domenski-specifičnim sistemima koji imaju loše performanse kada se primene na korpuse na kojima nisu razvijani. Iako je adaptacija takvih sistema moguća, ona zahteva previše vremena i drugih resursa. U tom kontekstu, značajna mana sistema zasnovanih na rečnicima i pravilima je to što za početni razvoj kao i dalje održavanje zahtevaju domenske eksperte, odnosno velike finansijske resurse (Goyal et al., 2018).

Sledeći logičan metodološki korak u razvoju medicinskog NER bio je upotreba tehnika mašinskog učenja, koje bi, po svojoj definiciji, trebalo da imaju veću moć generalizacije od ručno kreiranih pravila. Međutim, razvoj metoda zasnovanih na mašinskom učenju u medicini kaskao je za NER sistemima iz opšteg domena kao što su novinski članci ili Web sajtovi (Dehghan et al., 2013). Najveća prepreka bila je nedostupnost velikih korpusa koji su potrebni za obučavanje i međusobno poređenje sistema. Vremenom, kroz veliki trud istraživačke zajednice pojavili su se prvi javno dostupni korpusa od kojih su najznačajniji MIMIC i THYME. Sledeći važan korak u razvoju metoda baziranih na mašinskom učenju su istraživački izazovi (i2b2/n2c2, CCKS (X. Li et al., 2021), SemEval (Elhadad et al., 2015), itd.). Pored podsticaja istraživačima da povećaju performanse svojih sistema ovi izazovi su svoje korpuse učinili javno dostupnim pa se zato smatraju jednim od najznačajnijih faktora koji je uticao na napredak medicinskog NLP-a (Spasic et al., 2020; S. Wu et al., 2020).

Nakon pojave javno dostupnih korpusa u kojima su medicinski imenovani entiteti ručno obeleženi, metode mašinskog učenja sve više dobijaju na značaju, dok se pristupi zasnovani na pravilima smatraju zastarelim (Wang et al., 2018). Tadašnje metode zasnovane na mašinskom učenju zavisile su od velikog broja leksičkih, lingvističkih, ortografskih i semantičkih osobina ekstrahovanih iz medicinskih tekstova. Svaka od osobina tipično je zahtevala poseban alat za ekstrakciju koji su razvijali eksperti za NLP uz konsultacije sa lekarima. Da bi se postigle vrhunske performanse, eksperti za NLP i mašinsko učenje ulagali su veliki trud kako bi izabrali podskup osobina za svaki zaseban slučaj. Takođe, čak i sa pogodno izabranim skupom osobina, prvi sistemi mašinskog učenja nisu pružali vrhunske performanse pa su istraživači kreirali tzv. hibridne sisteme. Hibridni sistemi integrišu više različitih bazičnih metodologija, kao što su metode bazirane na mašinskog učenja, rečnici i ručno kreirana pravila. Ovakvi sistemi su na istraživačkim izazovima imali performanse na nivou ljudskih eksperata (Aleksandar Kovačević, 2011; Sun et al., 2013b). Međutim, zbog ograničene moći generalizacije hibridnih sistema, nisu mogli direktno da se koriste nad korpusima iz drugih kliničnih okruženja, već su zahtevali intervenciju prilagođavanja od eksperta različitih profila. Zbog navedenog ograničenja generalizacije, ni hibridni sistemi nisu imali masovnu upotrebu u praksi.

Upotreba metoda dubokog učenja, koje se mogu primeniti direktno nad sirovim podacima bez ekstrakcije osobina, predstavlja pravac koji trenutno najbrže vodi ka masovnom razvoju i korišćenju medicinskih NER sistema. Prvi značajan u ovom kontekstu je iz 2017. gde su autori (Dernoncourt et al., 2017), na dva značajna korpusa, demonstrirali da sistem baziran na dubokom učenju ima bolje performanse od trenutno najboljih sistema. Autori (Y. Wu et al., 2017) su pokazali slične rezultate u okviru studije u kojoj su uporedili performanse tada najboljeg modela koji zahteva ručno ekstrahovane osobine (CRF modela) sa performansama tada aktuelnih metoda dubokog učenja (LSTM i CNN – konvolutivne neuronske mreže). Oni su pokazali da LSTM model nadmašuje performanse CRF i CNN modela. Nakon toga upotreba dubokog učenja postepeno preuzima dominaciju medicinskom NER (Bose et al., 2021; S. Wu et al., 2020). Relativno noviji napreci ogledaju se u primeni transformer arhitekture, gde su autori (J. Lee et al., 2020) obučili BERT model na bio-medicinskom korpusu od 4.5 milijardi reči, sa kojim su uspeli da nadmaše performanse postojećih modela na nekoliko bio-medicinskih zadataka poput ekstrakcije relacija, odgovaranja na pitanja, i NER-a. Slične rezultate su dobili autori (Kim & Lee, 2020), kada su

poredili performanse BERT modela sa bidirekcionim LSTM modelom za medicinski NER.

Rezultati metoda dubokog učenja su sve češće uporedivi sa rezultatima ljudskih eksperata, koji se tipično mere pomoću međusobne saglasnosti anotatora. Nad korpusom koji je proistekao iz istraživačkog izazova i2b2 2010. godine, autori (Zhou et al., 2021) su, pomoću modela zasnovanog na dubokom učenju, ostvarili F1-meru od 0.874 što je za nijansu manje od F1-mere dobijene proverom međusobne saglasnosti anotatora eksperata od 0.880. Sličan rezultat dobili su i (J. Lee et al., 2020) nad korpusom BC5CDR, gde je razlika F1-mera između eksperata i modela bila samo 0.04. Iako su performanse metoda dubokog učenja slična performansama ljudi nad korpusima na Engleskom jeziku, na drugim jezicima još uvek nisu dosegle taj nivo (Akhtyamova et al., 2020; W. Lee et al., 2018b).

3.3. Prepoznavanje imenovanih entiteta u medicinskim dokumentima na srpskom jeziku

Prilagođavanje i razvoj medicinskih NLP tehnika za srpski jezik predstavlja kompleksan zadatak kome do sada nije posvećeno puno pažnje. U nastavku su opisane sve značajne studije koje se bave razvojem sistema za medicinski NER na srpskom jeziku.

Autori (Jaćimović et al., 2015) predstavili su sistem za automatsku deidentifikaciju kliničkih dokumenata na srpskom jeziku koji predstavlja adaptaciju postojećeg NER sistema razvijenog za tekstove opšteg domena (Krstev et al., 2014; Šandrih et al., 2019). Sistem je implementiran uz pomoću metoda zasnovanih na pravilima. Korpus koji je upotrebljen za razvoj i evaluaciju sastoji se od 200 nasumično izabranih kliničkih dokumenata sa ukupnim brojem od 143,378 reči. Kategorije entiteta za deidentifikaciju koje sistem detektuje su: osobe, datumi, geografske lokacije, organizacije, i brojevi. Sistem je postigao ukupnu F1 meru od 0.94. Iako su vrednosti mere performansi visoke, predstavljeni sistem poseduje sva gorenavedena ograničenja vezana za metodologije bazirane na ručno kreiranim pravilima. Dakle, glavno ograničenje je to što pravila sadrže veliki broj jezičkih konstrukcija koje su specifične za korpus na kome su razvijena što znači da neće imati tako dobre rezultate na medicinskim tekstovima koji nemaju date jezičke konstrukcije.

Statistički metod za NER nad medicinskim dokumentima na srpskom jeziku je razvijen od strane autora rada (Puflović et al., 2016). Model se sastojao od slovnih n-grama i n-grama reči kojima su dodeljene numeričke vrednosti zasnovane na frekvenciji datih n-grama u korpusu koji je služio za razvoj. Na osnovu datih vrednosti sistem vrši označavanje imenovanih entiteta u tekstu. Upotrebljeni korpus sastojao se od dokumenata sa neurološke klinike. Sistem je dizajniran da prepozna pet tipova entiteta: ICD 10 kodove bolesti, nazive lekova, medicinske skraćenice, brojeve (doze lekova, datume i vremena), i indikator koji označava da li je terapija uspešno završena. Prijavljena tačnost modela, evaluirana na 100 ručno proverenih dokumenata, je u opsegu od 0.64 to 0.90. U rezultatima nije jasno definisano na kojim kategorijama i koji tip grešaka je uticao da sistem ima lošije rezultate, kao ni na kojim kategorijama je sistem imao dobre performanse. U svakom slučaju i sistem obučen na ovaj način takođe u veoma velikoj meri zavisi od jednog korpusa i imaće dobre rezultate na drugim korpusima samo ako oni sadrže odgovarajuće n-grame.

Statistički metod za NER nad medicinskim dokumentima na srpskom jeziku je razvijen od strane autora rada (Avdic et al., 2020). Model se sastojao od slovnih n-grama i n-grama reči kojima su dodeljene numeričke vrednosti zasnovane na frekvenciji datih n-grama u korpusu koji je služio za razvoj. Na osnovu datih vrednosti sistem vrši označavanje imenovanih entiteta u tekstu. Za svaki tip entiteta prikupljen je zaseban rečnik. Druga metoda koristi relativnu frekvenciju termina u obučavajućem korpusu. Treća metoda je proširenje druge metode sa ručno kreiranim pravilima za ispravljanje čestih grešaka i detekciju skraćenica. Predložene metode testirane su na korpusu od 2000 medicinskih izveštaja, sa deset različitih tipova dijagnoza iz 32 medicinska centra. Medicinski izveštaji u korpusu sastoje se od struktuiranog dela i nestruktuiranog teksta sa anamnezom pacijenta koja u proseku sadrži 12 reči. Treća metoda je navedena kao metoda sa najboljim performansama, sa prosečnom F1 merom za sve termine od 0.937, a najviša F1 mera za medicinske termine je 0.896. Visoke performanse predstavljenog sistema mogu se pripisati samoj metodologiji koja je veoma zavisna od korpusa kao i neuobičajeno kratkim medicinskim izveštajima u samom korpusu. Mogućnost izdvajanja dvosmislenih i složenih medicinskih termina, prisutnih u dužim kliničkim tekstovima kao što su otpusne liste nije očigledna iz rezultata rada.

U literaturi postoji ograničen broj naučnih publikacije posvećenih primeni savremenih metoda dubokog učenja na NER zadatak na srpskom jeziku, pri čemu se postojeći radovi većinom fokusiraju na opšti domenu poput rada (Ljubešić & Lauc, 2021).

Na osnovu analize postojećih studija može se zaključiti da je medicinski NER za srpski jezik i dalje nedovoljno istražena oblast. Svi razvijeni pristupi oslanjaju se na rečnike i pravila koje veoma teško adaptirati tako da se mogu uspešno primeniti na tekstove iz drugih medicinskih domena ili ustanova. Metode mašinskog i dubokog učenja za medicinski NER, po najboljem saznanju autora, nisu primenjivane ni evaluirane na korpusima na srpskom jeziku.

4. Korpus zlatnog standarda

U kontekstu obrade prirodnog jezika, za obučavanje algoritama mašinskog učenja potrebno je imati anotiran korpus podataka na osnovu koga ti algoritmi mogu da se obuče za rešavanje zadataka NLP-a. Prilikom anotiranja korpusa podataka veoma je bitno da je formiran na način sa kojim je smanjena mogućnost pojave greška u anotacijama, jer greške iz anotiranog korpusa utiču na greške i tačnost obučenih algoritama. Korpus zlatnog standarda je korpus koji je anotiran od najmanje dva anotatora i u čijem procesu formiranja su vršeni koraci za osiguranje pouzdanosti anotacija (poput iterativnog formiranja uputstva anotiranja, obučavanje i evaluacija saglasnosti anotatora) (Pustejovsky & Stubbs, 2012; Wissler et al., 2014).

Za potrebe ovog istraživanja prikupljen je skup tekstualnih medicinskih dokumenata (korpus) sa Klinike za nefrologiju Univerzitetskog kliničkog centra Srbije. Korpus se sastoji od 17929 medicinskih dokumenata, obuhvatajući period od 1994. do 2016. godine.

Svi medicinski dokumenti korišćeni u ovom istraživanju su obrađeni u skladu sa relevantnim standardima etičke prakse i propisima o zaštiti privatnosti pacijenata. Primenjena su sva pravila definisana od strane Univerzitetskog kliničkog centra Srbije, uz poštovanje procedura i smernica nadležnih etičkih tela. Metodologija i sprovođenje akademskog istraživanja odobreno je od strane Etičkog odbora Univerzitetskog kliničkog centra Srbije pre početka istraživanja, čime je verifikovano da su sve planirane faze prikupljanja i analize podataka u skladu sa visokim standardima etičke prakse u medicini.
Pre preuzimanja dokumenata sa klinike izvršena je njihova deidentifikacija, odnosno svi identifikacioni podaci o pacijentima u tekstu su zamenjeni sa odgovarajućim surogatima kako bi se onemogućila identifikacija pacijenta i očuvala njihova privatnost.

U pripremnim fazama ovog istraživanja, lekari sa Klinike za nefrologiju su identifikovali medicinske dokumente hroničnih i akutnih bubrežnih bolesnika kao dokumente od interesa za ovo istraživanje. Uzimajući u obzir da su u korpusu sadržani medicinski dokumenti svih pacijenata koji su posetili Kliniku za nefrologiju, uključujući nefrološke pacijente kao i pacijente sa drugih klinika kojima je rađena dijaliza, pre pristupa procesu anotiranja identifikovani su svi dokumenti koje se odnose na hronične i akutne bubrežne bolesnike. Navedeni dokumenti su izabrani kao

dokumenti od interesa zbog činjenice da dokumenti pacijenata koji su upućeni na Kliniku za nefrologiju, sa drugih klinika, u sebi sadrže retke termine specifične za te klinike koji se u korpusu pojavljuju samo jednom čime bi u korpus bio unet nepotreban šum[9].

Dokumenti su identifikovani na osnovu pretrage fraza na početku dokumenta koje su jedinstvene za hronične i akutne bolesnike (fraza poput: „Dugogodišnji bubrežni bolesnik“, „Primljen kao hitan slučaj zbog visokih vrednosti azotnih materija“). Zbog čestih tipografskih greški pronalaženje fraza je izvršeno upotrebom *Jaro* algoritma za približno poklapanje reči koji se pokazao kao najefektivniji za medicinske dokumente na srpskom jeziku (Kaplar et al., 2019).

Zbog ograničenih resursa ovog istraživanja, u vidu dostupnih ljudskih resursa i ograničenog vremena, po uzoru na (W. Lee et al., 2018), za anotiranje nasumično je izabran je podskup od 203 otpusne liste hroničnih i akutnih bubrežnih bolesnika. Na osnovu srodnih istraživanja ovaj korpus se može smatrati korpusom srednje veličine. Pri čemu veličine korpusa zavise od cilja i potrebe istraživanja, pri čemu veličine korpusa korišćene za istraživanja medicinskog NLP-a variraju od 40 dokumenta do nekoliko desetina hiljada, gde je većina korpusa u opsegu od nekoliko stotina do par hiljada dokumenata (Spasic et al., 2020). Dominantan trend u istraživanju medicinskog NLP-a je korišćenje istaknutih javno dostupnih skupova podataka (Durango et al., 2023). Veličina istaknutih skupova podataka, korišćenih u sklopu istraživačkih izazova za medicinski NLP, se kreće od 300 dokumenta korišćenih u sklopu ShARE/CLEF eHealth izazovu iz 2013. godine (Suominen et al., 2013) do 1304 dokumenta korišćena 2014. godine u sklopu i2b2/UTHealth (Stubbs et al., 2015) istraživačkog izazova (Spasic et al., 2020).

4.1. Anotaciona šema

Nakon prikupljanja korpusa dokumenata, potrebno je definisati tipove anotacija (labela) koje želimo da anotiramo, njihovu strukturu i njihove

[9] Pod šumom u podacima smatraju se podaci koji su deformisani, imaju nizak odnos između signala i šuma, ili u sistem unose mnoštvo nekorisni informacija. U konkretnom slučaju termini koji se pojavljuju samo jednom u korpusu ne mogu se iskoristiti za obučavanje algoritama mašinskog učenja jer ne postoji dovoljno primera za obučavanje i testiranje.

relacije. Navedene definicije anotacija nazivamo anotacionom šemom ili modelom anotacija.

Kako bi dobili konzistentne i kvalitetne anotacije, pored dobro smišljenog modela anotacija, potrebno je razviti detaljno uputstvo koje će definisati pravila za anotiranje tipova anotacija iz anotacione šeme (uputstvo za anotiranje). U anotacionom uputstvu potrebno je što detaljnije navesti pravila anotiranja kako bi se izbegle moguće nedoumice anotatora. Razvoj anotacionog uputstva je iterativan proces, u kome se obučavanje anotatora vrši u fazama, gde se na kraju svake faze anotaciono uputstvo dopunjava kako bi se razrešile nedoumice otkrivene u tekućoj fazi. Faze obučavanja se ponavljaju dok god postoje nejasnoće u anotacionom uputstvu.

S obzirom na činjenicu da zadatak ovog istraživanja približno odgovara zadatku postavljenom na i2b2 izazovu za vremenske relacije iz 2012. godine (Sun et al., 2013a), izvršena je detaljna analiza anotacione šeme i anotacionih uputstava definisanih u tom izazovu. Nakon detaljne analize odlučeno je da se izvrši prilogođavanje anotacione šeme i anotacionog uputstva i2b2 izazova iz 2012. godine. Odluka za prilagođavanje i2b2 anotacione šeme i instrukcija za anotatore, kao i sam postupak prilagođavanja, odrađen je uz konsultacije sa nefrologom sa višegodišnjim kliničkim iskustvom.

Postojeća i2b2 šema sastoji se od tri tipa taga: EVENT, TIMEX3 i TLINK.

EVENT tagovi obeležavaju klinički relevantne koncepte, fraze koje preciziraju izvor informacije, kliničke departmane, i ostale klinički bitne događaje. Pod kliničkim relevantnim konceptima se smatraju: problemi (PROBLEMS) poput simptoma na koje se pacijent žali, povrede i dijagnostikovana oboljenja; testovi (TEST) koji su izvršeni u postupku dijagnostike poput laboratorijskih nalaza, biopsija ili radioloških snimaka; tretmani (TREATMENT) poput prepisanih terapija, hirurških i drugih zahvata u cilju lečenja oboljenja. Fraze koje preciziraju izvor informacije (EVIDENTIAL) određuju da li je određeni podatak dobijen od pacijenta (npr. pacijent se žali na prisustvo određenog simptoma) ili dijagnostičkim putem (npr. pregled pacijenta, rentgenski snimci i sl.). Kako bi se ispratio kompletan tok lečenja pacijenata obeležavaju se i klinički departmani (CLINICAL DEPARTMENT) koji su primili pacijenta, sa ili na koji departman se šalje pacijent za dalje lečenje. Događaji koji su relevantni za lečenje pacijenata a ne pripadaju jednoj od gore navedenih događaja se

obeležavaju kao dodatni događaji (OCCURRENCE), npr. da li je pacijent primljen kao hitan slučaj.

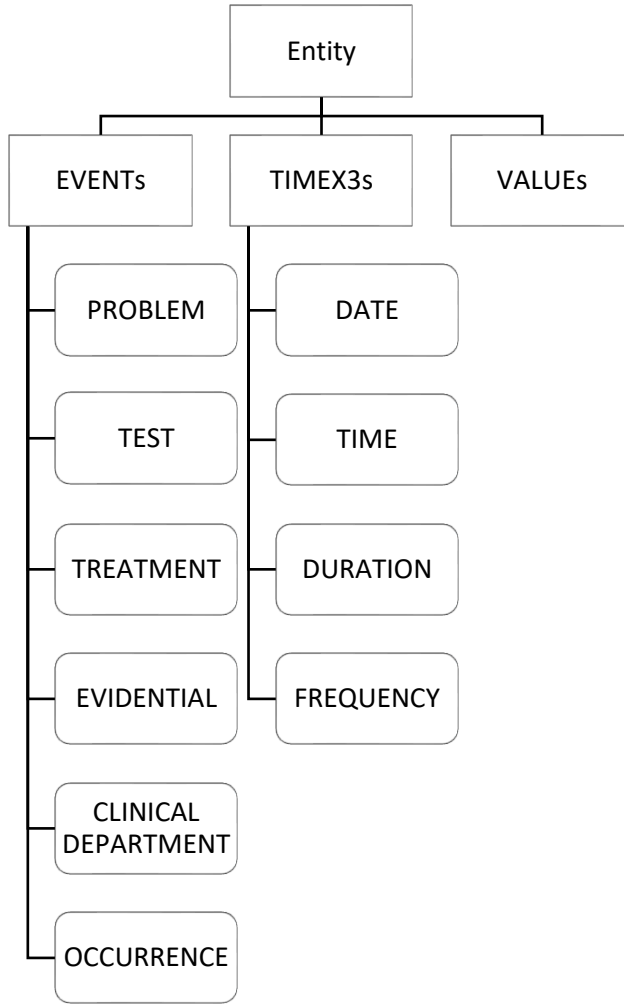
Dodatne informacije o EVENT tagovima su dati u vidi tri atributa: polaritet (eng. *polarity*), modalitet (eng. *modality*) i tip (eng. *type*). Polaritet obeležava da li je događaj pozitivan ili negativan. Na primer ukoliko se pacijent žali na prisustvo određenog simptoma EVENT tag tog simptoma će imati pozitivan polaritet, ukoliko pacijent tvrdi da nema određeni simptom onda će polaritet odgovarajućeg taga biti negativan. Modalitet obeležava da li se događaj zaista dogodio, npr. ukoliko je pacijentu predložena hirurška intervencija taj događaj će imati modalitet *predložen*, a ukoliko je reč o već izvršenoj hirurškoj intervenciji taj događaj će biti imati modalitet *stvaran*. Moguće vrednosti za modalitet su: *stvaran* (eng. *actual*), *hipotetički* (eng. *hypothetical*), *zavistan* (eng. *hedged*) i *predložen* (eng. *proposed*). Atribut tip obeležava kog tipa je EVENT događaj, moguće vrednosti su: PROBLEM, TEST, TREATMENT, EVIDENTIAL, CLINICAL DEPARTMENT, OCCURRENCE.

TIMEX3 tagovi obeležavaju vremenske izraze poput datuma (DATE), doba dana (TIME), trajanja (DURATION), i frekvencija terapija (FREQUENCY). Atributi TIMEX3 tagova su tip, VAL i MOD. Atribut tip je atribut koji obeležava kog tipa je TIMEX3 tag (DATE, TIME, DURATION, FREQUENCY).

VAL predstavlja vrednost TIMEX3 taga u ISO8601 formatu. Prema ISO8601 standardu datumi se navode u formatu YYYY-MM-DD (2022-04-19), vremena se dodaju po formatu Thh:hh.ss (T11:01:00), ili u kombinaciji sa datumom YYYY-MM-DDThh:mm:ss (2022-04-19T11:01.00). Trajanja su data u formatu P[n]Y[n]M[n]DT[n]H[n]M[n]S, odnosno trajanja počinju sa karakterom P praćenom bilo kojom kombinacijom oznake trajanja perioda (Y – godine, M – meseci, W – nedelje, D – dani), ili oznaka trajanja vremena gde se za vremena dodaje oznaka T nakon čega idu oznake za vremenska trajanja (H – sati, M – minuti, S – sekunde). Na primer, ukoliko pacijent treba da pije terapiju dve nedelje navodi se P2W, ili ukoliko davanje lek infuzijom treba da traje 12 sati navodi se PT12H. Ponavljanja se navode sa oznakom R praćeno trajanjem ponavljanja (npr. tri puta dnevno RPT8H).

Ukoliko je potrebno navesti dodatne informacije o TIMEX3 tagu koristi se MOD atribut. MOD atribut predstavlja TimeML (Pustejovsky et al., 2003) TIMEX3 modifikator sa mogućim vrednostima:

- NA - ukoliko nema dodatnih informacija,
- MORE - kada želimo da navedemo da je vremenski period trajao duže od navedenog (npr. duže od nedelju dana, vrednost bi bila nedelju dana a modifikator MORE),
- LESS - kada želimo da navedemo da je vremenski period trajao kraće od navedenog,
- APPROX – kada želimo da navedemo da je reč o približnoj vrednosti (npr. skoro nedelju dana),
- START – opisuje početak vremenskog perioda (npr. početkom 2010. godine),
- END – opisuje kraj vremenskog perioda (npr. krajem 2010.),
- MIDDLE – navodi da se radi o sredini vremenskog perioda (npr. sredinom maja 2010.).

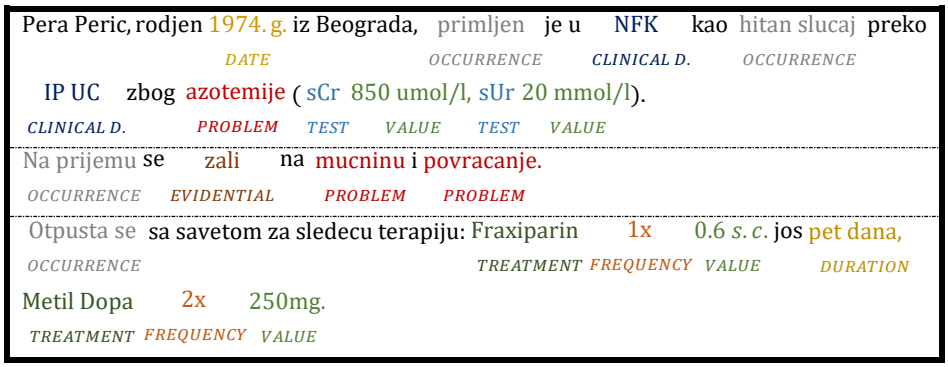


Slika 19. Anotaciona šema

U prilagođenoj šemi TLINK tagovi namenjeni za podzadatak prepoznavanja temporalnih relaciji, koji nije u fokusu ovog istraživanja, su zamenjeni VALUE tagovima koji su namenjeni za prepoznavanje vrednosti EVENT tagova poput doza definisanih u (Uzuner et al., 2010). Anotaciona šema je data na slici 19.

VALUE tagovi obeležavaju vrednosti povezane sa određenim EVENT tagovima, pre svega vrednosti doza prepisanih terapija i numeričkih vrednosti testova. Jedini atribut VALUE taga je ID atribut koji referencira identifikator EVENT taga sa kojim je taj VALUE tag povezan, identifikator EVENT taga se automatski generiše prilikom procesa anotiranja.

Primer anotiranih rečenica, u kojima su podaci pacijenata zamenjeni surogatima, su dati na slici 20.



Slika 20. Primeri anotiranih rečenica

4.2. Proces anotacije

Dva anotatora, doktorandi iz oblasti računarstva, su anotirala korpus od 203 otpusne liste. Nijedan od anotatora nije imao formalno obrazovanje iz medicinskih nauka. Iako su istraživanja pokazala da anotatori sa prethodnim medicinskim iskustvom (zdravstveni radnici) mogu da identifikuju i anotiraju medicinske podatke sa većom preciznošću (Xia & Yetisgen-Yildiz, 2012), autori (Uzuner, Solti, Xia, et al., 2010) su pokazali na i2b2 istraživačkom izazovu da osobe bez prethodnog medicinskog iskustva, uz adekvatnu obuku, mogu da anotiraju medicinske dokumente sa zadovoljavajućom preciznošću.

Dokumenti su anotirani upotrebom „*multi-document annotation environment*“ alatom za anotiranje (Rim, 2016). Anotatori su obučeni u nekoliko odvojenih obučavajućih sesija, u kojima su medicinski termini objašnjeni od strane medicinskih stručnjaka. Nakon svake sesije anotatori su anotirali skup od po pet otpusnih listi, koji prethodno nisu videli, i izračunata je saglasnost anotatora (odnosno izračunati su preciznost, odziv i F1 mera). Sve nejasnoće koje su imali anotatori su dodatno objašnjene u sledećoj sesiji i instrukcije za anotatore su dodatno prilagođene. Obučavanje je ponavljano sve dok nije postignuta saglasnost anotatora, odnosno dok u dve uzastopne sesije nije postignuta vrednost F1 mere od

barem 0.90.[10] Proces obučavanja anotatora prikazan je na slici 21, obučavanje anotatora je trajalo nedelju dana a sam proces anotiranja korpusa je trajao 9 nedelja.

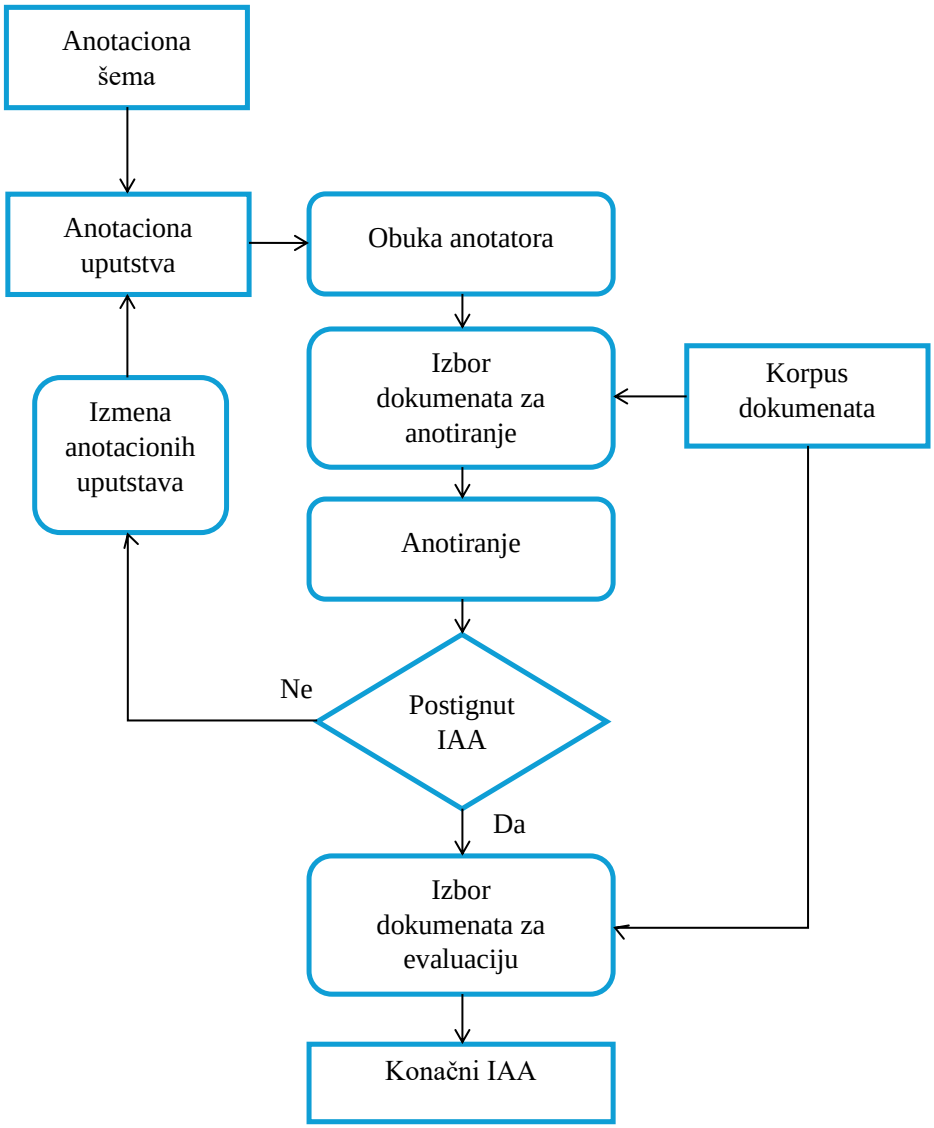
Kako bi odredili kvalitet anotacija korpusa anotatori su dobili novi skup od pet dokumenata (evaluacioni skup), koji nisu korišćeni u procesu obuke, da anotiraju. Na osnovu evaluacionog skupa je izračunata preciznost, odziv, i F1 mera između anotatora. Računanje je izvršeno za tačno poklapanje entiteta (obeleženi tekst entiteta je isti kod oba antoatora) i delimično poklapanje entiteta (obeleženi tekst entiteta ima barem jednu reč koja je ista kod oba anotatora). U tabeli 5. prikazan je konačni IAA.

	Tačno poklapanje entiteta			**Delimično poklapanje entiteta**		
Klasa	**Preciznost**	**Odziv**	**F1**	**Preciznost**	**Odziv**	**F1**
EVENT	0.930	0.852	0.889	0.906	0.941	0.923
TIMEX3	0.971	0.944	0.958	0.976	0.976	0.976
VALUE	0.953	0.906	0.929	0.965	0.954	0.960
Prosečno	0.939	0.873	0.904	0.927	0.947	0.937

Tabela 5. Preciznost, Odziv i F1 mera anotiranih klasa

Za evaluaciju atributa klasa, kako bi bili uporedivi sa IAA atributa i2b2 izazova, dodatno je iskorišćena statistička mera Cohen's Kappa (Pustejovsky & Stubbs, 2012) i tačnost. IAA za atribute EVENT i TIMEX3 klase su date u tabeli 6.

[10] Saglasnost anotatora od 90% kao zadovoljavajući stepen saglasnosti prilikom obuke izabran je na osnovu rezultata sličnih anotacionih zadataka navedenih u literaturi.



Slika 21. Proces obuke anotatora

Prilikom evaluacije atributa VALUE klasa je izuzeta, zbog činjenice da jedini atribut VALUE klase je referenca na jedinstveni identifikator entiteta koji se generiše od strane alata te ih nije moguće uporediti. [11]

[11] Ukoliko se u anotiranom dokumentu nalazi tekst „hct 0,26“ gde je „hct“ EVENT tag a „0,26“ VALUE tag. Kada anotator obeleži EVENT tag on će automatski dobiti

EVENT			TIMEX3		
Attribute	**Kappa**	**Accuracy**	**Attribute**	**Kappa**	**Accuracy**
Polarity	0.799	0.997	MOD	1.000	1.000
Modality	0.908	0.997	VAL	-	0.941
Type	0.996	0.997	Type	1.000	1.000

Tabela 6. *Cohen's Kappa* i Tačnost za atribute EVENT i TIMEX3 klasa

Rezultati IAA, prikazani u tabelama iznad, prikazuju vrednosti uporedive sa rezultatima prikazanih na i2b2 izazovima, na osnovu kojih je šema u ovom radu zasnovana. Vrednosti F1 mere za EVENT i TIMEX3 tagove na i2b2 izazovu su 0.83 i 0.73 za tačno poklapanje. Razlike u rezultatima, gde su u ovom radu predstavljene blago bolje vrednosti, su dobijene zbog manjeg broja anotatora (dva anotatora u okviru ovog istraživanja su izvršili anotiranja, dok je osam anotatora na i2b2 izazovu vršilo anotiranje) i manjeg broja anotiranih dokumenata.

Najveći stepen slaganja anotatora je nad entitetima TIMEX3 klase, to se može objasniti sa činjenicom da su entiteti TIMEX3 klase jasno definisani i ne zahtevaju domensko znanje za njihovu identifikaciju.

Na osnovu IAA rezultata odlučeno je da je kvalitet anotacija korpusa zadovoljavajuću i da se korpus može koristi u daljem toku istraživanja za obučavanje modela mašinskog učenja.

4.3. Karakteristike korpusa

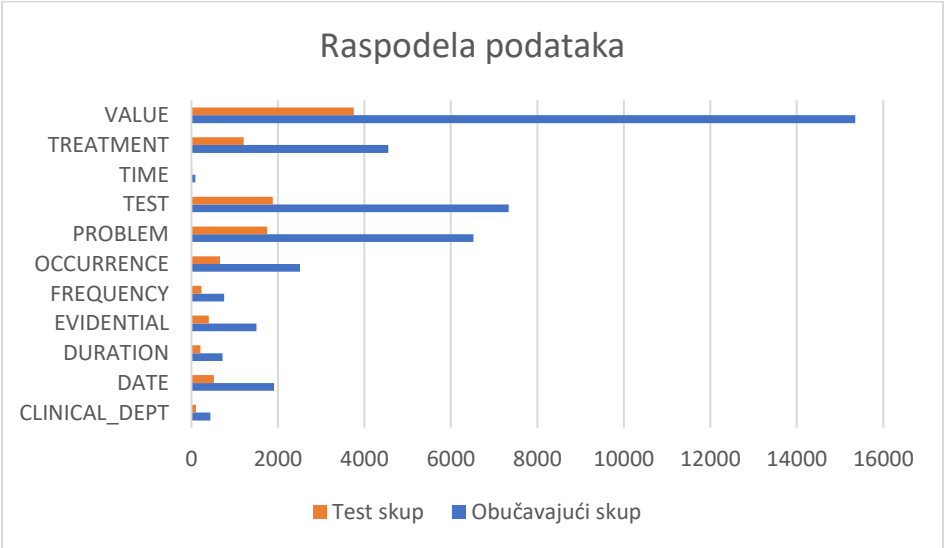
Konačni anotirani korpus se sastojao od 203 anotirane otpusne liste, sa 6,893 rečenice i 123,574 tokena (reči). Korpus je podeljen upotrebom metode stratifikovanog slučajnog uzorka na obučavajući skup podataka i test skup podataka. Podela je izvršena da u proseku test skup podataka sadrži oko 20% svih individualnih klasa entiteta. Konačan broj entiteta po skup dat je u tabeli 7 i grafički predstavljeni na slici 22.

jedinstveni ID (npr. E3), nakon obeležavanja VALUE taga anotator u atributu ID VALUE taga unosi taj ID, odnosno E3. Drugi anotator kada bude anotirao isti dokument će odraditi isti postupak, ali ID EVENT taga ne mora da znači da će biti E3 može da bude npr. E5. Zbog navedenog problema evaluacija atributa EVENT klase je izuzeta.

Statistički podaci na nivou tokena su predstavljeni u tabeli 8, dok je odnos između broja primera u obučavajućem skupu i prosečne dužine entiteta predstavljen na slici 23.

Klasa	Obučavajući skup	Test skup	Obučavajući skup (procentualno)	Test skup (procentualno)
CLINICAL_DEPT	440	105	80.73%	19.27%
DATE	1909	519	78.62%	21.38%
DURATION	719	206	77.73%	22.27%
EVIDENTIAL	1505	400	79.00%	21.00%
FREQUENCY	759	233	76.51%	23.49%
OCCURRENCE	2513	661	79.17%	20.83%
PROBLEM	6526	1747	78.88%	21.12%
TEST	7334	1878	79.61%	20.39%
TIME	87	16	84.47%	15.53%
TREATMENT	4557	1210	79.02%	20.98%
VALUE	15356	3752	80.36%	19.64%
Other	57327	14239	80.10%	19.90%

Tabela 7. Broj entiteta po klasama za obučavajući i test skup



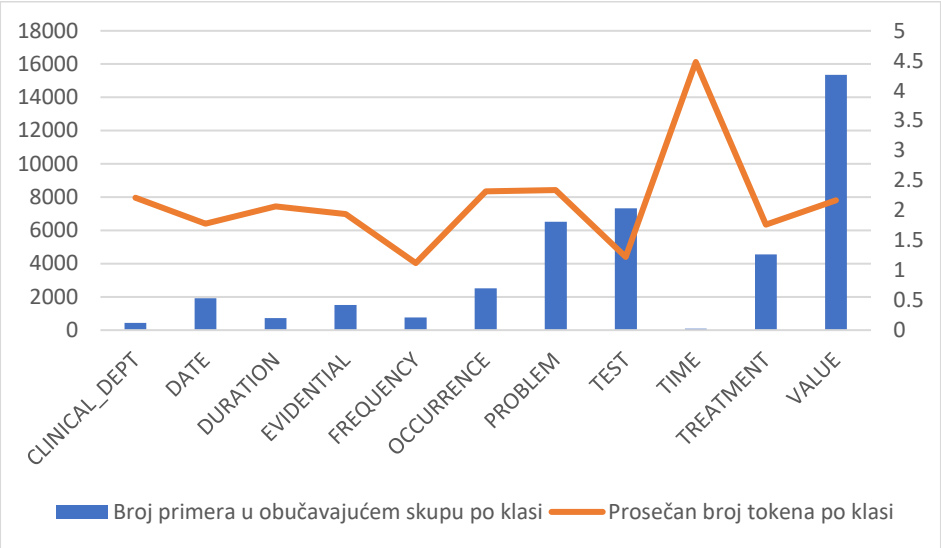
Slika 22. Raspodela podataka po skupovima

Na osnovu podataka o entitetima se vidi da iako postoje entiteti koji se sastoje od više od deset tokena, najčešće su se sastojali od dva tokena.

Klasa	**Tokeni**			
	Min	**Max**	**Mean**	**Mod**
CLINICAL_DEPT	1	7	2.21	2
DATE	1	10	1.78	2
DURATION	1	4	2.07	2
EVIDENTIAL	1	5	1.94	2
FREQUENCY	1	5	1.12	1
OCCURRENCE	1	18	2.32	2
PROBLEM	1	15	2.34	2
TEST	1	11	1.22	1
TIME	2	8	4.48	2
TREATMENT	1	13	1.76	1
VALUE	1	21	2.17	1

Tabela 8. Broj tokena po klasi

U podacima kao i na slici 22 jasno se vidi da klase nisu ravnomerno zastupljene u skupovima podataka, a klasa koja se ističe kao potencijalno problematična je klasa TIME.



Slika 23. Odnos broja primeraka po klasi sa prosečnim brojem tokena po klasi

Entiteti klase TIME u anotiranim otpusnim listovima obično predstavljaju vremena zakazanih kontrola, poput "13:00", ili datume i vremena određenih kliničkih događaja, kao što je "21.08.2017.g. u 13h25min". Ovi entiteti su znatno ređi od entiteta ostalih klasa u korpusu, a u proseku se sastoje 4.48 tokena (Slika 23) što je znatno odstupanje od broja tokena ostalih klasa. Mali broj primera klase TIME kao i relativno veliki broj tokena po entitetu klase TIME rezultovao je sa odstupanja prilikom podele na obučavajući i test skup gde u test skupu imamo samo 15% primera. Iz navedenih problema možemo da pretpostavimo da će algoritmima mašinskog učenja klasa TIME biti izazovna za detekciju. Pre sve algoritmi mašinskog učenja će videti jako mali broj primera klase TIME u toku svog obučavanja, pri čemu entiteti imaju znatno veći broj tokena po klasi u odnosu na ostale entitete.

Još jedan potencijalno problematičan faktor za klasifikacione algoritme je taj što, za razliku od klasičnog NER gde su entiteti (poput imena osoba, organizacija i sl.) imenice koje počinju sa velikim slovom, u korpusu medicinskih dokumenata iskorišćenih za ovo istraživanje većina entiteta počinje sa malim slovom.

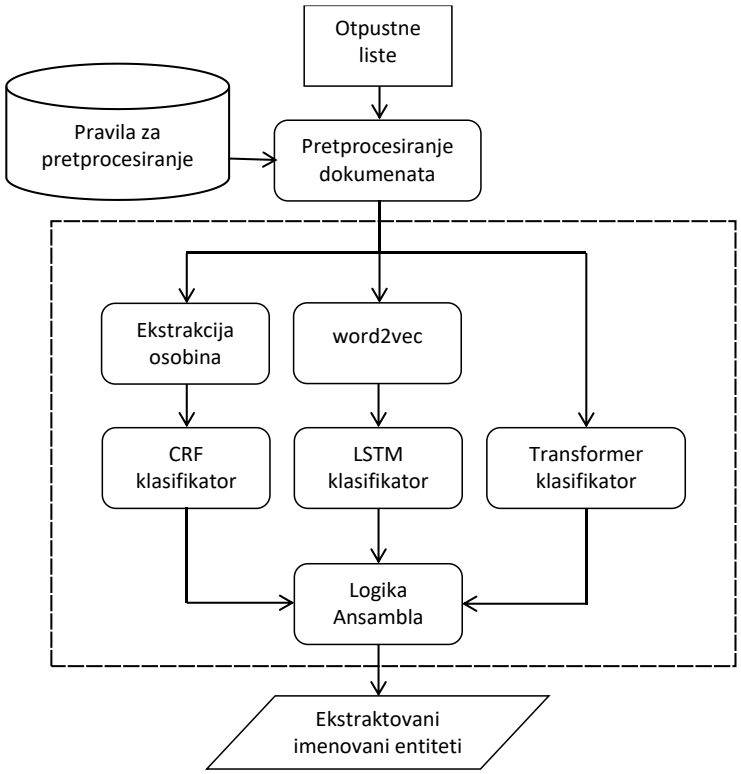
5. Model sistema za prepoznavanje imenovanih entiteta

U ovom poglavlju prvo je predstavljena arhitektura modela sistema. Nakon predstavljanja arhitekture, opisan je način pretprocesiranja dokumenata potreban za dalje obučavanje i korišćenje modela mašinskog učenja. Na kraju ovog poglavlja dat je opis načina formiranja i obučavanja modela mašinskog učenja koji su iskorišćeni za formiranje arhitekture sistema.

5.1. Arhitektura sistema

Arhitektura sistema za automatsko prepoznavanje imenovanih entiteta data je na slici 24, prepoznavanje imenovanih entiteta u medicinskim dokumentima se sastoji iz četiri koraka: pretprocesiranje dokumenata, priprema dokumenata za obradu sa obučenim modelom, individualno prepoznavanje imenovanih entiteta, kombinovanje individualnih rezultata sa ansambl strategijom. Medicinski dokumenti se prvo pretprocesiraju unapred definisanim pravilima, nakon čega se šalju modelima mašinskog učenja na dalju obradu i prepoznavanje entiteta. Sistem koristi tri različita tipa modela mašinskog učenja, a to su: CRF, LSTM i transformeri. Dodatna obrada pretprocesiranih dokumenata predstavlja transformisanje originalnog teksta u format koji je pogodan za obradu sa izabranim algoritmima mašinskog učenja. Svaki od navedenih modela prepoznaje imenovane entitete nad prosleđenim podacima tako što svaki token obeleži sa odgovarajućom klasom entiteta. Kako bi poboljšali tačnost sistema, modeli su kombinovani u ansamblu većinskim glasanjem. Za svaki token u dokumentu, ansambl poredi rezultate individualnih modela i odlučuje koja klasa će biti dodeljen procesiranom tokenu. Za konačnu odluku dodeljivanja klase tokenu koristi strategiju većinskog glasanja, odnosno tokenu se dodeljuje klasa za koju je većina modela glasala (dala kao rezultat).

Obučavanje i evaluacija performansi modela, koji su navedeni u ovom poglavlju, je vršena na računaru sa sledećim hardverskim specifikacijama: 32 GB RAM memorije, Intel i7 procesor, NVIDIA GeForce GTX 1080 Ti grafička kartica sa 11 GB RAM memorije.



Slika 24. Arhitektura sistema

5.2. Pretprocesiranje korpusa

Korpus je organizovan u tekstualne dokumente koje sadrže tačno jednu otpusnu listu. Tekst u otpusnim listama je napisan sa karakterima u UTF-8 formatu, većinski u latiničnom pismu, u jednoj liniji (bez znakova za prelaz u novi red). Kao prvi korak pretprocesiranja, svi dokumenti koji napisani sa ćiriličnim pismom su prevedeni na latinično pismo sa postojećim alatima za transliteraciju. U određenim dokumentima, za određena latinična slova, su korišćene simbolične zamene (poput „[“ i „{“ za „š“, „]“ i „}“ za „ć“, „@“ za „ž“ i sl.), dok je većina dokumenata bila napisana u formi bez dijakritika. Navedeni simboli, kao i slova sa dijakriticima, su zamenjena sa slovima bez dijakritika.

U periodu vršenja ovog ispitivanja, alati za podelu teksta medicinskih dokumenata u rečenice nisu postojali, te je za potrebe ovog istraživanja napisan alat za podelu teksta u rečenice. Alat upotrebom jednostavnih pravila, formiranih na osnovu osobina dokumenata u korpusu, deli tekst na individualne rečenice. Nakon podele teksta u rečenice, a pre

tokenizacije, reči greškom spojene sa brojevima su rastavljene u individualne reči. Na primer, tekst „htco,26“ zapravo predstavlja laboratorijski nalaz pacijentovog hematokrita (hct) gde je vrednost spojena sa nazivom testa, i dodatno gde je slovo „o“ iskorišćeno umesto broja nula. Nakon ispravke grešaka, izvršena je tokenizacija sa NLTK (Bird et al., 2009) jezički nezavisnim tokenizerom koji rastavlja reči na osnovu interpunkcijskih simbola u rečenici. Primer pretprocesiranja rečenice dat je na slici 25.

Sr}ani ritam nepravilan po tipu absoulute,tonovi tihi ali jasni,{umove ne }ujem, TA160/100.										
Srcani	ritam	nepravilan		po	tipu	absoulte	,	tonovi	ali	
jasni	,	sumove	ne	cujem	,	TA	160	/	100	.

Slika 25. Primer pretprocesiranja rečenice.

5.3. Obučavanje modela mašinskog učenja

5.3.1. Podela korpusa za obučavanje

Nakon pretprocesiranja korpus od 203 anotirana dokumenata je podeljen na obučavajući skup i na test skup dokumenata. Obučavajući skup je iskorišćen za obučavanje modela za NER zadatak, dok su performanse obučenih modela evaluirane na test skupu.

Pre samog procesa obučavanja, određeni modeli poput LSTM modela i modela zasnovanih na transformer arhitekturi, su zahtevali proces pre-treniranja. Pre-treniranje vektora značenja reči za LSTM model i pre-treniranje zadatka jezičkog modelovanja za transformer modele je izvršeno na preostalih 17 000 dokumenata koji nisu iskorišćeni u toku procesa anotiranja.

5.3.2. Uslovna slučajna polja (CRF)

CRF pripada grupi probabilističkih grafovskih metoda za segmentaciju i labeliranje sekvencijalnih podataka (Goyal et al., 2018; Lafferty et al., 2001). U NER domenu, CRF se istakao kao metod sa dobrim performansama (Goyal et al., 2018; Patrick & Li, 2010), i u sistematskim pregledima literature se pojavljuje kao najčešće korišćen metod za ekstrakciju EVENT i TIMEX entiteta iz kliničkih tekstova (Alfattni et al., 2020; Keretna et al., 2015; Wang et al., 2018). Iz prethodno navedenih razloga, poput autora (Goyal et al., 2018; Habibi et al., 2017; Moharasan

& Ho, 2017), u okviru ovog istraživanja odlučeno je da se CRF koristi kao osnova za poređenje performansi drugih NER modela.

Inicijalni skup atributa za CRF model formiran je po uzoru na često korišćene atribute u literaturi i prikazan je u tabeli 9.

Grupa atributa	Atribut	Primer
Leksički atributi	Token malim slovima	slabost
	Poslednja 3 slova token (identifikacija dužih sufiksa)	ost
	Poslednja 2 slova tokena (identifikacija kraćih sufiksa)	st
	Koren reči	slab
Šablonski atributi	Token je zarez	False
	Oblik reči[12]	wwwwwww
	Token je broj	False
	Token počinje sa velikm slovom	False
	Token je u alfanumeričkoj formi	True
Kontekstne osobine	Osobine okolinih tokena u rasponu od 5	

Tabela 9. Atributi CRF algoritma sa primerom za token "slabost" u rečenici "Za bubrežnu slabost zna od 3.2021. godine"

Hiper-parametri modela (*c1* – koeficijent *L1* regularizacije, *c2* – koeficijent *L2* regularizacije) su određeni algoritmom za nasumičnu pretragu unakrsnom validacijom iz *scikit-learn* biblioteke. Algoritam za nasumičnu pretragu unakrsnom validacijom nasumično bira vrednosti hiperparametara iz unapred datog opsega i trenira model sa tim hiperparametirma. Nakon treniranja vrši poređenje performansi modela. Algoritam ponavlja postupak nekoliko puta (u ovom disertacije broj ponavlja je postavljen na 100) i kao rezultat vraća vrednosti hiperparametara modela koji je imao najbolje performanse. Kao najbolje vrednosti za *c1* i *c2* parametre dobijeni su vrednosti 0.121 i 0.242.

[12] Oblik reči je osobina koja preslikava token u šablon korišćenih alfa-numeričkih karaktera npr. „sCr“ u „wWw“, „D3“ u „Wd“ i sl.

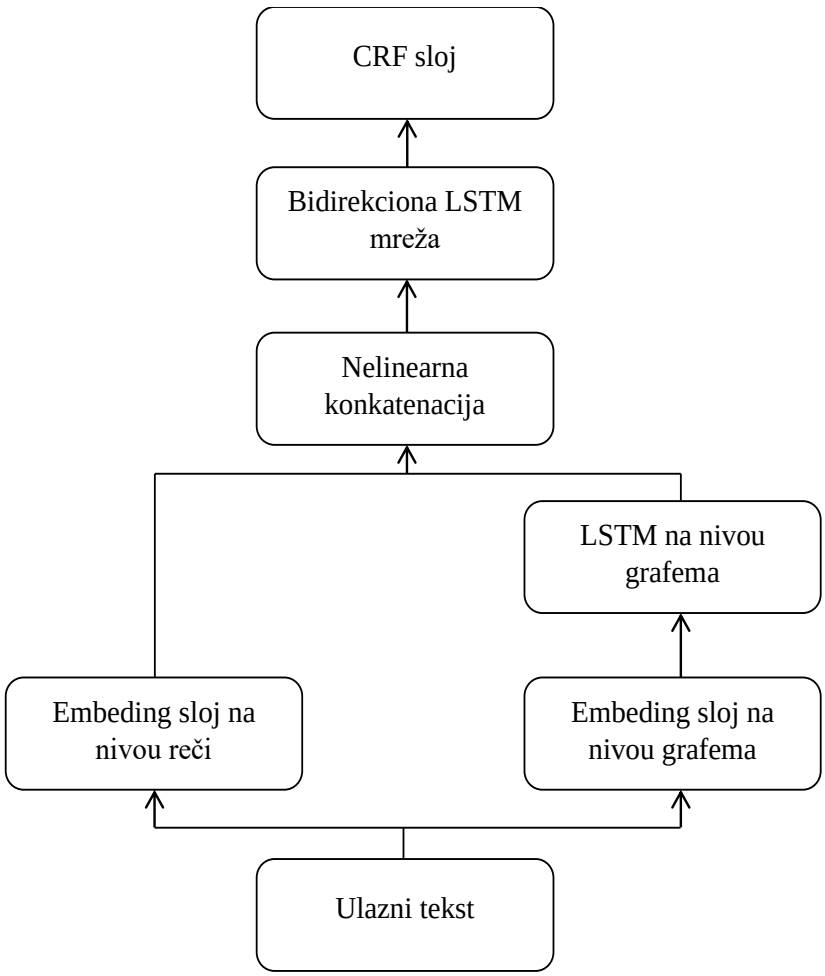
Nakon što je određen skup atributa sa najboljim performansama, izvršena je ablaciona studija atributa kako bi se odredila relevantnost individualnih atributa na performanse modela. Kako bi se video uticaj atributa na performanse CRF modela, atributi koje je CRF algoritam koristio sistematično su bili uklanjani i obučavanje CRF algoritma ponavljano. Ablaciona studija je rezultovala uklanjanjem dva šablonska atributa (token je zarez i token je broj) iz skupa atributa jer njihova upotreba je imala uticaj manji od 0.1% na konačni rezultat. Takođe vršena je provera uticaja broja okolnih tokena na rezultat, gde se pokazalo da je raspon od 5 (dva tokena pre trenutnog tokena i dva tokena posle) rezultovao sa najboljim performansama.

5.3.3. Rekurentna neuronska mreža sa dugotrajnom kratkoročnom memorijom (LSTM)

Jedan od glavnih nedostataka tradicionalnih metoda mašinskog učenja, poput CRF modela, je njihova uslovljenost ručno izabranim atributima. Moguće rešenje navedenog problema je upotreba metoda dubokog učenja, odnosno, upotreba veštačkih neuronskih mreža poput RNN ili CNN algoritama (L. Liu et al., 2018). Relativno skoriji napreci, u domenu dubokog učenja, rezultirali su povećanom primenom metoda mašinskog učenja za rešavanje problema kliničkog NLP. Autori (S. Wu et al., 2020) su naveli da je 2018. godine došlo do povećanja objavljenih radova iz domena dubokog učenja za 200%. Konkretno za NER zadatak, autori su istakli LSTM kao najčešće primenjivani algoritam. Metode dubokog učenja bazirane na LSTM algoritmu nadmašuju rezultate CRF pristupa kada je obučavajući skup dovoljno veliki, dok CRF algoritam ima bolje performanse nad manjim obučavajućim skupovima (Ramos-Flores et al., 2020). Štaviše, dodavanje CRF sloja LSTM modelu omogućava upotrebu informacija o tagovima na nivou rečenica što pospešuje performanse samog modela (Z. Huang et al., 2015). Bidirekcioni LSTM modeli sa CRF slojem imaju najbolju ili približno najbolju preciznost za zadatke tagovanja sekvenci poput tagovanja vrsta reči ili NER (Z. Huang et al., 2015; Su et al., 2019).

Neuronske mreže poput LSTM mreža, kao što je napomenuto u sekciji 2.3, zahtevaju numeričke vrednosti kao ulaze u mrežu koji se dobijaju kao vektori jedinice ili vektori značenja reči. U literaturi je dodatno pokazano da upotreba pre-treniranih vektora značenja reči, embeding pristup, u kombinaciji sa algoritmima dubokog učenja, poboljšava preciznost

modela za NER zadatak (Habibi et al., 2017; Z. Huang et al., 2015; Su et al., 2019). U praksi prilikom obučavanja neuronskih mreža koriste se prethodno obučeni embeding modeli koji se dodaju kao ulazni sloj u novu arhitekturu. U ređim situacijama moguće je obučiti embeding sloj prilikom obučavanja LSTM modela, ali mana ovog pristupa je što sam proces obučavanja neuronske mreže i embeding modela traje mnogo duže nego kada se koristi prethodno obučeni embeding model. U toku vršenja eksperimenta u okviru ove disertacije nisu postojali prethodno obučeni embednig modeli sa medicinskom terminologijom srpskom jeziku. Iz navedenog razloga, pre obučavanja LSTM-CRF modela, prvo je izvršeno pre-treniranje embeding modela, upotrebom word2vec modela iz gensim biblioteke (Rehurek & Sojka, 2010), na neanotiranim skupu deidentifikovanih medicinskih dokumenta.



Slika 26. Arhitektura LM-LSTM-CRF mreže

U okviru ovog istraživanja, korišćena je LM-LSTM-CRF arhitektura predložena od strane autora (L. Liu et al., 2018). LM-LSTM-CRF je bidirekciona LSTM mreža sa CRF slojem (Slika 26).

Korišćena LM-LSTM-CRF arhitektura proširuje bidirekcionu LSTM-CRF arhitekturu, predstavljenu u sekciji 2.2, sa LSTM mrežom na nivou grafema (karaktera). Dodavanjem LSTM mreže na nivou grafema konačan model može da prepozna leksičke osobine jezika i stil pisanja. Dodavanje navedenih osobina je odrađeno tako što se izlazi iz LSTM mreže na nivou karaktera konkateniraju sa vektorima značenja reči, uz pomoć jednostavne neuronske mreže koja koristi nelinearnu aktivacionu funkciju, pre ulaza u bidirekcionu LSTM mrežu.

Sloj	Parametar	Vrednost
Embeding sloj na nivou grafema	dimenzija	30
LSTM na nivou grafema	dimenzija	300
	broj skrivenih slojeva	1
Nelinearna konkatenacija	broj skrivenih slojeva	1
Embeding sloj na nivou reči	dimenzija	300
Bidirekciona LSTM mreža	dimenzija	300
	broj skrivenih slojeva	1
Optimizacija	brzina obučavanja	0.015
	optimizator	SGD[13]
	broj epoha	200[14]

Tabela 10. Hiperparametri LM-LSTM-CRF mreže

Bitno je napomenuti da bez upotrebe pre-treniranih vektora značenja reči nije bilo moguće uspešno obučiti LM-LSTM-CRF model. Odnosno, na relativno malom skupu anotiranih podataka, model nije bilo moguće obučiti sa upotrebom vektora jedinice ili vektora značenja reči koji su obučeni na samo anotiranom skupu.

[13] https://pytorch.org/docs/stable/generated/torch.optim.SGD.html
[14] Broj epoha je izabran na osnovu preporuke iz rada (L. Liu et al., 2018)

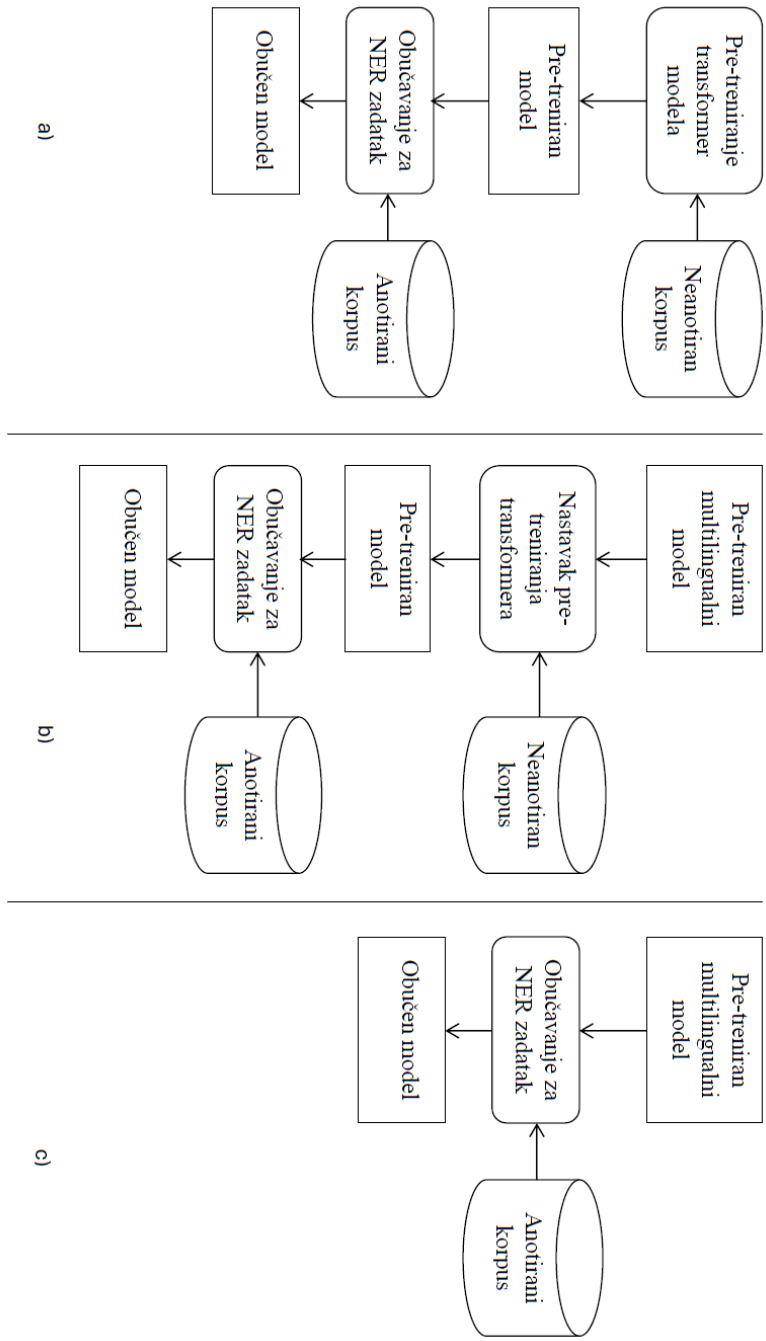
Obučavanja LM-LSTM-CRF mreže je izvršeno upotrebom hiperparametra predloženih od strane (L. Liu et al., 2018) koji su predstavljeni u tabeli 10. Prilikom obučavanja eksperimentisano je i sa različitim optimizatorima i sa dužinom trajanja obučavanja, odnosno sa brojem epoha, gde je korišćen broj od 400 i 600 epoha. U oba slučaja performanse krajnjeg modela nisu davali bolje performanse od modela obučavanog na 200 epoha sa SGD optimizatorom.

5.3.4. Modeli zasnovani na transformerima

Arhitektura transformera prvi put je predložena od strane autora (Vaswani et al., 2017) sa namenom da se pojednostave postojeći modeli za modelovanje sekvence-na-sekvencu zamenom kompleksnih rekuretnih i konvolutivnih neuronskih mreža samo sa mehanizmom pažnje (sekcija 2.4.2). U navedenom radu, autori (Vaswani et al., 2017) su pokazali da predloženi model postiže mnogo bolje rezultate od postojećih modela na zadatku prevođenja teksta.

Najzastupljeniji transformer model je BERT koji se pre-trenira na zadatku jezičkog modelovanja uz pomoć transformer arhitekture zasnovane na enkoder sloju (sekcija 2.4.4). Odnosno, BERT učeći na velikoj količini nelabeliranog teksta može da poboljša svoju efektivnost nad zadacima obrade prirodnog jezika. U domenu kliničkog NLP, dokumentovano je da BERT modeli mogu da nadmaše rezultate prethodno najboljih modela (Kim & Lee, 2020; J. Lee et al., 2020; Peng et al., 2019).

BERT modeli su obično pre-trenirani na ogromnim količinama podataka, kako bi naučili duboke bidirekcione reprezentacije teksta, i samo njihovi poslednji slojevi se obučavaju (ili do-obučavaju) za specifičan NLP zadatak poput odgovaranja na pitanja ili NER (sekcija 2.4.4). Postojeći pre-trenirani transformer modeli mogu da nastave svoje pre-treniranje na novom korpusu, odnonosno da izvrše do-obučavanje jezičkog modela, kako bi se poboljšale performanse tog modela nad tim novim korpusom. Nekoliko različitih pristupa primene transformer modela za klinički NER na srpskom jeziku je analizirano u ovom istraživanju (Slika 27): pre-treniranje jezičkog modelovanja, korišćenje pre-treniranih više-jezičkih modela i nastavljanje zadatka pre-treniranja jezičkog modelovanja postojećih više-jezičkih modela na korpusu neanotiranih medicinskih dokumenata.



Slika 27. Pristupi za korišćene transformer modela: a) pre-treniranje jezičkog modelovanja, b) nastavljanje zadatka pre-trerniranja jezičkog modelovanja postojećih više-jezičkih modela, c) korišćenje pre-treniranih više-jezičkih modela

Pre-treniranje novog modela na korpusu neanotiranih medicinskih dokumenata[15] je odrađeno kako bi utvrdili da li je kontekst reči, koji model može da nauči samo na konkretnom korpusu, dovoljan za prepoznavanje imenovanih entiteta u iskorišćenom korpusu. Više-jezički modeli su pre-trenirani na člancima sa različitih internet stranica, gde su korišćeni članci na oko 100 različitih jezika, koja u sebi obično ne sadrže istu medicinsku terminologiju kao što se koristi u kliničkim uslovima. Iz tog razloga više-jezički modeli su iskorišćeni direktno bez nastavka zadatka pre-treniranja i sa nastavkom zadatka pre-treniranja nad korpusom neanotiranih medicinskih dokumenata kako bi bio procenjen uticaj nastavka zadatka pre-treniranja.

U okviru ovog istraživanja za pre-treniranje novog jezičkog modela korišćen je RoBERTa model (Y. Liu et al., 2019). RoBERTa model ima istu arhitekturu kao BERT model, gde su razlike između ova dva modela u tome što je RoBERTa model: pre-treniran na većem korpusu dokumenata, pri čemu je zadatak predviđanja sledeće rečenice uklonjen. Dodatno, maskiranje prilikom zadatka maskirnog jezičkog modelovanja je odrađeno na dinamički način, odnosno u originalnom BERT radu maskiranje se radi samo u toku pripreme podataka za pre-treniranje dok RoBERTa model svaki put kada obrađuje sekvencu prvo odradi nasumično maskiranje tokena. Modifikacije RoBERTa modela su rezultovale povećanjem performansi modela, u odnosu na BERT model, na GLUE i SQuAD referentnim tačkama (Y. Liu et al., 2019).

Postojeći više-jezički BERT modeli su dostupni u dve varijante: modeli koji prave razliku između malih i velikih slova, i modeli koji su obučeni na tekstu koji je prethodno pretvoren u sva mala slova. Oba tipa modela su iskorišćena u okviru ovog istraživanja.

Svi modeli su dodatno do-obučeni za NER zadatak na anotiranom korpusu medicinskih dokumenata. Za formiranje novog modela na osnovu korpusa neanotiranih medicinskih dokumenata je iskorišćen RoBERTa model, u daljem tekstu T-RoBERTa. Dva više-jezička BERT modela i dva više-jezička RoBERTa modela su iskorišćeni direktno tako što su samo do-obučeni, dok je za jedan od BERT modela i za jedan od RoBERTa model do-obučen jezički model (nastavak zadatka pre-treniranja). BERT više-jezički modeli se razlikuju po tome da li prave razliku između malih i

[15] Pod korpusom neanotiranih medicinskih dokumenata smatraju se dokumenti koji nisu iskorišćeni u procesu anotrianja za kreiranje korpusa zlatnog standarda.

velikih slova. BERT Multilingual Cased, model koji pravi razliku između malih i velikih slova, je obučen na Wikipedia korpusu sa podrškom za 104 različita jezika, dok BERT Multilingual Uncased, model koji ne pravi razliku između malih i velikih slova, je obučen sa podrškom za 102 različita jezika. XLM RoBERTa modeli su obučeni na *CommonCrawl* korpusu veličine 2.5 terabajta sa podrškom za 100 različitih jezika (Conneau et al., 2019). Dve varijante XML RoBERTa modela se razlikuju po veličini treniranog modela gde se manji model, XML RoBERTa Base, sastoji od 270 miliona parametara a veći, XML RoBERTa Large, od 550 miliona parametara. Svi više-jezički modeli su do-obučeni za NER zadatak nad anotiranim korpusom. Nad BERT Multilingual Cased i XML RoBERTa Base je dodatno izvršen nastavak pre-treniranja za zadatak jezičkog modelaovanja nad neanotiranim korpusom nakon čega su do-obučeni nad anotiranim korpusom za NER zadatak. Modeli kojima je nastavljen zadatak pre-treniranja su u daljem tekstu obeleženi sa PT prefiksom (PT - BERT Multilingual Cased, i PT - XML RoBERTa Base). Zbog hardverskih ograničenja obučavajućeg računara do-obučavanje jezičkog modela nije izvršeno nad XML RoBERTa Large modelom. Spisak modela dat je u tabeli 11.

Naziv modela	**Arhitektura**	**Samo mala slova**	**Korpus za zadatak jezičkog modelovanja**
Pre-treniranje jezičkog modelovanja			
T-RoBERTa	RoBERTa	ne	Neanotiran korpus
Korišćenje pre-treniranih više-jezičkih modela			
BERT Multilingual Cased	BERT	ne	*Wikipedia* - 104 jezika
BERT Multilingual Uncased	BERT	da	*Wikipedia* - 102 jezika
XML RoBERTa Base	RoBERTa	ne	*CommonCrawl – 100 jezika*
XML-RoBERTa-Large	RoBERTa	ne	*CommonCrawl – 100 jezika*
Nastavljanje zadatka pre-trerniranja jezičkog modelovanja postojećih više-jezičkih modela			

PT - BERT Multilingual Cased	BERT	ne	*Wikipedia* i neanotiriani korpus dokumenata
PT - XML RoBERTa Base	RoBERTa	ne	*CommonCrawl* i *neanotirani korpus dokumenata*

Tabela 11. Spisak korišćenih modela i korpusa za pre-treniranje

Za implementacija transformer modela je korišćena Huggingface biblioteke (Wolf et al., 2019). Transformer modeli su do-obučeni za NER zadatak za brzinom obučavanja od $2e^{-5}$, kroz 5 epoha[16] sa AdamW optimizatorom[17] i verovatnoćom maskiranja tokena od 0.15. Kao i sa LSTM modelom, eksperimentisano je i sa većim brojem epoha pri čemu dužim treniranjem modela nisu postignute bolje performanse.

5.3.5. Modeli zasnovani na ansamblu

Kao što je napomenuto u sekciji 2.5, modeli zasnovani na ansamblu predstavljaju sisteme koji kombinuju više različitih klasifikatora kako bi dobili precizniji klasifikacioni model (Nayel & Shashirekha, 2017; Raza, 2019; Speck René and Ngonga Ngomo, 2014). Klasifikatori se mogu kombinovati na više različitih načina kako bi se formirao ansambl. Jednostavna i direktna strategija koja se koristi kada su predikcije klasifikatora diskretne vrednosti je strategija većinskog glasanja (Raza, 2019; Speck René and Ngonga Ngomo, 2014), koja je iskorišćena u okviru ovog istraživanja. U ansamblu sa većinskim glasanjem, ulaznom tokenu se dodeljuje labela (klasa) koja je predviđena od većine klasifikatora korišćenih u ansamblu.

Napravljena su dva ansambla sa većinskim glasanjem (Tabela 12), prvi u kome su kombinovani svi prethodno navedeni klasifikatori (u daljem tekstu Ensemble All), i drugi u čijoj kombinaciji su isključeni osnovni transformer modeli koji su imali pre-treniranu varijantu (u daljem tekstu Ensemble Best), odnosno modele za koje je dodatno izvršen nastavak zadatka pre-treniranja. Osnovni transformer modeli koji su imali pre-treniranu varijantu su isključeni uz pretpostavku da će kombinacija

[16] Broj epoha za transformer modele je izabran na osnovu preporuke iz korišćene *huggingface* biblioteke (Wolf et al., 2019)

[17] https://huggingface.co/docs/transformers/v4.38.2/en/main_classes/optimizer_schedules#transformers.AdamW

osnovne i pre-trenirane varijante samo dodatno pojačati grešku tog tipa modela. Kao primer možemo posmatrati XML RoBERTa Base model koji ima problema sa prepoznavanjem numeričkih vrednosti, tako da prilikom klasifikacije numeričkog tokena navedeni model će najverovatnije uvek glasati za pogrešnu klasu. Ukoliko se u ansamblu nalazi XML RoBERTa Base i njegova pre-trenirana varijanta (PT – XML RoBERTa Base), prilikom glasanja dva modela će pogrešno glasati, radi istog problema sa modelom, i time nagnuti ansambl ka pogrešnoj klasi. Konačni klasifikatori iskorišćeni za Ensemble Best model su: CRF, LM-LSTM-CRF, PT-BERT Multilingual Cased, PT-XLM RoBERTa Base, XLM RoBERTa Large.

ANSAMBL	MODELI U ANSAMBLU
ENSEMBLE ALL	CRF
	LM-LSTM-CRF
	T – RoBERTa
	BERT Multilingual Cased
	BERT Multilingual Uncased
	XML RoBERTa Base
	XML RoBERTa Large
	PT – BERT Multilingual Cased
	PT – XML RoBERTa Base
ENSEMBLE BEST	CRF
	LM-LSTM-CRF
	PT – BERT Multilingual Cased
	PT – XML RoBERTa Base
	XML RoBERTa Large

Tabela 12. Spisak modela po ansamblu

6. Eksperimentalni rezultati

Eksperimentalni rezultati, odnosno performanse iskorišćenih modela date su u ovom poglavlju. Prvo je dat opis evaluacije performansi modela. Nakon čega, kao poslednji deo ovog poglavlja dat je detaljan opis rezultata.

6.1. Eksperimentalna postavka za evaluaciju modela

U cilju formiranja sistema za prepoznavanje imenovanih entiteta na medicinskim dokumentima napisanim na srpskom jeziku, izvršena je evaluacija prethodno navedenih modela mašinskog učenja.

Modeli su evaluirani sa „hold-out“ metodom gde je 20% anotiranog korpusa iskorišćeno za test skup a 80% za obučavanje.

Kao što je navedeno, CRF model je zahtevao ručno formiranje atributa, dok su ostali modeli zahtevali korake pre-treniranja. Pre obučavanja LM-LSTM-CRF modela prvo su pre-trenirani vektori značenja reči koje je model koristio prilikom obučavanja (sekcija 5.3.3). Za potpunu procenu performansi transformer modela, izvršena je njihova evaluacija bez pre-treniranja i sa pre-treniranjem. Pre-treniranje je izvršeno nad skupom od 17000 neanotiranih dokumenata (sekcija 5.3.1), odnosno nad dokumentima koji nisu deo obučavajućeg i test skupa.

Nastavak zadatka pre-treniranja XML RoBERTa Large modela nije izvršen zbog nekompatibilnosti modela sa hardverskim specifikacijama obučavajućeg računara, tačnije količina RAM memorije potrebna u toku obučavanja modela je bila veća od raspoložive RAM memorije na obučavajućem računaru.

Za evaluaciju performansi modela odabrane su metrike preciznost, odziv i F1 mera koje se smatraju standardom za evaluaciju klasifikacionih modela, pogotovo kada klase nisu ravnomerno zastupljene. Dodatno, u medicinskom domenu želimo da dobro detektujemo probleme (bolesti), odnosno da imamo visoku vrednost odziva, ali istovremeno ne želimo da imamo lažno pozitivne rezultate, odnosno visoku preciznost. Imajući u vidu da se navedene metrike određuju za svaku klasu zasebno, jedinstvena vrednost ovih metrika za modele je dobijena uz pomoć mikro proseka (sekcija 2.6). Uzimajući u obzir da prepoznavanje imenovanih entiteta nije trivijalan zadatak gde jedan entitet može da se sastoji od više tokena

rezultati su predstavljeni za tačno poklapanje imenovanih entiteta i za poklapanje na nivou tokena (sekcija 2.7). Za evaluaciju rezultata za tačno poklapanje, kao i rezultata za poklapanje na nivou tokena, izabrana je metrika mikro preciznosti kako bi ostvareni rezultati bili uporedivi sa srodnim istraživanjima. Rezultati mikro prosečne vrednosti preciznosti, odziva i F1 mere su dati u tabeli 13.

6.2. Preciznost, odziv i F1 mera na nivou modela

CRF model je imao najbolju preciznost, dok je najbolji odziv imao PT-BERT Multilingual Cased model. Sveukupno, kao najbolji model pokazao se LM-LSTM-CRF model sa najboljom vrednošću za F1 meru. Model koji je imao najlošije performanse je T-RoBERTa model, transformer model čiji je jezički model formiran samo na osnovu neanotiranih medicinskih dokumenata. Kada se T-RoBERTa model uporedi sa sličnim model obučenim na medicinskim podacima, BioBERT modelom (J. Lee et al., 2020) loše performanse T-RoBERTa modela se mogu se pripisati veličini korpusa koji je korišćen za formiranje jezičkog modela. Preciznije jezički model T-RoBERTa je formiran na korpusu od 8.2 miliona reči dok je jezički model BioBERT modela, i sličnih modela, obučen na korpusu od 4.5 milijardi reči.

Model	***Preciznost***	***Odziv***	***F1 mera***
CRF	**0.890**	0.845	0.867
LM-LSTM-CRF$^{(PT)}$	0.879	0.884	0.882
T – RoBERTa	0.706	0.733	0.719
BERT Multilingual Cased	0.834	0.867	0.850
BERT Multilingual Uncased	0.819	0.862	0.840
XLM RoBERTa Base	0.767	0.823	0.794
XLM RoBERTa Large	0.849	0.876	0.862
PT – BERT Multilingual Cased$^{(PT)}$	0.867	0.886	0.876
PT – XLM RoBERTa Base$^{(PT)}$	0.785	0.840	0.812
Ensemble All	0.874	0.890	0.881
Ensemble Best	0.889	**0.895**	**0.892**

Tabela 13. Preciznost, odziv i F1 mera za tačno poklapanje entiteta. Modeli obeleženi sa $^{(PT)}$ su pretrenirani na skupu od 17000 neanotiranih dokumenata.

Kako bi se procenile opšte performanse modela, u tabeli 14 su prikazane vrednosti za preciznost, odziv i F1 meru modela na nivou tokena (sekcija 2.7). Rezultati na nivou tokena pokazuju da CRF model i dalje ima najbolju preciznost, dok najbolji odziv ima PT-XML RoBERTa Base, a

najbolje performanse je imao PT-BERT Multilingual Cased model sa najboljom vrednosti za F1 meru. Transformer modeli kojima je do-obučen jezički model (PT-BERT Multilingual Cased i PT-XLM RoBERTa Base modeli) imaju za nijansu bolje rezultate F1 mere od LM-LSTM-CRF modela na nivou tokena, dok je obrnuta situacija na nivou entiteta.

Poput rezultata u (Virtanen et al., 2019), (Dumitrescu et al., 2020) i (Koutsikakis et al., 2020), predstavljeni rezultati pokazuju da BERT Multilingual Cased model ima bolje performanse u odnosu na model koji ne pravi razliku između malih i velikih slova. Iz rezultata se takođe vidi da nastavak zadatka pre-treniranja, odnosno nastavak obučavanja jezičkog modela, transformer modela na neanotiranom korpusu rezultuje povećanjem performansi modela i na nivou tokena i za tačno poklapanje entiteta.

Model	***Preciznost***	***Odziv***	***F1 mera***
CRF	0.892	0.834	0.862
LM-LSTM-CRF	0.878	0.868	0.873
T – RoBERTa	0.852	0.830	0.841
BERT Multilingual Cased	0.856	0.858	0.857
BERT Multilingual Uncased	0.858	0.852	0.855
XLM RoBERTa Base	0.845	0.855	0.849
XLM RoBERTa Large	0.876	0.864	0.870
PT – BERT Multilingual Cased	0.873	0.879	0.876
PT-XLM RoBERTa Base	0.865	0.884	0.874
Ensemble All	0.898	0.886	0.893
Ensemble Best	**0.906**	**0.893**	**0.899**

Tabela 14. Preciznost, odziv i F1 mera za modele na nivou tokena

Jedna od strategija za poboljšanje klasifikacionih rezultata je upotreba ansambla (sekcija 2.5). Dva ansambl modela su formirana strategijom većinskog glasanja, Ensamble All koji kombinuje sve prethodno obučene klasifikacione modele i Ensamble Best u kome su isključeni osnovni transformer modeli koji imaju pre-treniranu varijantu (sekcija 5.3.5). Rezultati preciznosti, odziva i F1 mere za ansambl modela (Ensemble All i Ensemble Best) za tačno poklapanje entiteta dati su u tabeli 13, u tabeli 14. su prikazani njihovi rezultati na nivou tokena. Ensemble Best model je rezultovao sa najboljim performansama i za tačno poklapanje entiteta i za poklapanje na nivou tokena. U rezultatima se vidi poboljšanje od oko 1% za tačno poklapanje entiteta i oko 2.5% za poklapanje na nivou tokena,

kada se Ensemble Best model uporedi sa prethodno najboljim modelima po navedenim strategijama poklapanja entiteta i tokena.

Klasa	***Preciznost***	***Odziv***	***F1 mera***
CLINICAL_DEPT	0.784	0.833	0.808
EVIDENTIAL	0.835	0.881	0.857
OCCURRENCE	0.813	0.769	0.790
PROBLEM	0.730	0.721	0.725
TEST	0.926	0.944	0.935
TREATMENT	0.856	0.872	0.864
DATE	0.950	0.957	0.953
DURATION	0.931	0.960	0.945
FREQUENCY	0.947	0.975	0.961
TIME	0.750	1.0	0.857
VALUE	0.927	0.947	0.937

Tabela 15. Performanse Ensemble Best model za tačno poklapanje entiteta klasa

Klasa	***Preciznost***	***Odziv***	***F1 mera***
CLINICAL_DEPT	0.835	0.867	0.850
EVIDENTIAL	0.871	0.898	0.884
OCCURRENCE	0.882	0.746	0.808
PROBLEM	0.841	0.786	0.812
TEST	0.934	0.940	0.937
TREATMENT	0.913	0.890	0.902
DATE	0.971	0.967	0.969
DURATION	0.934	0.966	0.950
FREQUENCY	0.946	0.983	0.964
TIME	0.889	1.0	0.941
VALUE	0.953	0.966	0.960

Tabela 16. Performanse Ensemble Best model na nivou tokena

Sa obzirom na to da je najbolju mikro prosečnu F1 meru imao Ensemble Best model u nastavku su prikazani njegovi rezultati za svaku od klasa entieta. Rezultati Ensemble Best modela na nivou klasa predstavljeni su u tabeli 15. i tabeli 16. Najbolje vrednosti za F1 meru su postignuti za jednostavne numeričke vrednosti TIMEX3 tagova (DATE, DURATION, FREQUENCY i TIME) i VALUE tagova. Kod kompleksnijih EVENT tagova može se primetiti da su najbolje performanse postignute nad dobro

definisanim i dobro zastupljenim klasama poput TEST i TREATMENT klase. Dok se kod klasa sa najvećim stepenom jezičke dvosmislenosti, poput OCCURRENCE i PROBLEM klase, primećuju lošiji rezultati.

Rezultati za modele CRF, LM-LSTM-CRF, T-RoBERTa, BERT Multilingual Cased, BERT Multilingual Uncased, XLM RoBERTa Base, XLM RoBERTa Large, PT – BERT Multilingual Cased, PT – XLM RoBERTa Base za tačno poklapanje entiteta po klasama i za rezultate na nivou tokena prikazani su u Prilogu A i B.

7. Diskusija

Prepoznavanje imenovanih entiteta je ključan zadatak prilikom analize teksta i pred zadatak za mnoge zadatke mašinskog učenja. Razvoj novih savremenih NLP modela kontinuirano poboljšava lakoću korišćenja i tačnost prethodno korišćenih pristupa. Trenutno dominantini pristupi koji ostvaruju vrhunske rezultate za prepoznavanje imenovanih entiteta su zasnovani na transformer arhitekturi (Gaschi et al., 2023), dok se klasični pristupi poput CRF modela obično koriste kao osnova za poređenje rezultata. U okviru ovog istraživanja evaluirana je mogućnost primene trenutno najsavremenijih modela, kao i potrebnih koraka za njihovu primenu, za razvoj novog modela za klinički NER na srpskom jeziku. Pre svega za razvoj su upoređene performanse NER sistema baziranih na CRF, rekurentnim neuronskim mrežama (LSTM), transformerima (BERT i RoBERTa) i njihovim ansamblom.

Najbolji pristup na nivou tokena kao i za tačno poklapanje entiteta je kombinacija više klasifikatora u ansambl sa većinskim glasanjem, odnosno Ensemble Best model. Za tačno poklapanje entiteta Ensamble Best model je postigao vrednost F1 mere od 0.892 što je uporedivo sa IAA F1 merom sa vrednosti od 0.904. Na osnovu sličnosti rezultata modela i saglasnosti anotatora, može se zaključiti da je obučeni model dostigao performanse anotatora i da se može iskoristiti kao deo kliničkih sistema kao: podrška sistemima za odlučivanje, za opservacijske studije, za poređenje pacijenata i za zadatke koje zavise od NER rezultata.

Ostvareni rezultati su uporedivi sa rezultatima savremenih modela koji postižu vrhunske rezultate nad sličnim skupovima podataka (Tabela 17). Za prikaz performansi izabrani su savremeni modeli (SciBERT (Beltagy et al., 2019) , BioBERT (J. Lee et al., 2020), ClinicalBERT (Alsentzer et al., 2019), BioMed-RoBERTa (Gururangan et al., 2020), PTLMBCT (Lewis et al., 2020), kao i u radu (Lewis et al., 2020), koji su prijavljivali vrhunske performanse nad javno dostupnim skupovima podataka za medicinski NER.

Iz navedene tabele se vidi da su performanse Ensamble Best modela na nivou performansi savremenih modela. Odstupanja modela između različitih skupovima podataka zavisi od samog NER problema koji je obuhvaćen skupom podataka i kvaliteta anotacija u navedenom skupu (saglasnosti anotatora):

- BC5CDR-Disease skupu

  - zadatak je prepoznavanje naziva bolesti u naučnim člancima
  - prosečna saglasnost anotatora - 0.875
- i2b2-2010 skup
  - zadatak prepoznavanja lekova i doza u kliničkim dokumentima
  - prosečna saglasnot anotatora - 0.924
- i2b2-2012 skup
  - zadatak obuhvata kliničke događaje, temporalne relacije i vremenske događaje (detaljni opis i2b2-2012 izazova je dat u sekciji 4.)
  - prosečna saglasnost anotatora – 0.780

Skup / Model	BC5CDR-Disease (Engleski jezik)	i2b2-2010 (Engleski jezik)	i2b2-2012 (Engleski jezik)	MEDRS[18] (Srpski jezik)
SciBERT	0.836	0.863	0.776	-
BioBERT	0.833	0.867	0.776	-
ClinicalBERT	0.813	0.863	0.780	-
BioMed-RoBERTa	0.806	0.850	0.764	-
PTLMBCT	**0.838**	**0.881**	**0.795**	-
Ensamble Best	-	-	-	**0.892**

Tabela 17. F1 mere savremenih modela za medicinski NER nad reprezentativnim skupovima podataka.

Osnovni zaključak prisutan u radovima za sve gore navedene savremene modele je da upotreba velikih količina podataka za zadatak pre-treniranja pozitivno utiče na performanse tih modela na NER zadatku. Dodatno se može zaključiti i da je kvalitet anotacija jako bitan prilikom obučavanja modela.

Na osnovu ostvarenih može se zaključiti da se vrhunski rezultati za medicinski NER mogu postići, sa savremenim više-jezičkim modelima, i na srpskom jeziku što ranija istraživanja nisu pokazala. Primena savremenih modela pokazuje da nije potrebno ručno definisanje semantičkih i sintaksnih pravila za prepoznavanje imenovanih entiteta nad

[18] MEDRS se odnosi na korpus dokumenata anotiranih u sklopu ove disertacije.

medicinskim dokumentima, već se uspešno mogu koristiti savremeni modeli zasnovani na dubokom učenju.

Rezultati prikazani u tabeli 15 i tabeli 16 pokazuju da je Enamble Best model imao najbolje performanse sa dobro definisanima klasama, odnosno sa klasama koje u sebi nemaju dvosmislene ili kontekstno zavisne termine poput klasa FREQUENCY, DURATION, VALUE i TEST. Sa klasama koje u sebi imaju kontekstno zavisne termine poput klase OCCURRENCE i PROBLEM model je imao lošije performanse, što je i vidljivo po razlici rezultata za tačno poklapanje i rezultata na nivou tokena (Prilog A. i B.) i na osnovu analize grešaka (sekcija 7.3). Slična odstupanja rezultata na nivou klasa su vidljiva i kod ostalih modela samo sa nižim vrednostima metrika u odnosu na Enasmble Best. Na osnovu rezultata za klasu TIME može se zaključiti da i zastupljenost (broj primera) klase u obučavajućem skupu ima određeni uticaj na performanse modela. Iz rezultata se vidi da model uvek obeleži dobro sve primere klase TIME (visoka vrednost odziva) ali termine koji pripadaju drugim klasama pomeša sa klasom TIME (vrednost preciznosti), i kao što je uočeno u sekciji 4.3 model ima lošije rezultate za tačno poklapanje entiteta.

7.1. Prednosti i mane modela dubokog učenja

Modeli zasnovani na dubokom učenju su imali bolje rezultate za odziv i F1 meru, dok je CRF model (klasičan model mašinskog učenja) imao najbolje rezultate preciznosti.

Preciznost CRF modela može se objasniti izabranim atributima koji su ograničili CRF model da klasifikuje samo dobro definisane entitete koje nemaju varijabilnosti i kontekstne zavisnosti što je dovelo do visoke vrednosti preciznosti, dok su modeli dubokog učenja imali više fleksibilnosti što ih je omogućilo da budu tačniji što se odražava u visokim vrednostima odziva. To se može videti na primeru rečenice „sternalna punkcija ukazala na toksicno ostecenje sa sa ocuvanom celularnoscu“ gde je „toksicno ostecenje“ obeleženo kao PROBLEM od strane anotatora. Termin „toksicno ostecenje“ se ne pojavljuje u obučavajućem skupu, dok se reč „toksicno“ pojavljuje samo tri puta (osam puta ukoliko se uključi i reč „nefrotoksicno“) u obučavajućem skupu dok se reč „ostecenje“ pojavljuje sedam puta. U navedenom primeru CRF nije obeležio ni jednu reč kao entitet klase PROBLEM, LSTM model je obeležio samo reč „ostecenje“, a transformer modeli su uspešno obeležili ceo termin kao entitet klase PROBLEM. Jedino odstupanje koje se može primetiti sa CRF

rezultatima (Prilozi A. i B.) je u tome što CRF uspešno obeležava sve primere klase TIME, dok ostali modeli imaju problema sa prepoznavanjem entiteta te klase. To se može objasniti sa izborom atributa „oblika reči“ (sekcija 5.3.2) koji omogućava CRF modelu da sa malim brojem primera razgraniči TIME klasu od ostalih klasa.

Kada se uporede rezultati modela dubokog učenja, jedini modeli koji značajno odstupaju od ostalih modela, po rezultatima tačnog poklapanja entiteta, su osnovni RoBERTa modeli (XML RoBERTa Base i PT – XLM RoBERTa Base). Razlike u performansama osnovnih RoBERTa modela se mogu objasniti poteškoćom tih modela u procesiranju numeričkih vrednosti (Thawani et al., 2021). Odnosno, ograničenjem BPE tokenizera (sekcija 2.4.4) koji prilikom obrade numeričkih vrednosti numeričke tokene rastavi na pod-tokene na nekonzistentan način, npr. broj 1991 u jednom primeru nasumično predstavlja kao 1-991, a broj 1992 kao 19-92, što je dodatno potvrđeno sa značajno boljim rezultatima na nivou tokena (Prilog A. i B.).

Poput autora (Ramos-Flores et al., 2020), na osnovu rezultata istraživanja može se doći do zaključka da je CRF model pogodniji za primenu na manjim skupovima podataka od modela dubokog učenja, ukoliko nisu dostupni adekvatni korpusi nad kojima se može izvršiti pre-treniranje. Dodatno eksperimentalni rezultati su pokazali da modeli dubokog učenja, poput transforemera i RNN, sa adekvatnim koracima pre-treniranja mogu da budu precizniji od CRF modela na relativno malim skupovima podataka kao što je skup od 203 anotirana dokumenta.

7.2. Uticaj pre-treniranja na modele dubokog učenja

Bitno je napomenuti da je za pre-treniranje potrebno da bude dostupna veća količina podataka na kojima se može vršiti pre-treniranje. Autori (J. Lee et al., 2020) su pokazali važnost pre-treniranja BERT modela na biomedicinskom korpusu za poboljšanje rezultata NER zadatka. U svojim eksperimentima došli su do zaključka da povećanje veličine korpusa dovodi do povećanja performansi na NER zadatku do optimalne tačke od oko 4.5 milijardi reči u korpusu. Slično njima, u okviru ovog istraživanja, pokazano je da je pre-treniranje transformer modela (sekcija 2.4.4) i prethodno obučavanje vektora značenja reči za LM-LSTM-CRF model (sekcija 2.3) na biomedicinskom korpusu od 8.2 miliona reči bitan korak koji povećava krajnje performanse za NER zadatak. Sa pre-treniranjem F1 mera za BERT model je 2.5% veća od modela koji nije pre-treniran (Prilog

A.), dok LSTM model nije bilo moguće obučiti bez prethodnog obučavanja vektora značenja reči. Dodatno je pokazano da je za obučavanje novog jezičkog modela, kako bi se dobili konkurentni rezultati, potreban značajno veći korpus od korpusa sa 8.2 miliona reči.

U domenu biomedicinske analize teksta, teoretisano je da su bidirekcione reprezentacije, koje transformer modeli uče u toku obučavanja, ključne zbog kompleksnih relacija koje postoje između biomedicinskih termina (J. Lee et al., 2020). Te biomedicinske reprezentacije mogu objasniti zašto su individualno najbolje rezultate odziva, na nivou tokena, postignute od strane transformer modela. Rezultati takođe pokazuju da individualno ne postoji model koji je značajno bolji od drugih i iz analize grešaka može se zaključiti da različiti tipovi modela imaju potencijal da dopunjuju jedni druge. Iz tog razloga modeli su kombinovani u ansambl sa većinskim glasanjem koji je proizveo najbolje rezultate. Na primer CRF model dobro prepoznaje tokene klase DURATION (Tabela B.1.) dok XLM RoBERTa Large dobro prepoznaje tokene klase FREQUENCY (Tabela B.7.), njihovom kombinacijom dobijamo model koji može dobro da prepoznaje oba klase tokena.

7.3. Analiza grešaka

Slika 28. Osnovna podela grešaka sistema.

Analiza grešaka je izvršena nad rezultatima modela sa najbolji performansama, Ensemble Best modela, za tačno poklapanje entiteta. Većina grešaka, oko 72%, su lažno negative greške (Slika 28). Radi detaljnije analize, LN greške su podeljene u dve grupe: neobeležene entitete i netačno obeležene entitete. Neobeleženi entiteti su anotirani entiteti koje model uopšte nije prepoznao, a od ukupnog broja LN grešaka 33% su neobeleženi entiteti. Netačno obeleženi entiteti predstavljaju sve ostale LN greške poput anotiranih entiteta kojima modeli nije prepoznao tačne granice, odnosno u kojima prepoznat entitet nema tačno preklapanje, i obeleženih entiteta sa pogrešno identifikovanim tipom.

Tip greške	**Primer**					
Neobeležen entitet	C	lecenje		u		barokomori
	T	B-TREA..		I-TREA..		I-TREA..
	P	O		O		O
Netačne granice	C	suva	gangrena	II	i	III
	T	B-PROB..	I-PROB..	O	O	O
	P	B-PROB..	I-PROB..	I-PROB..	O	O
Netačana klasa	C	kolegijum	NFK	prekida	PD	
	T	B-OCCU..	I-OCCU..	I-OCCU..	I-OCCU..	
	P	O	I-CLIN..	O	O	

Tabela 18. Primeri lažno negativnih grešaka. Skraćenice: C – kontekst greške, T – očekivane klase, P – klase dodeljene od strane modela, TREA - TREATMENT, PROB – PROBLEM, OCCU – OCCURRENCE, CLIN – CLINICAL_DEPT

Među neobeleženim entitetima kao glavni uzročnik greške identifikovani su termini koji su nedovoljno zastupljeni u korišćenom korpusu sa udelom od ~53% netačno obeleženih entiteta. Od nedovoljno zastupljenih termina, ~16% su termini koji se pojavljuju samo u test skupu (poput termina „pneumobilija“) dok se ostali u proseku pojavljuju manje od pet puta u celom korpusu. Iako modeli mašinskog/dubokog učenja imaju dobru sposobnost generalizacije, za sam proces postizanja te generalizacije je potreban adekvatan broj primera. Na primer, modeli dubokog učenja imaju sposobnost da dobro klasifikuju primere koje nikada nisu videli (*few shot*) ali to zavisi od konteksta navedenih klasa.

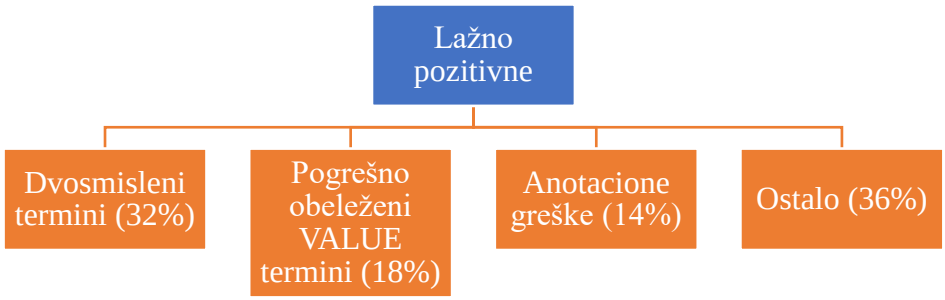
Odnosno, ukoliko je model obučen za prepoznavanje osoba i u primerima na kojima je obučavan se pojavljuje forma „Ime Prezime je CEO kompanije“ algoritam će na osnovu konteksta „je CEO kompanije“ moći da odredi ime i prezime osobe, koje nije mogao da vidi u obučavajućem skupu, zapravo pripada klasi osoba. Ali ukoliko ne postoji kontekst koji ukazuje određenu klasu, kao što može da se javi u slučaju malog broja kliničkih tekstova, onda modeli neće moći da dobro generalizuju. Problem nedovoljno zastupljenih termina, u korpusu dokumenata korišćenom u okviru ove disertacije, je naglašen zbog prirode medicinskog domena gde se terminologija razlikuje na osnovu specijalizacije klinika. Na primer, pacijent upućen na kliniku za nefrologiju radi dijalize, sa klinike za neurohirurgiju, će u pratećem izveštaju imati terminologiju koja nije uobičajena za kliniku za nefrologiju.

Jedna od kategorija grešaka identifikovanih prilikom analize neobeleženih entiteta, sa ~6% neobeleženih grešaka, predstavljaju vrednosti VALUE entiteta, vrednosti koje su obično dobro prepoznate u drugim delovima test skupa, ali koji se ovaj put nalaze odmah nakon neobeleženih EVENT entiteta sa kojima su povezani. Na primer, vrednost „500 ml“ koja se navodi kao vrednost u terapijama i preporukama za unos tečnosti je obično dobro klasifikovana. Ali u situacijama kada terapija ili preporuka, za koju se navodi vrednost, nije prepoznata u određenim primerima dolazi do greške sa obeležavanjem vrednosti. Ova kategorija greške ukazuje da je model identifikovao zavisnosti između VALUE i EVENT entiteta.

Od netačno obeleženih entiteta, 76% su entiteti čiji tip je tačno obeležen ali nemaju tačno poklapanje sa anotiranim entitetima (Tabela 18.). Četvrtina grešaka u kojima model nije prepoznao tačne granice entiteta sadrže dvosmislene (kontekstno višeznačne) termine koji nisu uvek anotirani i pogrešno obeležene interpunkcijske znakove. Kao primer kontekstno višeznačnih reči anotatori su reč „odgovarajućom“ anotirali kada se nalazi u kontekstu „*odgovarajućom simptomatskom terapijom*“, dok kada se nalazi pored konkretnog naziva tretmana poput „odgovarajuća terapija primenom *Vancogalom* per os“ je anotirana samo terapija od interesa (podvučene reči u primerima su prepoznati entiteti). Reprezentativan primer za pogrešno obeležene interpunkcijske znakove je PROBLEM entitet “Effusio pleurae bilat.” za koji je model nije obeležio pripadajući interpunkcijski simbol kao deo entiteta[19].

Veliki broj grešaka u kome model nije obeležio tačan tip entiteta se javljao sa dvosmislenim tipovima entiteta, odnosno sa PROBLEM i OCCURRENCE entitetima. Jedan od primera takvog tipa greške sa OCCURRENCE entitetom može se videti u entitetu “*zapocinjanja lecenja hemodijalziama*”, koji sugeriše da je pacijent primljen primljen u bolnicu radi lečenja hemodijalizama, model je prepoznao samo „*hemodijalziama*“ kao TREATMENT entitet. Jedan od potencijalnih uzroka zbog kojeg model ima problema sa kontekstno zavisnim primerima je u kompleksnim kontekstnim zavisnostima medicinskih dokumenta i u relativno malom obučavajućem skupu.

[19] Anotator je obeležio interpunkcijski simbol kao deo entiteta jer u konkretnom slučaju „bilat.“ je skraćenica za „bilateralis“.



Slika 29. Podela lažno pozitivnih grešaka.

Većina lažno pozitivnih grešaka (LP, Slika 29), oko 32%, su takođe izazvana od strane kontekstno zavisnih (dvosmislenih) termina i fraza. Dodatnih 18% grešaka su prouzrokovana od strane teksta koji je obično asociran sa VALUE entitetima, a koji zapravo predstavlja neku drugu medicinsku informaciju (npr. „147 mm“ kada se koristi u kontekstu veličine parijetalnog režnja koji ne pripada konkretnom EVENT entitetu te se ne obeležava sa VALUE entitetom).

Potrebno je naglasiti da oko 14% grešaka su greške prouzrokovane anotacionim greškama ili preferencijama anotatora, što je i prikazano prilikom provere saglasnosti anotatora. Na osnovu provere saglasnosti anotatora utvrđeno je da anotatori imaju IAA od 0.904 te je očekivano da razlika od 0.096 unese određeni nivo greške u model i ne može se očekivati da model ima bolje performanse od vrednosti IAA. Greške anotatora nisu dovoljno učestale da bi se mogle klasifikovati u kategorije i odnose se na pojedinačne primere. Kao konkretan primer anotacije koja je svrstane kao problem sa anotiranjem je u rečenici „Visegodisnji je dijabeticar na OAD, a sada na odgovarajucem rezimu ishrane.“. U navedenoj rečenici anotator je iz konteksta rečenice obeležio „dijabeticar na OAD“ kao PROBLEM zbog dugogodišnjeg uzimanja lekova sa potencijalnim uticajem na progresiju hronične bubrežne insuficijencije, a model je prepoznao da je „dijabeticar“ entitet tipa PROBLEM a „OAD[20]“ entitet tipa TREATMENT. Bitno je napomenuti da se u korpusu termin „OAD“ pojavljuje samo pet puta i u četiri primera se koristi kao aktivna terapija koju pacijent koristi te je i anotiran kao TREATMENT u tim primerima. Prilikom računanja grešaka sistema, za navedeni primer, greška je računata dva puta: jednom za netačne granice entiteta PROBLEM i jednom za pogrešno obeležen entitet TREATMENT.

[20] OAD – Oralni antidijabetici (*Oral Antidiabetic Drug*)

7.4. Ograničenja

Adaptiranje savremenih NLP modela u jezike poput srpskog jezika, koji imaju manjak dostupnih leksičkih resursa (Avdic et al., 2020), i dalje može biti težak zadatak. Jedna poteškoća prouzrokovana nedostatkom leksičkih resursa u okviru ovog istraživanja je nedostatak tagova vrste reči. Tagovi vrste reči su često korišćen atribut za NER sa CRF modelom, i oni mogu da poboljšaju performanse modela (Nayel & Shashirekha, 2017). Zbog nedostatka adekvatno dostupnih alata za tagovanje i lematizacaciju, u toku trajanja ovog istraživanja, nisu korišćeni tagovi za vrste reči ni leme reči.

Kao najveće ograničenje ove studije se može smatrati veličina i raznovrsnosti korišćenog korpusa podataka.

Korpus neanotiranih dokumenata koji je iskorišćen za zadatak pre-treniranja je relativno mali kada se uporedi sa korpusima korišćenim u relativno novijim studijama poput, BioBERT-a i BioMed-a. Sa većom količinom podataka, korišćenim u zadatku pre-treniranja, model bi imao više kontekstnih primera iz kojih bi mogao bolje da nauči međusobne zavisnosti reči i fraza što bi dodatno pomoglo sa performansama kod kontekstno dvosmislenih pojmova. Dodatno, sa povećanjem veličine korpusa smanjuje se broj retkih termina koji mogu da predstavljaju određeni vid šuma u sistemu. Na osnovu analize grešaka sistema, primećeno je da su među čestim greškama modela greške izazvane od strane opštih medicinskih pojmova koji nisu bili dovoljno zastupljeni u relativno malom skupu podataka koji je korišćen. Dok je očekivano da u kompleksnim poljima postoje retki termini, povećanje veličine skupa podataka može da smanji količinu nedovoljno zastupljenih opštih termina i time da se smanji greška koju oni unose u sistem.

Veličina korpusa zlatnog standarda, odnosno anotiranih dokumenata, ima veliki uticaj na performanse modela za prepoznavanja imenovanih entiteta. Povećanje broja anotiranih primera omogućava modelu da bolje nauči karakteristike različitih klasa entiteta i time poboljša kvalitet klasifikacija. Ključna prepreka za povećanje broja anotiranih primera je dostupnosti resursa. Anotiranje je izazovan zadatak koji zahteva dosta vremena i angažovanje domenskih eksperta.

Istraživanje u okviru ove studije je pokazalo da se, za zadatak prepoznavanja imenovanih entiteta u medicinskim dokumentima na srpskom jeziku, mogu koristiti savremeni modeli dubokog učenja koji konstantno napreduju i ostvaruju sve bolje rezultate. Kvalitet više-jezičkih modela, kao i njihove performanse na različitim jezicima, se konstantno poboljšavaju te nedostatak u veličinama korpusa na specifičnim jezicima, poput srpskog jezika, u budućnosti neće biti toliko veliki problem zbog mogućnosti upotrebe korpusa iz više različitih jezika.

Dodatno, raznovrsnost korišćenih podataka je jedan od faktora koji utiče na sposobnost generalizacije modela dubokog učenja. U okviru ovog istraživanja korišćeni su dokumenti sa jedne klinike. Obično lekari u jednoj klinici pišu kliničke dokumente na način specifičan za tu kliniku, što znači da se struktura dokumenata između različitih klinika, a posebno različitih medicinskih ustanova, može značajno razlikovati. Iako metode dubokog učenja imaju jako dobre sposobnosti generalizacije, promena strukture dokumenata u odnosu na strukturu sa kojom je model obučen može da rezultuje smanjenjem klasifikacionih performansi. Trenutno ne postoje univerzalni modeli (takozvani *zero shot learning* modeli) koji bi mogli da se obuče po formatu iz jedne klinike i da uspešno vrše klasifikaciju u dokumentima koji su napisani po formatu iz drugih ustanova.

Prilikom obučavanja modeli dubokog učenja zahtevaju dosta računarski resursa u vidu procesorskih jezgara, koji su obično u vidu grafičkih procesorskih jedinica (GPU) ili tenzorskih procesorskih jedinica (TPU) sa odgovarajućom količinom RAM memorije. Na primer, BERT Large model je obučavan 4 dana na 16 Cloud TPU akceleratora[21], pri čemu jedan akcelerator se sastoji od 4 TPU čipa i 32 GB RAM memorije. Korišćenje navedenog hardvera je obično finansijski nepristupačno za projekte sa manjim budžetom. Dodatno ograničenje ovog istraživanja je u vidu korišćenog hardvera (sekcija 5.1) koji nije imao dovoljno RAM memorije zahtevan od XML RoBERTa Large modela za nastavak zadatka pre-treniranja. Prilikom nastavka zadatka pre-treniranja prvo je izvršena evaluacija pre-treniranja BERT Multilingual Cased modela sa dužinom trajanja od osam sati, dva dana i četiri dana. Razlike u krajnjim performansama modela koji su pre-trenirani dva i četiri dana nisu bile

[21] Cloud TPU akcelerator je specijalizovan hardver razvije od strane kompanije Google namenjen za ubrzanje obučavanja modela dubokog učenja.

primetne, tako da je dužina od dva dana izabrana za pre-treniranje ostalih modela.

Nakon što je završeno ovo istraživanje javno su objavljeni veliki jezički modeli (*large language models*, LLM) poput GPT-2 modela (Radford et al., 2019), koji predstavlja proširenje GPT modela (sekcija 2.4.4) u vidu količine slojeva u arhitekturi čime se povećao broj obučavanih parametara sa 117 miliona na 1542 miliona. Nakon uspeha GPT-2 modela, dolazi do pojave modela sa nekoliko milijardi obučavajućih parametara poput LLaMA modela (Touvron et al., 2023) sa varijantama od 7 do 65 milijardi parametara i GPT-3 modela (Brown et al., 2020) sa 175 milijardi parametara. U poslednje vreme, LLM modeli su postali dominanti u određenim NLP zadacima poput generisanja, sumarizacije i razumevanja teksta. Efektivnost njihove primene, za zadatak prepoznavanja imenovanih entiteta, još nije dovoljno istražena.

8. Zaključak

Osnovna tema ove doktorske disertacije je sistem za automatsko prepoznavanje imenovanih entiteta u medicinskim dokumentima na srpskom jeziku. Automatsko prepoznavanje imenovanih entiteta je jedan od ključnih zadataka oblasti obrade prirodnog jezika. U okviru ove disertacije, prepoznavanje imenovanih entiteta odnosi se na prepoznavanje medicinski relevantnih termina iz tekstova napisanih od strane lekara u kliničkim uslovima. Klinički relevantni termini su kategorizovani u tri osnovne kategorije: klinički događaji (EVENT), vremenske odrednice (TIMEX3) i vrednosti događaja (VALUE). Metodologija istraživanja se bazira na metodama mašinskog učenja i dubokog učenja koje su primenjena za razvoj sistema.

Prototip sistema za prepoznavanje imenovanih entiteta je evaluiran na ručno anotiranom korpusu. Anotirani korpus se sastoji od klinički tekstova sa Klinike za nefrologiju Univerzitetsko kliničkog centra Srbije. Ukupno je anotirano 203 dokumenta od strane dva anotatora.

Prilikom pregleda relevantne literature za medicinski NER na srpskom jeziku, ustanovljeno je da nisu korišćeni savremeni pristup dubokog učenja koje ne zavise od eksperata za formiranje rečnika i pravila. Predstavljeni rezultati evaluacije sistema pokazuju da upotreba savremenih metoda dubokog učenja daje obećavajuće performanse u prepoznavanju imenovanih entiteta, koji su na nivou rezultata anotatora. Iz ostvarenih rezultata može se napraviti pozitivan zaključak za polaznu hipotezu ovog istraživanja, odnosno moguće je iskoristi savremene modele dubokog učenja za prepoznavanje imenovanih entiteta na srpskom jeziku. Dodatno se može zaključiti da prototip sistem, na osnovu rezultata, ne samo da omogućava direktno korišćenje medicinskih podataka i uvid u semantiku medicinskih dokumenata, već predstavlja i osnovu koja se može koristi za razvoj složenijih alata koji bi dodatno mogli unaprediti kvalitet kliničke prakse i istraživanja.

Ograničenja ovog istraživanja su predstavljena u poslednjoj sekciji sedmog poglavlja. Veličina neanotiranog korpusa podataka kao i anotiranog korpusa zlatnog standarda se ističe kao najveće ograničenje ovog istraživanja. Povećanjem veličine korpusa obično dovodi do opšteg poboljšanja performansi modela dubokog učenja, što je i navedeno kao pretpostavka u diskusiji ograničenja. Greške prikazane u analizi su najčešće vezane za kontekstno zavisne entitete čiji kontekst sistem nije

mogao da nauči zbog ograničene količine dostupnih podataka. Otklanjanjem navedenog ograničenja, može se samo očekivati poboljšanje performansi sistema.

Dalji pravci razvoja i istraživanja koji mogu da dovedu do poboljšanja performansi sistema se ogledaju u otklanjanju navedenih ograničenja. Pre svega neophodno je formirati veći skup anotiranih medicinskih dokumenata i otkloniti hardverska ograničenja. Povećanje skupa anotiranih dokumenata bi rezultovalo manjim brojem retkih entiteta i termina te bi u određenoj meri bila smanjena greška. Uklanjanje hardverskih ograničenja bi dovelo do mogućnosti korišćenja i do-obučavanja većih modela poput XML-RoBERTa-Large čime bi imali objektivniju sliku o mogućnostima modela mašinskog učenja i uticaja veličine modela na konačne performanse.

Literatura

Akhtyamova, L., Martinez, P., Verspoor, K., & Cardiff, J. (2020). Testing contextualized word embeddings to improve NER in Spanish clinical case narratives. *IEEE Access*, *8*, 164717–164726.

Aleksandar Kovačević. (2011). *Automatizovano izdvajanje semantike iz naučnih članaka u oblasti Informatike*. Univerzitet u Novom Sadu.

Alfattni, G., Peek, N., & Nenadic, G. (2020). Extraction of temporal relations from clinical free text: A systematic review of current approaches. *Journal of Biomedical Informatics*, *108*, 103488.

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. *ArXiv Preprint ArXiv:1904.03323*.

Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings of the AMIA Symposium*, 17–21.

Avdic, A., Marovac, U., & Jankovic, D. (2020). Automated labeling of terms in medical reports in Serbian. *Turkish Journal of Electrical Engineering & Computer Sciences*, *28*(6), 3285–3303.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. *ArXiv Preprint ArXiv:1607.06450*.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *ArXiv Preprint ArXiv:1409.0473*.

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. *ArXiv Preprint ArXiv:1903.10676*.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc."

Bose, P., Srinivasan, S., Sleeman, W. C., Palta, J., Kapoor, R., & Ghosh, P. (2021). A Survey on Recent Named Entity Recognition and Relationship Extraction Techniques on Clinical Texts. *Applied Sciences*, *11*(18), 8319.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123–140.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., & others. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2019). Unsupervised Cross-lingual Representation Learning at Scale. *CoRR*, *abs/1911.02116*. http://arxiv.org/abs/1911.02116

Cowie, M. R., Blomster, J. I., Curtis, L. H., Duclaux, S., Ford, I., Fritz, F., Goldman, S., Janmohamed, S., Kreuzer, J., Leenay, M., & others. (2017). Electronic health records to facilitate clinical research. *Clinical Research in Cardiology*, *106*(1), 1–9.

de Oliveira, J. M., da Costa, C. A., & Antunes, R. S. (2021). Data structuring of electronic health records: a systematic review. *Health and Technology*, *11*(6), 1219–1235.

Dehghan, A., Keane, J. A., & Nenadic, G. (2013). Challenges in Clinical Named Entity Recognition for Decision Support. *2013 IEEE International Conference on Systems, Man, and Cybernetics*, 947–951. https://doi.org/10.1109/SMC.2013.166

Dernoncourt, F., Lee, J. Y., Uzuner, O., & Szolovits, P. (2017). De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, *24*(3), 596–606.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint ArXiv:1810.04805*.

Dumitrescu, S. D., Avram, A.-M., & Pyysalo, S. (2020). *The birth of Romanian BERT*. arXiv. https://doi.org/10.48550/ARXIV.2009.08712

Durango, M. C., Torres-Silva, E. A., & Orozco-Duque, A. (2023). Named Entity Recognition in Electronic Health Records: A Methodological Review. *Healthc Inform Res*, *29*(4), 286–300.

Elhadad, N., Pradhan, S., Gorman, S., Manandhar, S., Chapman, W., & Savova, G. (2015). SemEval-2015 task 14: Analysis of clinical text. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 303–310.

Evans, R. S. (2016). Electronic health records: then, now, and in the future. *Yearbook of Medical Informatics*, *25*(S 01), S48–S61.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378.

Fleiss, J. L., Levin, B., & Paik, M. C. (2013). *Statistical methods for rates and proportions*. john wiley & sons.

Friedman, C., Alderson, P. O., Austin, J. H. M., Cimino, J. J., & Johnson, S. B. (1994). A General Natural-language Text Processor for Clinical Radiology. *Journal of the American Medical Informatics Association*, *1*(2), 161–174. https://doi.org/10.1136/jamia.1994.95236146

Ganaie, M. A., Hu, M., Malik, A. K., Tanveer, M., & Suganthan, P. N. (2022). Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, *115*, 105151.

Gaschi, F., Fontaine, X., Rastin, P., & Toussaint, Y. (2023). Multilingual Clinical NER: Translation or Cross-lingual Transfer? In T. Naumann, A. Ben Abacha, S. Bethard, K. Roberts, & A. Rumshisky (Eds.), *Proceedings of the 5th Clinical Natural Language Processing Workshop* (pp. 289–311). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.clinicalnlp-1.34

Goulart, R. R. V., Strube de Lima, V. L., & Xavier, C. C. (2011). A systematic review of named entity recognition in biomedical texts. *Journal of the Brazilian Computer Society*, *17*(2), 103–116. https://doi.org/10.1007/s13173-011-0031-9

Goyal, A., Gupta, V., & Kumar, M. (2018). Recent named entity recognition and classification techniques: a systematic review. *Computer Science Review*, *29*, 21–43.

Grishman, R., & Sundheim, B. (1996). Message Understanding Conference- 6: A Brief History. *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*. https://aclanthology.org/C96-1079

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. *ArXiv Preprint ArXiv:2004.10964*.

Habibi, M., Weber, L., Neves, M., Wiegandt, D. L., & Leser, U. (2017). Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, *33*(14), i37–i48.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hu, Z., & Ma, X. (2023). A novel neural network model fusion approach for improving medical named entity recognition in online health

expert question-answering services. *Expert Systems with Applications*, *223*, 119880. https://doi.org/https://doi.org/10.1016/j.eswa.2023.119880

Huang, K., Altosaar, J., & Ranganath, R. (2019). Clinicalbert: Modeling clinical notes and predicting hospital readmission. *ArXiv Preprint ArXiv:1904.05342*.

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *ArXiv Preprint ArXiv:1508.01991*.

Humphreys, K., Gaizauskas, R., Azzam, S., Huyck, C., Mitchell, B., Cunningham, H., & Wilks, Y. (1998). University of Sheffield: Description of the LaSIE-II system as used for MUC-7. *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998*.

Ibáñez-Garcia, S., Rodriguez-Gonzalez, C., Escudero-Vilaplana, V., Martin-Barbero, M. L., Marzal-Alfaro, B., la Rosa-Triviño, J. L., Iglesias-Peinado, I., Herranz-Alonso, A., & Sanjurjo Saez, M. (2019). Development and Evaluation of a Clinical Decision Support System to Improve Medication Safety. *Applied Clinical Informatics*, *10*(3), 513–520. https://doi.org/10.1055/s-0039-1693426

Jaćimović, J., Krstev, C., & Jelovac, D. (2015). A rule-based system for automatic de-identification of medical narrative texts. *Informatica*, *39*(1).

Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, *13*(6), 395–405.

Jurafsky, D., & Martin, J. H. (2020). *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition. 3rd Edition draft* (3rd ed.). https://web.stanford.edu/~jurafsky/slp3/

Kaplar, A., Aleksić, A., Stošović, M., Naumović, R., Brković, V., & Kovačević, A. (2019). Evaluating String Distance Metrics for Approximate Dictionary Matching: A Case Study in Serbian Electronic Health Records. *Proceedings of the 9th International Conference on Information Society and Technology*, 135–137.

Keretna, S., Lim, C. P., Creighton, D., & Shaban, K. B. (2015). Enhancing medical named entity recognition with an extended segment representation technique. *Computer Methods and Programs in Biomedicine*, *119*(2), 88–100.

Kim, Y.-M., & Lee, T.-H. (2020). Korean clinical entity recognition from diagnosis text using BERT. *BMC Medical Informatics and Decision Making*, *20*(7), 1–9.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *ArXiv Preprint ArXiv:1412.6980*.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Koutsikakis, J., Chalkidis, I., Malakasiotis, P., & Androutsopoulos, I. (2020). GREEK-BERT: The Greeks Visiting Sesame Street. *11th Hellenic Conference on Artificial Intelligence*, 110–117. https://doi.org/10.1145/3411408.3411440

Kovačević, A., Dehghan, A., Filannino, M., Keane, J. A., & Nenadic, G. (2013). Combining rules and machine learning for extraction of temporal expressions and events from clinical narratives. *Journal of the American Medical Informatics Association*, *20*(5), 859–866.

Kozareva, Z., Ferrández, O., Montoyo, A., Muñoz, R., Suárez, A., & Gómez, J. (2007). Combining data-driven systems for improving named entity recognition. *Data & Knowledge Engineering*, *61*(3), 449–466.

Krstev, C., Obradović, I., Utvić, M., & Vitas, D. (2014). A system for named entity recognition based on local grammars. *Journal of Logic and Computation*, *24*(2), 473–489.

Krupka, G., & Hausman, K. (1998). IsoQuest Inc.: description of the NetOwl^{TM} extractor system as used for MUC-7. *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998*.

Lafferty, J., McCallum, A., & Pereira, F. C. N. (2001). *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159–174.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234–1240.

Lee, W., Kim, K., Lee, E. Y., & Choi, J. (2018a). Conditional random fields for clinical named entity recognition: a comparative study using Korean clinical texts. *Computers in Biology and Medicine*, *101*, 7–14.

Lee, W., Kim, K., Lee, E. Y., & Choi, J. (2018b). Conditional random fields for clinical named entity recognition: A comparative study

using Korean clinical texts. *Computers in Biology and Medicine*, *101*, 7–14. https://doi.org/https://doi.org/10.1016/j.compbiomed.2018.07.019

Lewis, P., Ott, M., Du, J., & Stoyanov, V. (2020). Pretrained language models for biomedical and clinical tasks: understanding and extending the state-of-the-art. *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, 146–157.

Li, J., Sun, A., Han, J., & Li, C. (2020). A Survey on Deep Learning for Named Entity Recognition. *IEEE Transactions on Knowledge and Data Engineering*, *34*(1), 50–70. https://doi.org/10.1109/TKDE.2020.2981314

Li, X., Wen, Q., Lin, H., Jiao, Z., & Zhang, J. (2021). Overview of CCKS 2020 Task 3: named entity recognition and event extraction in Chinese electronic medical records. *Data Intelligence*, *3*(3), 376–388.

Liu, D. C., & Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, *45*(1–3), 503–528.

Liu, L., Shang, J., Ren, X., Xu, F., Gui, H., Peng, J., & Han, J. (2018). Empower sequence labeling with task-aware neural language model. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1).

Liu, P. J., Saleh, M., Pot, E., Goodrich, B., Sepassi, R., Kaiser, L., & Shazeer, N. (2018). Generating wikipedia by summarizing long sequences. *ArXiv Preprint ArXiv:1801.10198*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *ArXiv Preprint ArXiv:1907.11692*.

Ljubešić, N., & Lauc, D. (2021). BERTić–The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian. *ArXiv Preprint ArXiv:2104.09243*.

Marković, I. P. (2017). *Izbor atributa integracijom znanja o domenu primenom metoda odlučivanja kod prediktivnog modelovanja vremenskih serija nadgledanim mašinskim učenjem*. Univerzitet u Nišu.

Marovac, U. A., Avdić, A. R., & Milošević, N. L. (2023). A Survey of Resources and Methods for Natural Language Processing of Serbian Language. *ArXiv Preprint ArXiv:2304.05468*.

Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., & Hurdle, J. F. (2008). Extracting information from textual documents in the

electronic health record: a review of recent research. *Yearbook of Medical Informatics*, 128–144.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *ArXiv Preprint ArXiv:1301.3781*.

Moharasan, G., & Ho, T.-B. (2017). Extraction of temporal events from clinical text using semi-supervised conditional random fields. *International Conference on Data Mining and Big Data*, 409–421.

Mohit, B. (2014). Named Entity Recognition. In I. Zitouni (Ed.), *Natural Language Processing of Semitic Languages* (pp. 221–245). Springer. https://doi.org/10.1007/978-3-642-45358-8_7

Nayel, H., & Shashirekha, H. L. (2017). Improving NER for clinical texts by ensemble approach using segment representations. *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, 197–204.

Nikola Nikolić. (2021). *Automatsko izdvajanje mišljenja iz tekstualnih komentara studentskih anketa*. Univerzitet u Novom Sadu.

Olah, C. (2015). *Understanding LSTM Networks*. http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Ornstein, S. M., Oates, R. B., & Fox, G. N. (1992). Thecompu-ter-based medical record: Currentstatus. *The Journal of Family Practice*, *35*(5), 557.

Palshikar, G. K. (2013). Techniques for named entity recognition: a survey. In *Bioinformatics: Concepts, Methodologies, Tools, and Applications* (pp. 400–426). IGI Global.

Patrick, J., & Li, M. (2010). High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *Journal of the American Medical Informatics Association*, *17*(5), 524–527.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

Peng, Y., Yan, S., & Lu, Z. (2019). Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. *ArXiv Preprint ArXiv:1906.05474*.

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, *6*(3), 21–45. https://doi.org/10.1109/MCAS.2006.1688199

Puflović, D., Velinov, G., Stanković, T., Janković, D., & Stoimenov, L. (2016). A supervised named entity recognition for information extraction from medical records. *ICIST 2016 Proceedings*, 91–96.

Pustejovsky, J., Castano, J. M., Ingria, R., Sauri, R., Gaizauskas, R. J., Setzer, A., Katz, G., & Radev, D. R. (2003). TimeML: Robust specification of event and temporal expressions in text. *New Directions in Question Answering*, *3*, 28–34.

Pustejovsky, J., & Stubbs, A. (2012). *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. " O'Reilly Media, Inc."

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., & others. (2018). *Improving language understanding by generative pre-training*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., & others. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.

Ramos-Flores, O., Pinto, D., Montes-y-Gómez, M., & Vázquez, A. (2020). Probabilistic vs deep learning based approaches for narrow domain NER in Spanish. *Journal of Intelligent & Fuzzy Systems*, *39*(2), 2015–2025.

Ramshaw, L. A., & Marcus, M. P. (1995). Text Chunking using Transformation-Based Learning. *ArXiv*, *cmp-lg/9505040*. https://api.semanticscholar.org/CorpusID:725590

Ratnaparkhi, A. (1998). *Maximum entropy models for natural language ambiguity resolution*. University of Pennsylvania.

Raza, K. (2019). Improving the prediction accuracy of heart disease with ensemble learning and majority voting rule. In *U-Healthcare Monitoring Systems* (pp. 179–196). Elsevier.

Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.

Rim, K. (2016). Mae2: Portable annotation tool for general natural language use. *Proc 12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, 75–80.

Robinson, A. J., & Fallside, F. (1987). *The utility driven dynamic error propagation network* (Vol. 1). University of Cambridge Department of Engineering Cambridge.

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *WIREs Data Mining and Knowledge Discovery*, *8*(4), e1249. https://doi.org/https://doi.org/10.1002/widm.1249

Šandrih, B., Krstev, C., & Stanković, R. (2019). Development and evaluation of three named entity recognition systems for serbian-the case of personal names. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, 1060–1068.

Sang, E. F., & Veenstra, J. (1999). Representing text chunks. *ArXiv Preprint Cs/9907006*.

Segura-Bedmar, I., Mart\'\inez, P., & Herrero-Zazo, M. (2013). SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 341–350. https://aclanthology.org/S13-2056

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. In K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1715–1725). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1162

Si, Y., Wang, J., Xu, H., & Roberts, K. (2019). Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, *26*(11), 1297–1304.

Sim, J., & Wright, C. C. (2005). The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Physical Therapy*, *85*(3), 257–268. https://doi.org/10.1093/ptj/85.3.257

Spasic, I., Nenadic, G., & others. (2020). Clinical text data in machine learning: systematic review. *JMIR Medical Informatics*, *8*(3), e17984.

Speck René and Ngonga Ngomo, A.-C. (2014). Ensemble Learning for Named Entity Recognition. In T. and B. A. and W. C. and K. C. and V. D. and G. P. and N. N. and J. K. and G. C. Mika Peter and Tudorache (Ed.), *The Semantic Web – ISWC 2014* (pp. 519–534). Springer International Publishing.

Stubbs, A., Kotfila, C., Xu, H., & Uzuner, Ö. (2015). Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2. *Journal of Biomedical Informatics*, *58*, S67–S77.

Su, J., Hu, J., Jiang, J., Xie, J., Yang, Y., He, B., Yang, J., & Guan, Y. (2019). Extraction of risk factors for cardiovascular diseases from

Chinese electronic medical records. *Computer Methods and Programs in Biomedicine*, *172*, 1–10.

Sun, W., Rumshisky, A., & Uzuner, O. (2013a). Annotating temporal information in clinical narratives. *Journal of Biomedical Informatics*, *46*, S5–S12.

Sun, W., Rumshisky, A., & Uzuner, O. (2013b). Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *Journal of the American Medical Informatics Association*, *20*(5), 806–813. https://doi.org/10.1136/amiajnl-2013-001628

Suominen, H., Salanterä, S., Velupillai, S., Chapman, W. W., Savova, G., Elhadad, N., Pradhan, S., South, B. R., Mowery, D. L., Jones, G. J. F., & others. (2013). Overview of the ShARe/CLEF eHealth evaluation lab 2013. *International Conference of the Cross-Language Evaluation Forum for European Languages*, 212–231.

Sutton, C., McCallum, A., & others. (2012). An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, *4*(4), 267–373.

Thawani, A., Pujara, J., Szekely, P. A., & Ilievski, F. (2021). Representing numbers in NLP: a survey and a vision. *ArXiv Preprint ArXiv:2103.13136*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., & others. (2023). Llama: Open and efficient foundation language models. *ArXiv Preprint ArXiv:2302.13971*.

Uzuner, Ö., Solti, I., & Cadag, E. (2010). Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, *17*(5), 514–518.

Uzuner, Ö., Solti, I., Xia, F., & Cadag, E. (2010). Community annotation experiment for ground truth generation for the i2b2 medication challenge. *Journal of the American Medical Informatics Association*, *17*(5), 519–523.

Uzuner, Ö., South, B. R., Shen, S., & DuVall, S. L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, *18*(5), 552–556. https://doi.org/10.1136/amiajnl-2011-000203

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., & Pyysalo, S. (2019). *Multilingual is not enough: BERT for Finnish*. arXiv. https://doi.org/10.48550/ARXIV.1912.07076

Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., Liu, S., Zeng, Y., Mehrabi, S., Sohn, S., & others. (2018). Clinical information extraction applications: a literature review. *Journal of Biomedical Informatics*, *77*, 34–49.

Williams, R. J., & Zipser, D. (1989). A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, *1*(2), 270–280.

Williams, R. J., & Zipser, D. (1995). Gradient-based learning algorithms for recurrent networks and their computational complexity. *Back-Propagation: Theory, Architectures and Applications*, *13*, 433–486.

Wissler, L., Almashraee, M., D\'\iaz, D. M., & Paschke, A. (2014). The Gold Standard in Corpus Annotation. *IEEE GSC*, *21*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & others. (2019). Huggingface's transformers: State-of-the-art natural language processing. *ArXiv Preprint ArXiv:1910.03771*.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241–259. https://doi.org/https://doi.org/10.1016/S0893-6080(05)80023-1

Wu, S., Roberts, K., Datta, S., Du, J., Ji, Z., Si, Y., Soni, S., Wang, Q., Wei, Q., Xiang, Y., & others. (2020). Deep learning in clinical natural language processing: a methodical review. *Journal of the American Medical Informatics Association*, *27*(3), 457–470.

Wu, Y., Jiang, M., Xu, J., Zhi, D., & Xu, H. (2017). Clinical named entity recognition using deep learning models. *AMIA Annual Symposium Proceedings*, *2017*, 1812.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V, Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., & others. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv Preprint ArXiv:1609.08144*.

Xia, F., & Yetisgen-Yildiz, M. (2012). Clinical corpus annotation: challenges and strategies. *Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM'2012) in Conjunction with the International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey*, 67.

Xu, K., Zhou, Z., Hao, T., & Liu, W. (2018). A Bidirectional LSTM and Conditional Random Fields Approach to Medical Named Entity Recognition. In A. E. Hassanien, K. Shaalan, T. Gaber, & M. F. Tolba (Eds.), *Proceedings of the International Conference on*

Advanced Intelligent Systems and Informatics 2017 (Vol. 639, pp. 355–365). Springer International Publishing. https://doi.org/10.1007/978-3-319-64861-3_33

Zhou, Y., Ju, C., Caufield, J. H., Shih, K., Chen, C., Sun, Y., Chang, K.-W., Ping, P., & Wang, W. (2021). Clinical named entity recognition using contextualized token representations. *ArXiv Preprint ArXiv:2106.12608*.

Biografija

Aleksandar Kaplar rođen je 1991. godine u Bijeljini. Završio je srednju tehničku školu „Mihajlo Pupin“ u Bijeljini 2010. godine.

Četvorogodišnje akademske studije na Fakultetu tehničkih nauka u Novom Sadu upisao je 2010. godine na studijskom programu Računarstvo i automatika. Diplomirao je 2014. godine sa temom „Distribuirana aplikacija za klasifikaciju Twitter profila prema polu“.

Školske 2014/2015. godine je upisao master akademske studije na studijskom programu Računarstvo i automatika. Master rad odbranio je 2015. godine sa temom „Primena veštačkih neuronskih mreža u analizi visokodimenzionalnih skupova podataka“.

Od 2015. godine je student doktorskih akademskih studija na Fakultetu tehničkih nauka – smer Računarstvo i automatika. Autor je i koautor više naučnih radova predstavljenih na domaćim i međunarodnim konferencijama.

Prilozi

A. Tabele rezultata za tačno poklapanje entieta po klasama

U ovoj sekciji dati su rezultati po klasama entiteta za sledeće modele: CRF, LM-LSTM-CRF, T-RoBERTa, BERT Multilingual Cased, BERT Multilingual Uncased, XLM RoBERTa Base, XLM RoBERTa Large, PT – BERT Multilingual Cased, PT – XLM RoBERTa Base.

Klasa	***Preciznost***	***Odziv***	***F1 mera***
CLINICAL_DEPT	0.929	0.813	0.867
DATE	0.928	0.860	0.892
DURATION	0.936	0.889	0.912
EVIDENTIAL	0.914	0.846	0.879
FREQUENCY	0.927	0.941	0.934
OCCURRENCE	0.838	0.693	0.759
PROBLEM	0.721	0.601	0.656
TEST	0.933	0.905	0.919
TIME	01.0	01.0	01.0
TREATMENT	0.860	0.779	0.817
VALUE	0.922	0.916	0.919
macro avg	0.901	0.840	0.868
micro avg	0.892	0.840	0.865

Tabela A.1. Rezultati za tačno poklapanje entiteta CRF modela

Klasa	***Preciznost***	***Odziv***	***F1 mera***
CLINICAL_DEPT	0.741	0.833	0.784
DATE	0.914	0.930	0.922
DURATION	0.920	0.929	0.925
EVIDENTIAL	0.820	0.841	0.830
FREQUENCY	0.951	0.961	0.956
OCCURRENCE	0.825	0.751	0.786
PROBLEM	0.694	0.678	0.686
TEST	0.928	0.932	0.930
TIME	0.50	0.667	0.571
TREATMENT	0.858	0.860	0.859
VALUE	0.924	0.941	0.932
macro avg	0.825	0.847	0.835
micro avg	0.879	0.882	0.880

Tabela A.2. Rezultati za tačno poklapanje entiteta LM-LSTM-CRF modela

Klasa	**Preciznost**	**Odziv**	**F1 mera**
CLINICAL_DEPT	0.611	0.688	0.647
DATE	0.833	0.853	0.843
DURATION	0.763	0.879	0.817
EVIDENTIAL	0.768	0.791	0.779
FREQUENCY	0.723	0.759	0.740
OCCURRENCE	0.708	0.646	0.675
PROBLEM	0.495	0.568	0.529
TEST	0.772	0.766	0.769
TIME	0.0	0.0	0.0
TREATMENT	0.699	0.770	0.733
VALUE	0.762	0.779	0.770
macro avg	0.649	0.682	0.664
micro avg	0.717	0.745	0.731

Tabela A.3. Rezultati za tačno poklapanje entiteta T-RoBERTa modela

Klasa	Preciznost	Odziv	F1 mera
CLINICAL_DEPT	0.765	0.813	0.788
DATE	0.857	0.90	0.878
DURATION	0.80	0.929	0.860
EVIDENTIAL	0.742	0.831	0.784
FREQUENCY	0.829	0.906	0.866
OCCURRENCE	0.689	0.776	0.730
PROBLEM	0.636	0.681	0.657
TEST	0.898	0.936	0.917
TIME	0.333	0.333	0.333
TREATMENT	0.786	0.816	0.801
VALUE	0.892	0.934	0.912
macro avg	0.748	0.805	0.775
micro avg	0.825	0.873	0.849

Tabela A.4. Rezultati za tačno poklapanje entiteta BERT Multilingual Cased modela

Klasa	Preciznost	Odziv	F1 mera
CLINICAL_DEPT	0.735	0.750	0.742
DATE	0.890	0.916	0.903
DURATION	0.798	0.919	0.854
EVIDENTIAL	0.766	0.816	0.790
FREQUENCY	0.842	0.921	0.880
OCCURRENCE	0.681	0.733	0.706
PROBLEM	0.601	0.671	0.634
TEST	0.871	0.931	0.90
TIME	0.50	0.333	0.40
TREATMENT	0.748	0.819	0.782
VALUE	0.896	0.932	0.914
macro avg	0.757	0.795	0.773
micro avg	0.812	0.869	0.840

Tabela A.5. Rezultati za tačno poklapanje entiteta BERT Multilingual Uncased modela

Klasa	Preciznost	Odziv	F1 mera
CLINICAL_DEPT	0.698	0.771	0.733
DATE	0.769	0.866	0.814
DURATION	0.798	0.919	0.854
EVIDENTIAL	0.721	0.836	0.774
FREQUENCY	0.783	0.887	0.831
OCCURRENCE	0.697	0.729	0.713
PROBLEM	0.584	0.649	0.615
TEST	0.856	0.90	0.878
TIME	0.0	0.0	0.0
TREATMENT	0.751	0.789	0.770
VALUE	0.758	0.841	0.797
macro avg	0.674	0.744	0.707
micro avg	0.757	0.823	0.789

Tabela A.6. Rezultati za tačno poklapanje entiteta XLM RoBERTa Base modela

Klasa	Preciznost	Odziv	F1 mera
CLINICAL_DEPT	0.740	0.771	0.755
DATE	0.867	0.913	0.889
DURATION	0.913	0.949	0.931
EVIDENTIAL	0.784	0.866	0.823
FREQUENCY	0.929	0.970	0.949
OCCURRENCE	0.759	0.794	0.776
PROBLEM	0.707	0.692	0.699
TEST	0.883	0.916	0.899
TIME	0.40	0.667	0.50
TREATMENT	0.834	0.845	0.839
VALUE	0.862	0.912	0.886
macro avg	0.789	0.845	0.813
micro avg	0.840	0.871	0.855

Tabela A.7. Rezultati za tačno poklapanje entiteta XLM RoBERTa Large modela

Klasa	Preciznost	Odziv	F1 mera
CLINICAL_DEPT	0.667	0.833	0.741
DATE	0.881	0.913	0.897
DURATION	0.887	0.949	0.917
EVIDENTIAL	0.806	0.866	0.835
FREQUENCY	0.947	0.970	0.959
OCCURRENCE	0.762	0.776	0.769
PROBLEM	0.678	0.690	0.684
TEST	0.916	0.943	0.929
TIME	0.60	01.0	0.750
TREATMENT	0.842	0.864	0.853
VALUE	0.917	0.952	0.934
macro avg	0.809	0.887	0.842
micro avg	0.862	0.892	0.877

Tabela A.8. Rezultati za tačno poklapanje entiteta PT – BERT Multilingual Cased modela

Klasa	Preciznost	Odziv	F1 mera
CLINICAL_DEPT	0.741	0.833	0.784
DATE	0.787	0.866	0.825
DURATION	0.850	0.970	0.906
EVIDENTIAL	0.728	0.891	0.801
FREQUENCY	0.853	0.887	0.870
OCCURRENCE	0.717	0.751	0.734
PROBLEM	0.638	0.690	0.663
TEST	0.877	0.917	0.896
TIME	0.0	0.0	0.0
TREATMENT	0.781	0.840	0.810
VALUE	0.755	0.844	0.797
macro avg	0.702	0.772	0.735
micro avg	0.777	0.844	0.809

Tabela A.9. Rezultati za tačno poklapanje entiteta PT – XLM RoBERTa Base modela

B. Tabele rezultata na nivou tokena

U ovoj sekciji su dati rezultati na nivou tokena za iste modele kao u prethodnoj sekciji. Za razliku od rezultata za tačno poklapanje entieta, rezultati na nivou tokena su dati u IOB2 formatu (sekcija 2.7). Prilikom računanja rezultata za tačno poklapanje entiteta nisu pravljene razlike između „*B*-„ i „*I*-„ klasa tokena dok su za metrike na nivou tokena te razlike uzete u obzir.

Klasa	**Preciznost**	**Odziv**	**F1 mera**
B-CLINICAL_DEPT	0.952	0.833	0.889
B-DATE	0.942	0.876	0.908
B-DURATION	0.947	0.899	0.922
B-EVIDENTIAL	0.919	0.851	0.884
B-FREQUENCY	0.942	0.956	0.949
B-OCCURRENCE	0.90	0.744	0.814
B-PROBLEM	0.778	0.652	0.709
B-TEST	0.953	0.924	0.938
B-TIME	1.0	1.0	1.0
B-TREATMENT	0.891	0.809	0.848
B-VALUE	0.946	0.939	0.942
I-CLINICAL_DEPT	0.953	0.719	0.820
I-DATE	0.914	0.860	0.886
I-DURATION	0.959	0.879	0.917
I-EVIDENTIAL	0.930	0.864	0.896
I-FREQUENCY	0.781	0.833	0.806
I-OCCURRENCE	0.842	0.583	0.689
I-PROBLEM	0.735	0.615	0.670
I-TEST	0.791	0.729	0.759
I-TIME	1.0	1.0	1.0
I-TREATMENT	0.793	0.712	0.750
I-VALUE	0.914	0.942	0.928
macro avg	0.899	0.828	0.860
micro avg	0.892	0.835	0.862
weighted avg	0.888	0.835	0.859

Tabela B.1. Rezultati na nivou tokena za CRF model

Klasa	Preciznost	Odziv	F1 mera
B-CLINICAL_DEPT	0.870	0.833	0.851
B-DATE	0.950	0.953	0.951
B-DURATION	0.909	0.909	0.909
B-EVIDENTIAL	0.835	0.881	0.857
B-FREQUENCY	0.938	0.970	0.954
B-OCCURRENCE	0.826	0.791	0.808
B-PROBLEM	0.752	0.702	0.726
B-TEST	0.941	0.949	0.945
B-TIME	1.0	0.667	0.80
B-TREATMENT	0.874	0.873	0.873
B-VALUE	0.946	0.959	0.953
I-CLINICAL_DEPT	0.855	0.825	0.839
I-DATE	0.894	0.950	0.921
I-DURATION	0.915	0.907	0.911
I-EVIDENTIAL	0.839	0.889	0.863
I-FREQUENCY	0.774	0.80	0.787
I-OCCURRENCE	0.749	0.659	0.701
I-PROBLEM	0.696	0.650	0.672
I-TEST	0.809	0.791	0.80
I-TIME	1.0	0.692	0.818
I-TREATMENT	0.844	0.773	0.807
I-VALUE	0.935	0.944	0.939
micro avg	0.878	0.868	0.873
macro avg	0.870	0.835	0.849
weighted avg	0.876	0.868	0.871

Tabela B.2. Rezultati na nivou tokena za LM-LSTM-CRF model

Klasa	Preciznost	Odziv	F1 mera
B-CLINICAL_DEPT	0.850	0.708	0.773
B-DATE	0.924	0.954	0.939
B-DURATION	0.876	0.920	0.898
B-EVIDENTIAL	0.874	0.833	0.853
B-FREQUENCY	0.931	0.973	0.952
B-OCCURRENCE	0.872	0.663	0.753
B-PROBLEM	0.688	0.627	0.656
B-TEST	0.920	0.923	0.921
B-TIME	0.0	0.0	0.0
B-TREATMENT	0.801	0.808	0.804
B-VALUE	0.907	0.946	0.926
I-CLINICAL_DEPT	0.769	0.690	0.727
I-DATE	0.893	0.883	0.888
I-DURATION	0.898	0.874	0.886
I-EVIDENTIAL	0.912	0.806	0.856
I-FREQUENCY	0.917	0.355	0.512
I-OCCURRENCE	0.806	0.456	0.583
I-PROBLEM	0.651	0.605	0.627
I-TEST	0.682	0.628	0.654
I-TIME	0.667	0.615	0.640
I-TREATMENT	0.747	0.684	0.714
I-VALUE	0.880	0.876	0.878
micro avg	0.852	0.830	0.841
macro avg	0.794	0.719	0.747
weighted avg	0.848	0.830	0.837

Tabela B.3. Rezultati na nivou tokena za T-RoBERTa model

Klasa	Preciznost	Odziv	F1 mera
B-CLINICAL_DEPT	0.932	0.774	0.846
B-DATE	0.955	0.970	0.962
B-DURATION	0.868	0.926	0.896
B-EVIDENTIAL	0.824	0.833	0.828
B-FREQUENCY	0.899	0.968	0.932
B-OCCURRENCE	0.785	0.736	0.760
B-PROBLEM	0.791	0.788	0.789
B-TEST	0.909	0.935	0.922
B-TIME	0.0	0.0	0.0
B-TREATMENT	0.863	0.866	0.865
B-VALUE	0.934	0.939	0.936
I-CLINICAL_DEPT	0.778	0.766	0.772
I-DATE	0.943	0.921	0.932
I-DURATION	0.834	0.958	0.892
I-EVIDENTIAL	0.868	0.801	0.833
I-FREQUENCY	0.735	0.679	0.706
I-OCCURRENCE	0.782	0.606	0.683
I-PROBLEM	0.716	0.724	0.720
I-TEST	0.735	0.749	0.742
I-TIME	0.632	0.923	0.750
I-TREATMENT	0.817	0.719	0.764
I-VALUE	0.904	0.930	0.917
micro avg	0.858	0.852	0.855
macro avg	0.796	0.796	0.793
weighted avg	0.856	0.852	0.854

Tabela B.4. Rezultati na nivou tokena za BERT Multilingual Uncased model

Klasa	Preciznost	Odziv	F1 mera
B-CLINICAL_DEPT	0.805	0.786	0.795
B-DATE	0.947	0.973	0.960
B-DURATION	0.854	0.936	0.893
B-EVIDENTIAL	0.825	0.857	0.841
B-FREQUENCY	0.899	0.965	0.931
B-OCCURRENCE	0.761	0.769	0.765
B-PROBLEM	0.778	0.787	0.782
B-TEST	0.918	0.934	0.926
B-TIME	0.0	0.0	0.0
B-TREATMENT	0.858	0.868	0.863
B-VALUE	0.941	0.965	0.953
I-CLINICAL_DEPT	0.759	0.817	0.787
I-DATE	0.932	0.908	0.920
I-DURATION	0.854	0.951	0.90
I-EVIDENTIAL	0.827	0.840	0.833
I-FREQUENCY	0.620	0.574	0.596
I-OCCURRENCE	0.767	0.623	0.687
I-PROBLEM	0.717	0.716	0.716
I-TEST	0.752	0.729	0.740
I-TIME	0.684	1.0	0.813
I-TREATMENT	0.785	0.711	0.746
I-VALUE	0.920	0.935	0.928
micro avg	0.856	0.858	0.857
macro avg	0.782	0.802	0.790
weighted avg	0.854	0.858	0.856

Tabela B.5. Rezultati na nivou tokena za BERT Multilingual Cased model

Klasa	Preciznost	Odziv	F1 mera
B-CLINICAL_DEPT	0.880	0.779	0.826
B-DATE	0.948	0.981	0.964
B-DURATION	0.818	0.90	0.857
B-EVIDENTIAL	0.773	0.851	0.810
B-FREQUENCY	0.902	0.976	0.938
B-OCCURRENCE	0.801	0.720	0.759
B-PROBLEM	0.741	0.775	0.758
B-TEST	0.906	0.934	0.919
B-TIME	0.0	0.0	0.0
B-TREATMENT	0.855	0.877	0.866
B-VALUE	0.937	0.960	0.948
I-CLINICAL_DEPT	0.746	0.810	0.777
I-DATE	0.892	0.962	0.926
I-DURATION	0.830	0.938	0.881
I-EVIDENTIAL	0.828	0.866	0.846
I-FREQUENCY	0.725	0.537	0.617
I-OCCURRENCE	0.798	0.557	0.656
I-PROBLEM	0.701	0.716	0.708
I-TEST	0.725	0.762	0.743
I-TIME	1.0	0.154	0.267
I-TREATMENT	0.786	0.745	0.765
I-VALUE	0.916	0.940	0.928
micro avg	0.845	0.855	0.850
macro avg	0.796	0.761	0.762
weighted avg	0.844	0.855	0.848

Tabela B.6. Rezultati na nivou tokena za XML RoBERTa Base model

Klasa	Preciznost	Odziv	F1 mera
B-CLINICAL_DEPT	0.792	0.811	0.802
B-DATE	0.984	0.974	0.979
B-DURATION	0.902	0.918	0.910
B-EVIDENTIAL	0.802	0.846	0.823
B-FREQUENCY	0.959	0.989	0.974
B-OCCURRENCE	0.795	0.781	0.788
B-PROBLEM	0.813	0.771	0.792
B-TEST	0.931	0.925	0.928
B-TIME	0.615	1.0	0.762
B-TREATMENT	0.885	0.889	0.887
B-VALUE	0.954	0.963	0.959
I-CLINICAL_DEPT	0.652	0.776	0.709
I-DATE	0.935	0.956	0.945
I-DURATION	0.919	0.954	0.936
I-EVIDENTIAL	0.865	0.841	0.853
I-FREQUENCY	0.828	0.889	0.857
I-OCCURRENCE	0.740	0.682	0.710
I-PROBLEM	0.756	0.714	0.735
I-TEST	0.772	0.785	0.779
I-TIME	0.750	0.923	0.828
I-TREATMENT	0.820	0.767	0.793
I-VALUE	0.939	0.944	0.941
micro avg	0.876	0.864	0.870
macro avg	0.837	0.868	0.849
weighted avg	0.874	0.864	0.869

Tabela B.7. Rezultati na nivou tokena za XML RoBERTa Large model

Klasa	Preciznost	Odziv	F1 mera
B-CLINICAL_DEPT	0.788	0.857	0.821
B-DATE	0.957	0.971	0.964
B-DURATION	0.892	0.928	0.910
B-EVIDENTIAL	0.849	0.873	0.861
B-FREQUENCY	0.955	0.968	0.962
B-OCCURRENCE	0.811	0.757	0.783
B-PROBLEM	0.794	0.793	0.794
B-TEST	0.930	0.943	0.937
B-TIME	0.750	0.750	0.750
B-TREATMENT	0.888	0.902	0.895
B-VALUE	0.948	0.975	0.961
I-CLINICAL_DEPT	0.748	0.863	0.801
I-DATE	0.926	0.916	0.921
I-DURATION	0.882	0.951	0.915
I-EVIDENTIAL	0.850	0.881	0.865
I-FREQUENCY	0.803	0.907	0.852
I-OCCURRENCE	0.747	0.640	0.690
I-PROBLEM	0.742	0.736	0.739
I-TEST	0.755	0.789	0.772
I-TIME	0.765	1.0	0.867
I-TREATMENT	0.837	0.809	0.823
I-VALUE	0.934	0.949	0.942
micro avg	0.873	0.879	0.876
macro avg	0.843	0.871	0.856
weighted avg	0.872	0.879	0.875

Tabela B.8. Rezultati na nivou tokena za PT - BERT Multilingual Cased model

Klasa	Preciznost	Odziv	F1 mera
B-CLINICAL_DEPT	0.893	0.885	0.889
B-DATE	0.961	0.981	0.971
B-DURATION	0.860	0.945	0.90
B-EVIDENTIAL	0.761	0.887	0.819
B-FREQUENCY	0.946	0.963	0.954
B-OCCURRENCE	0.829	0.769	0.798
B-PROBLEM	0.788	0.810	0.799
B-TEST	0.932	0.948	0.940
B-TIME	0.0	0.0	0.0
B-TREATMENT	0.875	0.913	0.894
B-VALUE	0.937	0.972	0.954
I-CLINICAL_DEPT	0.701	0.871	0.777
I-DATE	0.938	0.956	0.947
I-DURATION	0.869	0.969	0.916
I-EVIDENTIAL	0.802	0.887	0.842
I-FREQUENCY	0.766	0.667	0.713
I-OCCURRENCE	0.783	0.633	0.70
I-PROBLEM	0.730	0.763	0.746
I-TEST	0.80	0.804	0.802
I-TIME	0.833	0.769	0.80
I-TREATMENT	0.843	0.815	0.828
I-VALUE	0.920	0.952	0.936
micro avg	0.865	0.884	0.874
macro avg	0.807	0.825	0.815
weighted avg	0.865	0.884	0.873

Tabela B.9. Rezultati na nivou tokena za PT – XML RoBERTa Base model

Klasa	Preciznost	Odziv	F1 mera
B-CLINICAL_DEPT	0.872	0.854	0.863
B-DATE	0.973	0.980	0.977
B-DURATION	0.941	0.960	0.950
B-EVIDENTIAL	0.844	0.891	0.867
B-FREQUENCY	0.957	0.985	0.971
B-OCCURRENCE	0.888	0.801	0.843
B-PROBLEM	0.824	0.783	0.803
B-TEST	0.949	0.961	0.955
B-TIME	1.0	1.0	1.0
B-TREATMENT	0.899	0.903	0.901
B-VALUE	0.950	0.966	0.958
I-CLINICAL_DEPT	0.742	0.807	0.773
I-DATE	0.963	0.946	0.955
I-DURATION	0.929	0.972	0.950
I-EVIDENTIAL	0.895	0.899	0.897
I-FREQUENCY	0.818	0.90	0.857
I-OCCURRENCE	0.845	0.680	0.753
I-PROBLEM	0.793	0.731	0.761
I-TEST	0.823	0.798	0.810
I-TIME	0.867	1.0	0.929
I-TREATMENT	0.868	0.810	0.838
I-VALUE	0.942	0.952	0.947
micro avg	0.906	0.893	0.90
macro avg	0.890	0.890	0.889
weighted avg	0.904	0.893	0.898

Tabela B.10. Rezultati na nivou tokena za Ensamble Best model

Овај Образац чини саставни део докторске дисертације, односно докторског уметничког пројекта који се брани на Универзитету у Новом Саду. Попуњен Образац укоричити иза текста докторске дисертације, односно докторског уметничког пројекта.

План третмана података

Назив пројекта/истраживања
Аутоматско издвајање именованих ентитета из медицинских докумената на српском језику
Назив институције/институција у оквиру којих се спроводи истраживање
а) Факултет техничких наука, Универзитет у Новом Саду
Назив програма у оквиру ког се реализује истраживање
Рачунарство и аутоматика – докторска дисертација
1. Опис података
1.1 Врста студије *Укратко описати тип студије у оквиру које се подаци прикупљају* **Докторска дисертација** 1.2 Врсте података а) **квантитативни** б) **квалитативни** 1.3. Начин прикупљања података а) анкете, упитници, тестови б) **клиничке процене, медицински записи, електронски здравствени записи** в) генотипови: навести врсту _____

г) административни подаци: навести врсту _____
д) узорци ткива: навести врсту_____
ђ) снимци, фотографије: навести врсту_____
е) текст, навести врсту **Анонимизовани клинички извештаји, Литературни извори**
ж) мапа, навести врсту _____
з) остало: описати **Нумерички експерименти**

1.3 Формат података, употребљене скале, количина података

1.3.1 Употребљени софтвер и формат датотеке:
а) Excel фајл, датотека .xlsx_____
b) SPSS фајл, датотека _____
c) PDF фајл, датотека .pdf_____
d) Текст фајл, датотека .txt, .xml_____
e) JPG фајл, датотека _____
f) Остало, датотека _____

1.3.2. Број записа (код квантитативних података)

а) број варијабли **велики број**
б) број мерења (испитаника, процена, снимака и сл.) **велики број**

1.3.3. Поновљена мерења
а) да
б) **не**

Уколико је одговор да, одговорити на следећа питања:
а) временски размак између поновљених мера је _____
б) варијабле које се више пута мере односе се на _____
в) нове верзије фајлова који садрже поновљена мерења су именоване као _____

Напомене:
_____

Да ли формати и софтвер омогућавају дељење и дугорочну валидност података?

а) Да
б) ***Не***
Ако је одговор не, образложити
Софтвер јер формиран на основу приватних података тако да није могуће дељење података али је могуће слободно коришћење софтвера

2. Прикупљање података

2.1 Методологија за прикупљање/генерисање података

2.1.1. У оквиру ког истраживачког нацрта су подаци прикупљени?
а) експеримент, навести тип **Нумерички експеримент**
б) корелационо истраживање, навести тип _____
ц) анализа текста, навести тип **Прикупљање података анализом доступне литературе**
д) остало, навести шта **Анонимизовани медицински извештаји**

2.1.2 Навести врсте мерних инструмената или стандарде података специфичних за одређену научну дисциплину (ако постоје).
_____

_____

2.2 Квалитет података и стандарди

2.2.1. Третман недостајућих података
а) Да ли матрица садржи недостајуће податке? Да **Не**

Ако је одговор да, одговорити на следећа питања:
а) Колики је број недостајућих података?
_____
б) Да ли се кориснику матрице препоручује замена недостајућих података? Да Не
в) Ако је одговор да, навести сугестије за третман замене недостајућих података
_____

2.2.2. На који начин је контролисан квалитет података? Описати

Квалитет података је контролисан поређењем експерименталних и теоријских података

2.2.3. На који начин је извршена контрола уноса података у матрицу?
Контрола уноса података у матрицу је извршена од стране административног лица Факултета техничких наука која су уносила податке из извештаја у електронски облик

3. Третман података и пратећа документација

3.1. Третман и чување података

3.1.1. Подаци ће бити депоновани у _____ *репозиторијум.*
3.1.2. URL адреса
_____
3.1.3. DOI
_____

3.1.4. Да ли ће подаци бити у отвореном приступу?
а) Да
б) Да, али после ембарга који ће трајати до
_____
в) ***Не***

Ако је одговор не, навести разлог
Постоји ограничење о приступу подацима од стране Факултета техничких наука, Универзитета у Новом Саду, као и ризик од злоупотребе, неовлашћеног преузимања, обраде и објављивања целине или дела прикупљених и обрађених података истраживања.

3.1.5. Подаци неће бити депоновани у репозиторијум, али ће бити чувани.
Образложење
Подаци неће бити у отвореном приступу. Подаци се чувају у електронској форми на рачунарима одговорних и овлашћених лица.

3.2 Метаподаци и документација података
3.2.1. Који стандард за метаподатке ће бити примењен?

Не примењује се стандард за метаподатке.

3.2.1. Навести метаподатке на основу којих су подаци депоновани у репозиторијум.

Не примењује се стандард за метаподатке.

Ако је потребно, навести методе које се користе за преузимање података, аналитичке и процедуралне информације, њихово кодирање, детаљне описе варијабли, записа итд.

3.3 Стратегија и стандарди за чување података
3.3.1. До ког периода ће подаци бити чувани у репозиторијуму?
_____
3.3.2. Да ли ће подаци бити депоновани под шифром? Да **Не**
3.3.3. Да ли ће шифра бити доступна одређеном кругу истраживача? Да **Не**
3.3.4. Да ли се подаци морају уклонити из отвореног приступа после извесног времена?
Да **Не**
Образложити

---

4. Безбедност података и заштита поверљивих информација

Овај одељак МОРА бити попуњен ако ваши подаци укључују личне податке који се односе на учеснике у истраживању. За друга истраживања треба такође размотрити заштиту и сигурност података.

4.1 Формални стандарди за сигурност информација/података

Истраживачи који спроводе испитивања с људима морају да се придржавају Закона о заштити података о личности (*https://www.paragraf.rs/propisi/zakon_o_zastiti_podataka_o_licnosti.html*) и одговарајућег институционалног кодекса о академском интегритету.

4.1.2. Да ли је истраживање одобрено од стране етичке комисије?
Да Не
Ако је одговор Да, навести датум и назив етичке комисије која је одобрила истраживање
23.02.2017. Етички одбор Универзитетског клиничког центра Србије

4.1.2. Да ли подаци укључују личне податке учесника у истраживању? **Да** Не
Ако је одговор да, наведите на који начин сте осигурали поверљивост и сигурност информација везаних за испитанике:
а) Подаци нису у отвореном приступу
б) **Подаци су анонимизирани**
ц) Остало, навести шта
_____

5. Доступност података

5.1. Подаци ће бити
а) јавно доступни
б) доступни само уском кругу истраживача у одређеној научној области
ц) ***затворени***

Ако су подаци доступни само уском кругу истраживача, навести под којим условима могу да их користе:
_____

Ако су подаци доступни само уском кругу истраживача, навести на који начин могу приступити подацима:
_____

5.4. Навести лиценцу под којом ће прикупљени подаци бити архивирани.
Ауторство–некомерцијално–без прераде

6. Улоге и одговорност

6.1. Навести име и презиме и мејл адресу власника (аутора) података

Александар Каплар, aleksandar.kaplar@uns.ac.rs

6.2. Навести име и презиме и мејл адресу особе која одржава матрицу с подацима

Александар Каплар, aleksandar.kaplar@uns.ac.rs

6.3. Навести име и презиме и мејл адресу особе која омогућује приступ подацима другим истраживачима

Напомена: Подаци нису доступни.