Univerzitet u Sarajevu
Elektrotehnički fakultet Sarajevo
Odsjek za računarstvo i informatiku

Spacy model bosanskog jezika

Predmet: Vještačka inteligencija
Predmetni profesor i supervizor: Van. prof. dr. Amila Akagić
Studenti:

- Eldar Buzadžić
- Faruk Zahiragić

Faza 1: Izbor teme i opis problema

1.1 Naziv teme

SpaCy model bosanskog jezika

1.2 Opis problema

Prirodni jezik predstavlja jedan od najvažnijih oblika komunikacije među ljudima, ali njegova obrada putem računara je izuzetno složen zadatak zbog bogatstva strukture, višeznačnosti i konteksta. Iako postoje brojni alati i modeli za obradu jezika, većina njih je razvijena za engleski i druge svjetski dominantne jezike, dok su jezici poput bosanskog znatno manje zastupljeni u dostupnim NLP (Natural Language Processing) resursima.

Konkretan problem koji se nastoji riješiti ovim projektom je **nedostatak kvalitetnog jezičkog modela za bosanski jezik** koji podržava osnovne NLP zadatke kao što su tokenizacija, morfološka analiza, lematizacija, prepoznavanje dijelova govora (POS tagging), prepoznavanje zavisnosti (dependency parsing), segmentacija rečenica i prepoznavanje imenovanih entiteta (NER).

Bez pouzdanog jezičkog modela za bosanski jezik, razvoj aplikacija koje koriste obradu prirodnog jezika — poput analize sentimenta, automatskog sažimanja teksta, klasifikacije dokumenata, chatbota i sistema za pretragu — ostaje ograničen, nepouzdan ili u potpunosti onemogućen.

Cilj ovog projekta je stoga izgradnja i treniranje jezičkog modela zasnovanog na SpaCy okruženju, koji omogućava automatsku i tačnu analizu bosanskog jezika, čime se omogućava upotreba NLP tehnologija u širokom spektru praktičnih i istraživačkih domena za bosansko govorno područje.

Na ovom mjestu vrijedi definisati i standardne komponente svakog Spacy modela, a to su:

- `Tok2vec` - pretvaranje riječi u vektor oblik pogodan za dalje korištenje od strane drugih komponenti
- `Tagger` - označavanje dijela govora
- `Morphologizer` - morfološka analiza
- `Parser` - sintaksna analiza
- `Lemmatizer` - lematizacija
- `Senter` - detektor kraja rečenica.
- `Attribute_ruler` - modifikacija atributa tokena na osnovu pravila
- `Ner` - prepoznavanje imenovanih entiteta

1.3 Definicija osnovnih pojmova

- **Obrada prirodnog jezika (NLP):** Obrada prirodnog jezika (eng. *Natural Language Processing*, NLP) je oblast vještačke inteligencije koja se fokusira na interakciju između računara i ljudskog jezika. Cilj NLP-a je da računarima omogući razumijevanje, interpretaciju i generisanje jezika koji ljudi koriste za komunikaciju. NLP obuhvata zadatke poput segmentacije rečenica, tokenizacije, analize sintakse, semantike, prepoznavanja entiteta i još mnogo toga.
- **Tokenizacija:** Tokenizacija je proces razdvajanja teksta na manje jedinice, tzv. tokene. Tokeni mogu predstavljati riječi, interpunkcijske znakove, brojeve itd. Ovaj korak je osnovni preduslov za većinu NLP zadataka.
- **Lematizacija:** Lematizacija predstavlja proces svođenja riječi na njen osnovni, rječnički oblik – lemu. Na primjer, riječ "poslovali" se lematizira u "poslovati". Lematizacija omogućava da se različiti oblici iste riječi analiziraju kao jedna cjelina.
- **Analiza morfologije:** Morfološka analiza obuhvata identifikaciju gramatičkih osobina riječi (npr. rod, broj, padež, glagolski oblik) na osnovu njihovog oblika i konteksta u rečenici.
- **Označavanje dijelova govora (POS tagging):** Ova tehnika se koristi za automatsko dodjeljivanje gramatičkih kategorija (imenica, glagol, pridjev itd.) svakoj riječi u rečenici, što je ključno za razumijevanje njene sintaktičke funkcije.
- **Sintaksna analiza (Dependency Parsing):** Sintaksna analiza određuje gramatičke odnose između riječi u rečenici, identifikujući glavne riječi i njihove zavisne dijelove. Ovi odnosi se predstavljaju u obliku stablo-strukture koje prikazuje zavisnosti.
- **Prepoznavanje imenovanih entiteta (NER):** NER (eng. *Named Entity Recognition*) predstavlja zadatak identifikacije i klasifikacije specifičnih elemenata u tekstu, poput imena osoba (PER), lokacija

(LOC), organizacija (ORG), datuma i sl. Ova informacija je važna za razne aplikacije u analizi teksta i ekstrakciji znanja.
- **Jezički model:** Jezički model je statistički ili neuronski model treniran na velikim korpusima teksta s ciljem da "nauči" obrasce i strukture jezika. Takav model može se koristiti za razne zadatke, poput analize teksta, automatskog prevođenja, generisanja jezika i drugo.

1.4 Korist i primjena rješenja

Primjene u realnim sistemima:

- **Sistemi za analizu sentimenta:** Automatska detekcija emocija i stavova izraženih u tekstovima korisnika na društvenim mrežama, anketama ili recenzijama.

- **Chatbotovi i virtuelni asistenti:** Razvoj sistema koji razumiju i odgovaraju na upite korisnika na bosanskom jeziku, uz upotrebu lematizacije, označavanja dijelova govora i prepoznavanja entiteta.

- **Pretraživanje i ekstrakcija informacija:** Poboljšano pretraživanje tekstualnih baza podataka, npr. novinskih arhiva, zakonskih dokumenata ili medicinskih zapisa, kroz semantičku analizu i razumijevanje konteksta.

- **Automatsko označavanje i kategorizacija teksta:** Sistemi koji mogu klasifikovati tekstove po temama, žanrovima ili pravnim kategorijama.

- **Obrazovni alati:** Aplikacije koje pomažu u učenju jezika, gramatici i leksici kroz automatsku analizu i korekciju teksta.

Korist za istraživačku zajednicu:

- **Dalji razvoj NLP alata za bosanski jezik:** Model može poslužiti kao osnova za treniranje specijalizovanih modela za složenije zadatke, poput automatskog prevođenja, sažimanja teksta ili generisanja odgovora.

- **Uporedna jezička analiza:** Omogućava poređenje bosanskog sa srodnim jezicima (hrvatski, srpski, crnogorski) u lingvističkim i

računalnim istraživanjima.

- **Primjena u društvenim naukama:** Omogućava obradu velikih korpusa tekstova radi analize društvenih narativa, medijskog diskursa, političkih govora i dr.

1.5 Pregled postojećih dataset-ova

1.5.1 Training corpus SUK 1.0

SUK trening korpus sadrži oko milion tokena ručno označenih na nivoima tokenizacije, segmentacije rečenica, morfosintaktičkog označavanja i lematizacije, a neki dijelovi sadrže i daljnje ručno verificirane anotacije. Morfosintaktičke oznake i (gdje su prisutne) sintaktičke zavisnosti uključene su i u JOS/MULTEXT-East okvir, kao i u okvir Universal Dependencies. Jezički korpus slovenskog jezika je prvi objavljeni korpus južnoslavenskog jezika i služio je kao baza za razvoj korpusa ostalih južnoslavenskih jezika.
Izvor: https://www.clarin.si/repository/xmlui/handle/11356/1747

1.5.2 Training corpus hr500k

Trening korpus hr500k sadrži oko 500.000 tokena ručno označenih na nivoima tokenizacije, segmentacije rečenica, morfosintaktičkog označavanja, lematizacije i imenovanih entiteta. Oko polovine korpusa je također ručno označeno sintaktičkim zavisnostima. Nadalje, oko petine korpusa je označeno semantičkim oznakama uloga. Format korpusa je prilagođen CoNLL-U specifikaciji. Ovaj korpus je baziran na otkrićima slovenskog korpusa. Za potrebe treniranja bosanskog modela, ovaj dataset se može koristiti kao početna osnova uz pažljivu prilagodbu jezičkim razlikama između hrvatskog i bosanskog jezika. Budući da se oba jezika oslanjaju na sličnu gramatiku i leksiku, transfer učenja je efektivan, posebno u kontekstu morfološke obrade i parsiranja.

Izvor: https://github.com/reldi-data/hr500k

1.5.3 Bosanskohercegovački nacionalni korpus (BHNC)

BHNC predstavlja najobuhvatniji korpus savremenog bosanskog jezika. Obuhvata tekstove iz raznih domena — književnosti, novinarstva, nauke, pravnih i administrativnih izvora. Korpus je razvijen s ciljem normativnog i lingvističkog proučavanja jezika i obuhvata milione riječi.

Iako BHNC nije javno dostupan u CoNLL-U formatu, njegova velika vrijednost leži u tome što može poslužiti kao **izvor neanotiranog teksta** za pretreniravanje jezičkih reprezentacija (npr. tok2vec komponente) ili kao baza za ručnu ili automatsku anotaciju dodatnih podataka za trening `ner` komponente.

Izvor: https://bhnc.izj.unsa.ba/

Faza 2: Pregleda stanja u oblasti

2.1 Opći kontekst i značaj problema

Historijski gledano, većina modela i NLP alata za južnoslavenske jezike bili su ograničeni na osnovne zadatke i zasnivani na pravilima (rule-based pristupi) ili na manjem broju ručno anotiranih podataka. Neki od najvažnijih baseline-ova i pristupa su navedeni u sljedećoj sekciji. Vrijedi pomenuti da spomenuti radovi u nastavku nisu konkretno obrađivali bosanski jezik, ali su formirani korpusi i modeli za hrvatski i srpski jezik relevantni i za jezički model bosanskog jezika zbog jednostavnog transfera znanja između ta tri jezika.

2.2 Historijski razvoj južnoslavenskih jezičkih modela

2.2.1 Korištenje Universal Dependencies korpusa

Hrvatski i srpski jezik su uključeni u Universal Dependencies (UD) inicijativu [1]. Na taj način je kreiran *treebank* ovih jezika koji omogućava robusnu implementaciju POS tagova, morfoloških značajki i sintaksičkih zavisnosti.

2.2.2 Setovi pravila iz projekta MULTEXT-East

Hrvatski i srpski jezik su uključeni u projekte poput MULTEXT-East [2], koji su definisali morfološke oznake i oznake dijelova govora, ali nisu obezbijedili sveobuhvatan annotirani korpus potreban za treniranje modernih modela. Konkretno MULTEXT-East je odgovoran za paralelni korpus ovih jezika baziran na Goerge Orwellovom romanu “1984”.

2.2.3 Definisanje NER pravila za južnoslavenske jezike

NER pravila za južnoslavenske jezike su obuhvaćena definisanjem istih pravila za slovenski jezik u [3]. Tu je definisano pet kategorija imenovanih entiteta:

- osoba, PER
- izvedena osoba, DERIV-PER
- lokacija, LOC
- organizacija, ORG
- razno, MISC

2.2.4 Spacy model hrvatskog jezika

Kulminacija prijašnjih radova na jezičkim modelima južnoslavenskih jezika je hr500k [4], korpus koji je korišten za kreiranje Spacy hrvatskog jezičkog modela. Koristeći prijašnja dostignuća kod UD, MULTEXT-East i Janes-NER projekata kreiran je sveobuhvatan pipeline koji uključuje sljedeće komponente: *tok2vec, tagger, morphologizer, parser, lemmatizer, senter, attribute_ruler, ner*

2.3 Sažetak postignutih rezultata

Naziv modela/rada	Obuhvaćene komponente	Broj tokena
UD_Croatian-SET	`tagger`, `morphologizer`, `parser`	151,226
MULTEXT-East Serbian	`morphologizer`, `ner` (djelomično)	cca. 100,000
Janes-NER	`ner`	nije primjenjivo
hr500k	`tok2vec`, `tagger`, `morphologizer`, `parser`, `lemmatizer`, `senter`, `attribute_ruler`, `ner`	500,000

2.4 Zaključak

Dosadašnja postignuća su dobar temelj za formiranje bosanskog jezičkog modela, zbog već postojećeg modela za hrvatski jezik. Trenutna ideja za implementaciju je da se koristi hrvatski model kao ground truth za kreiranje modela bosanskog jezika. Inicijalno bi se za model bosanskog jezika koristila konfiguracija pipeline-a hrvatskog modela kroz koji bi prošao dio, a kasnije eventualno i čitav BHNC dataset u svrhu treniranja bosanskog modela (ovaj dio je urađen u sklopu projekta za kurs). Nakon toga bi se inferencije koje se razlikuju za bosanski i hrvatski model ručno označavale (ovaj dio nije obuhvaćen u ovome kursu). Poslije ovoga

postupka će se treniranje ponovno izvršiti ali tada bi se koristio ažurirani pipeline prilagođen specifičnostima bosanskog jezika.

Faza 3: Izbor analiza i pretprocesiranje dataset-a

Google Drive link projekta: VI_Projekat_Tim35

3.1 Izvor skupa podataka

Raw_data

3.2 Analiza

Za implementaciju Spacy modela bosanskog jezika korištenje kvalitetnog dataseta za treniranje modela je od ključnog značaja. S obzirom da nema mnogo značajnih istraživanja u smjeru NLP modela za bosanski jezik, izbor je bio ograničen na jedini nama poznat jezički korpus za bosanski jezik - Bosanskohercegovački nacionalni korpus (BHNC). Ovaj jezički korpus se još uvijek prikuplja i sastoji se od tekstova iz različitih žanrova i registara uključujući književne časopise, novine, naučne i stručne tekstove, te internet portale. Trenutno ovaj korpus sadrži 248.767.271 riječi distribuisanih na osnovu više izvora i po različitim godinama kada su prikupljeni podaci objavljeni. Prikaz distribucije podataka po vrsti publikacije i godini objavljivanja je prikazan na slici 3.1

Slika 3.1: Distribucija BHNC podataka po vrsti i godini objavljivanja publikacija

Iako je BHNC kvalitetan jezički korpus koji se definitivno planira koristiti u kasnijim fazama ovoga projekta nakon samog kursa, on sadrži previše podataka za treniranje modela za fazu projekta obuhvaćenu projektom ovoga kursa. Iz toga razloga je za treniranje projektnog modela korišten dataset prikupljen u svrhu dodavanja u BHNC od strane autora ovoga projekta. Ovaj dataset se sastoji od 3 tekstualna file-a koji predstavljaju rezultat web i pdf scrapinga obavljenog nad više izvora. Prvi izvor je web scraping portala banjaluka.com. Jedan od autora ovoga projekta je radio čitav web scraping ovoga portal međutim zbog prevelike količine scrapeanih podataka korišten je manji dio ovoga scrapea. Sljedeći csv file je kompletni web scraping portal cazin.net. Konačni csv file korišten za treniranje je pdf scraping godišnjaka Filozofskog fakulteta Sveučilišta u Mostaru gdje je scrapeano preko 80 članaka na razne teme iz društvenih nauka.

Iako je najjednostavnija metoda koju smo mogli koristiti bila da koristimo čitav web scrape portala banjaluka.com ipak smo se odlučili da koristimo raznovrsnije izvore podataka da možemo uzeti u obzire sve lingvističke varijacije koje se mogu pronaći u različitim dijelovima Bosne i Hercegovine. Ovakav postupak povećava kompleksnost analize ali omogućava da model bude robusniji i kvalitetniji.

Broj tokena po svakom file-u:

- Dio banjaluka.com web scrapinga - oko 1.480.000 tokena
- Web scraping cazin.net portala - 1.137.675 tokena
- Pdf scraping godišnjaka Filozofskog fakulteta Univerziteta u Mostaru - 113.263 tokena

Svekupno to čini 2.730.938 tokena za čitav dataset

Količina podataka po svakom file-u:

- banjalukacom_web_scraping - 9.4 MB
- cazinnet_web_scraping - 7.4 MB
- Sveuciliste_u_mostaru_pdf_scraping - 685 KB

Što čini da je ukupni dataset veličine 17,485 MB.

3.3 Metode pretprocesiranje podataka

Kako bi se trenirao kvalitetan jezički model, neophodno je prethodno izvršiti detaljno i pažljivo **pretprocesiranje podataka**. Ovo podrazumijeva čišćenje, segmentaciju, leksičku i sintaktičku analizu sirovih tekstova, kao i konverziju u format pogodan za treniranje u SpaCy okviru. Sljedeći koraci su korišteni za obradu sirovih podataka:

Ekstrakcija i čišćenje članaka

Ulazni tekstovi dolaze iz strukturiranih tekstualnih datoteka koje sadrže više članaka odvojenih posebnim markerima (***). Zbog toga je bilo potrebno izvršiti sljedeće:

- identifikovala početke članaka na osnovu zaglavlja poput AUTOR(I):,
- izdvojiti tijelo članka, isključujući meta-informacije,
- pripremiti tekst za lingvističku obradu.

NLP obrada rečenica i anotacija

Nakon segmentacije teksta u članke, koristi se unaprijed trenirani model za hrvatski jezik (hr_core_news_sm) kako bi se dobile osnovne lingvističke informacije za svaku rečenicu:

- **tokenizacija** (razdvajanje rečenica na riječi),
- **lema** (osnovni oblik riječi),
- **upos/xpos oznake** (gramatičke kategorije),
- **morfološke osobine** (broj, rod, padež, itd.),
- **zavisnosna sintaksa** (glagolske i imenske veze među riječima),
- **imenovani entiteti** (entiteti kao što su imena osoba, mjesta, organizacija i sl.).

Svi ovi podaci se pretvaraju u standardizovani **CoNLL-U format**, koji je proširen dodatnim oznakama za prepoznavanje imenovanih entiteta (NER=B-PER, NER=I-LOC, itd.) i informacijom da li riječ ima razmak nakon sebe (SpaceAfter=No).

Formiranje dataset-a u SpaCy formatu (.spacy)

Nakon kreiranja *.conllu datoteke manuelnim pretvaranjem rezultirajućeg .txt filea u .conllu jer je .txt file formatiran kao .conllu, vrši se konverzija u binarni .spacy format, što uključuje:

- parsiranje svake rečenice u CoNLL-U formatu,
- ponovno kreiranje Doc objekata uz očuvanje svih lingvističkih informacija,
- eksplicitno mapiranje token.dep_, token.head, token.tag_, token.lemma_, token.pos_ i morfoloških oznaka,
- rekonstrukciju entitetskih oznaka u doc.ents putem Span objekata.

Konačni rezultat je .spacy datoteka spremna za treniranje, uz očuvanje svih korisnih informacija potrebnih za učenje složenih NLP zadataka.

Konačno se dataset dijeli na trening i test split i to u omjeru 70-30. Split je izvršen random odabirima tako da svako ponavljanje skripte za split podataka može dati različite rezultate (bez veće razlike).

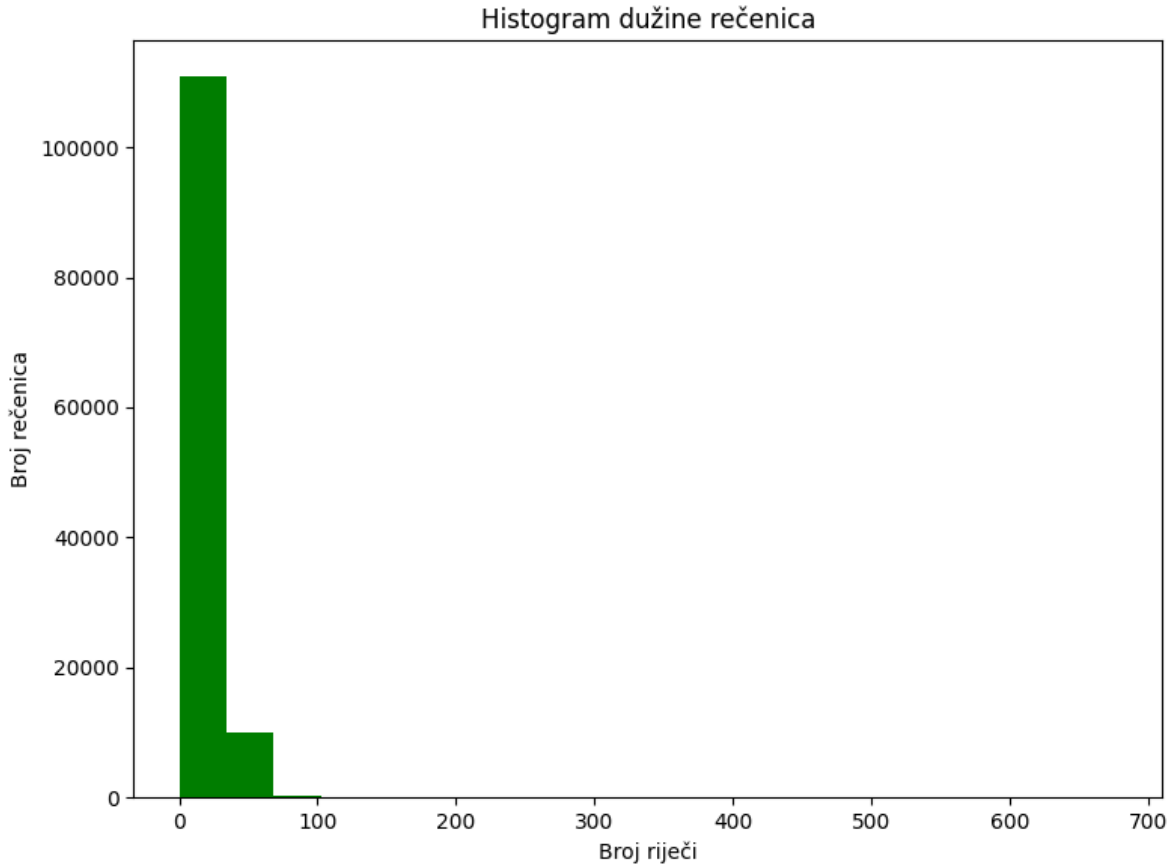
Organizacija foldera

Podaci su organizovani u sljedeće strukture unutar foldera `BS_TrainingData`:

- `/Raw_data/` – sadrži sirove članke (originalni tekst),
- `/Processed_data/` – sadrži generisani CoNLL-U i spacy dataset,
- `/Spacy_ready_data/` – sadrži `*.spacy` datoteku ta treniranje modela koja nije podijeljena na train i test dio,
- `/Spacy_proper_split_data/` – sadrži `*.spacy` datoteke za treniranje i testiranje modela

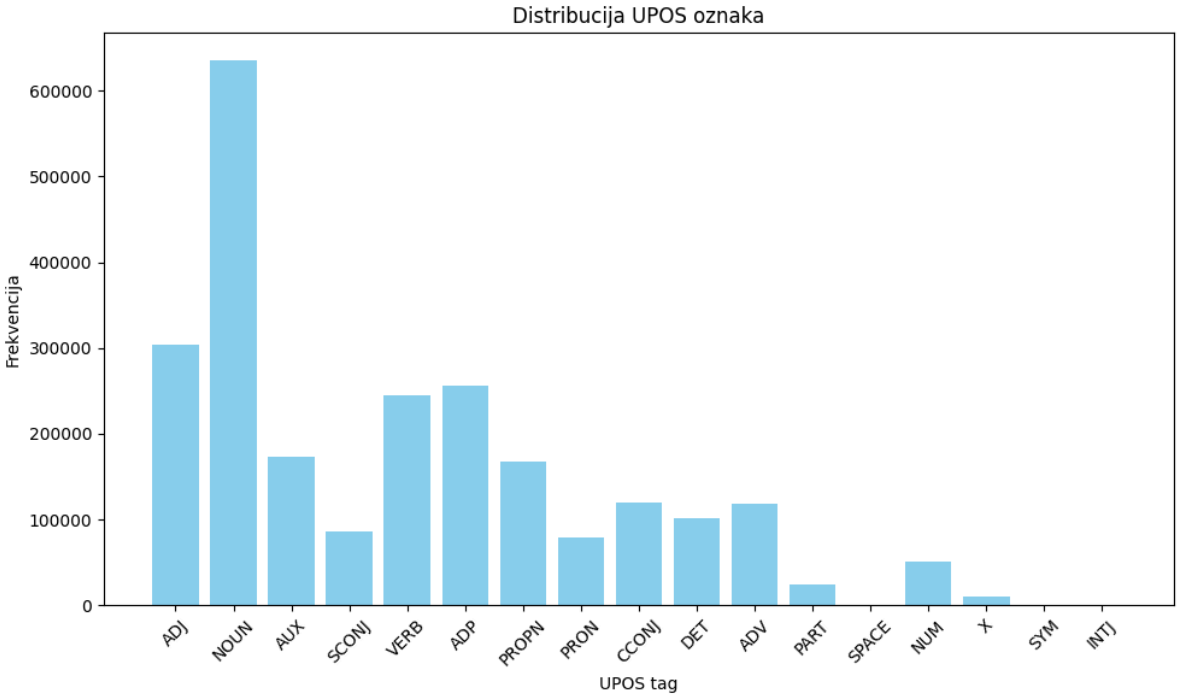
3.4 Analiza procesiranog dataseta

Kako bi se stekla dublja razumijevanja strukture i karakteristika korištenog korpusa za treniranje SpaCy modela bosanskog jezika, provedena je vizuelna analiza procesiranog .conllu dataseta. Vizualizacije su odabrane s ciljem da obuhvate osnovne jezične karakteristike, distribucije i potencijalne nepravilnosti unutar skupa podataka. Na osnovu te analize moguće je donijeti zaključke o uravnoteženosti, bogatstvu i kompleksnosti skupa, te eventualno donijeti odluke o dodatnim koracima obrade ili augmentacije podataka. Prije ove analize su izbačeni interpunkcijski znakovi jer nisu relevantni za ovu analizu.



Slika 3.4.1: Histogram dužine rečenica

Vizualizacija na slici 3.4.1 prikazuje raspodjelu dužine rečenica. Većina rečenica u korpusu sadrži između do 50 riječi, što je tipično za književno-publicističke tekstove i online članke. Općenito, prisutnost ekstremno kratkih i ekstremno dugih rečenica može utjecati na performanse modela, posebno pri učenju zavisnosti.

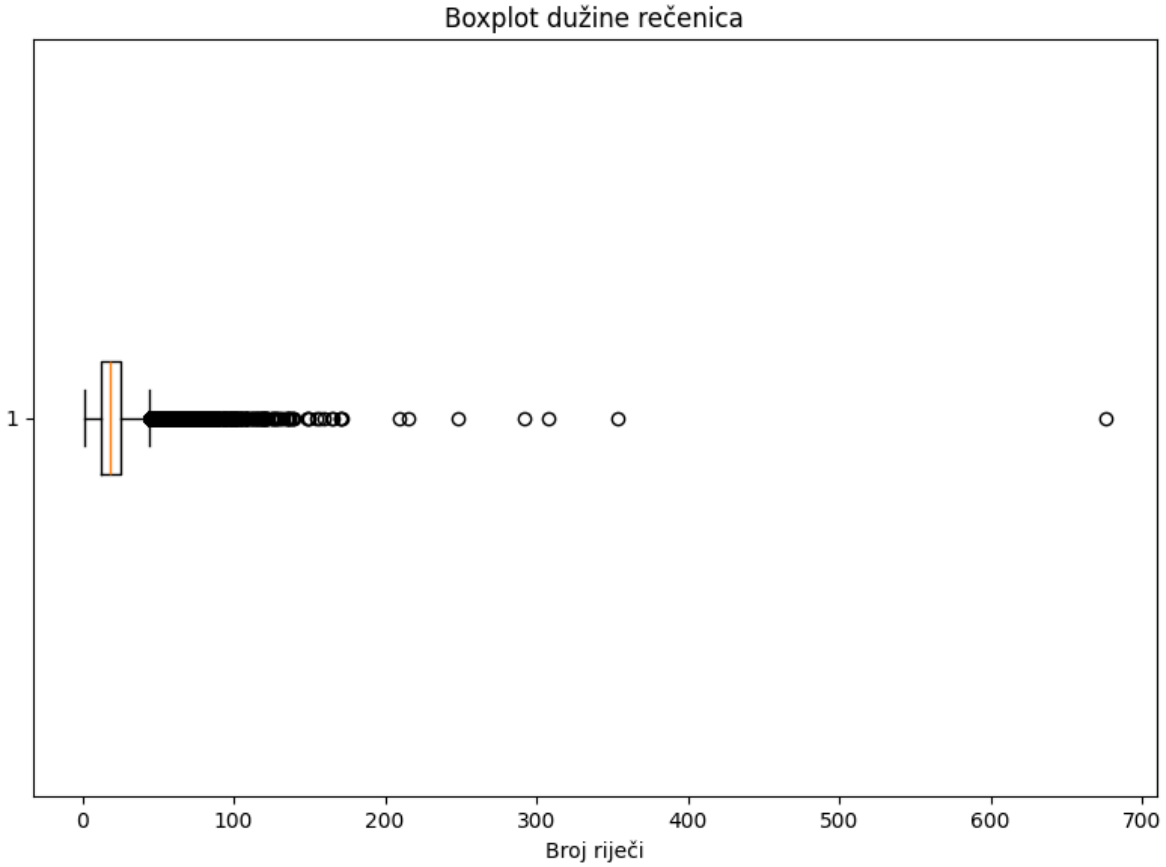


Slika 3.4.2: Histogram učestalosti POS tagova

Na slici 3.4.2 prikazana je učestalost različitih univerzalnih POS tagova (npr. NOUN, VERB, ADJ). Dominacija imenica potvrđuje narativni karakter teksta, dok je ravnomjerna prisutnost ostalih kategorija pokazatelj jezične raznolikosti u datasetu.

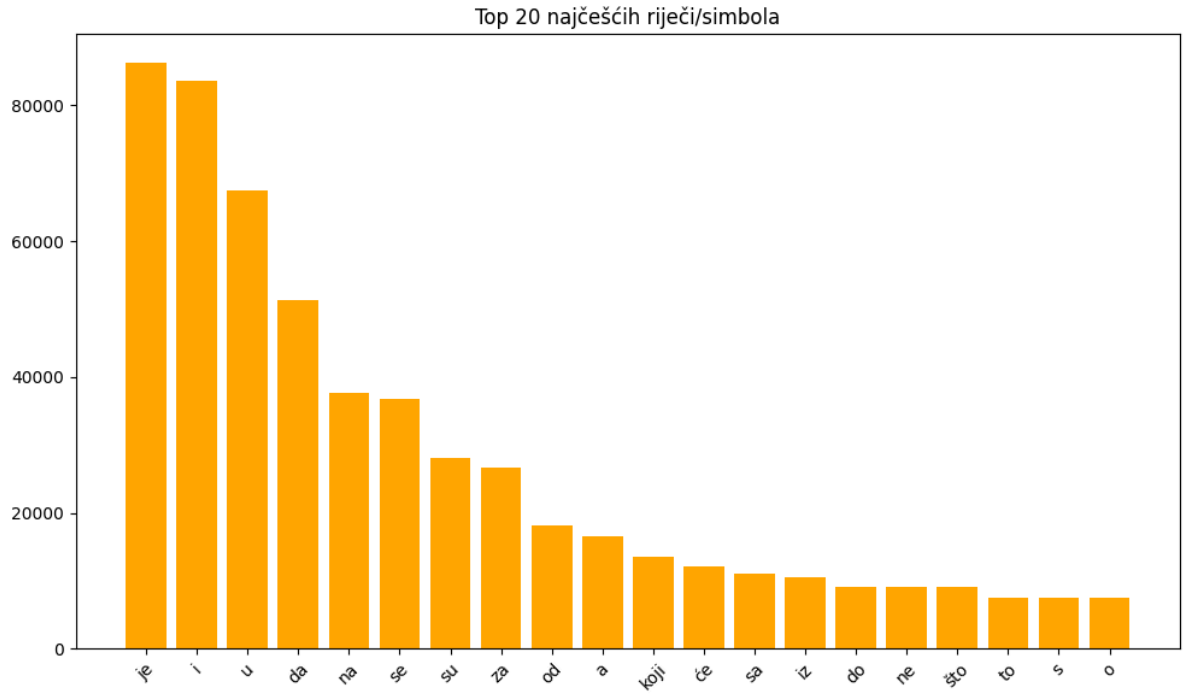
Slika 3.4.3: Word Cloud najfrekventnijih riječi

Oblak riječi (bez stop riječi) omogućava brz uvid u najčešće leme. Dominiraju riječi poput “da”, “na”, “je”, što su tipični pomoćni glagoli u bosanskom jeziku.



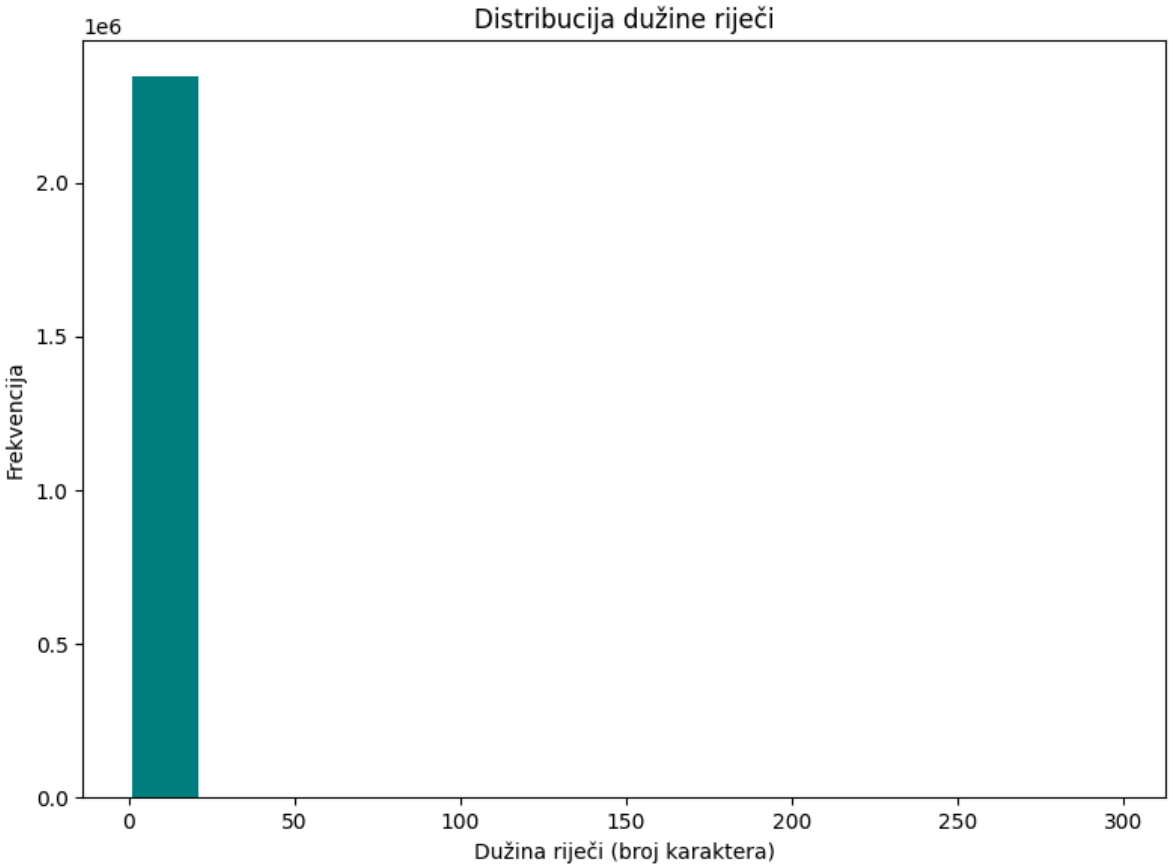
Slika 3.4.4: Boxplot broja riječi po rečenici

Boxplot na slici 3.4.4 vizualizira distribuciju dužine rečenica, s naglaskom na outliere. Vidljivo je da su rečenice uglavnom srednje dužine oko 30-40 riječi, ali s nekoliko izrazito dugih koje mogu otežati parsiranje. Ovo je korisno za odluke o eventualnom "truncatingu" ili segmentiranju. Outlier koji imaju preko 200 riječi su naizgled rezultat pdf scrapinga fajlova u kojima su se nalazile tabele koje su interpretirane kao jedna rečenica.



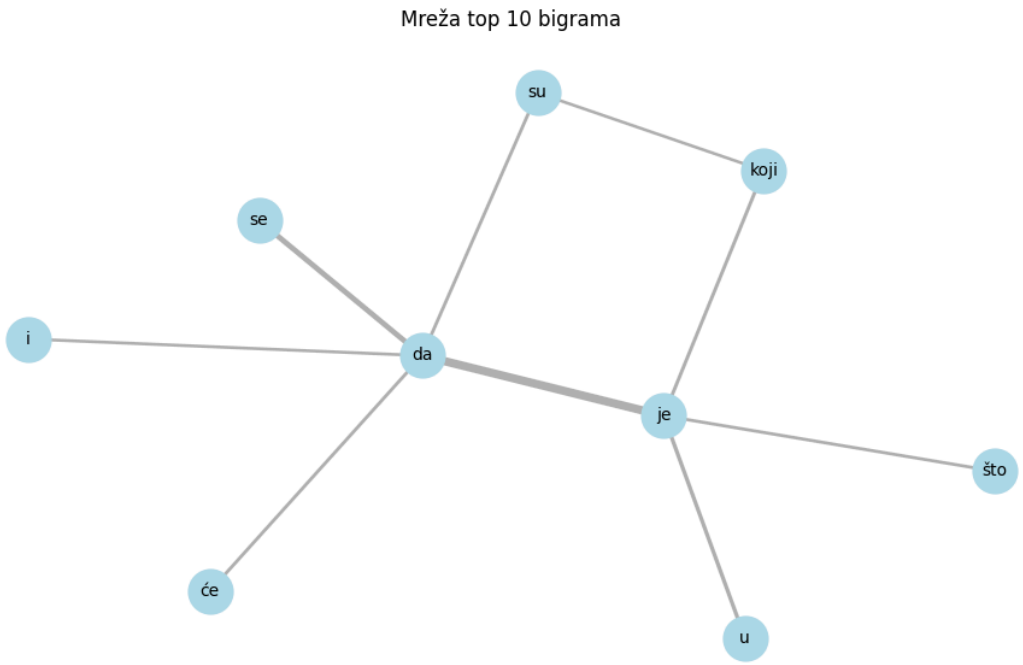
Slika 3.4.5: Histogram najčešćih riječi

Na slici 3.4.5 je vidljivo da kao i na word cloudu pomoćni glagoli i veznici dominiraju u datasetu.



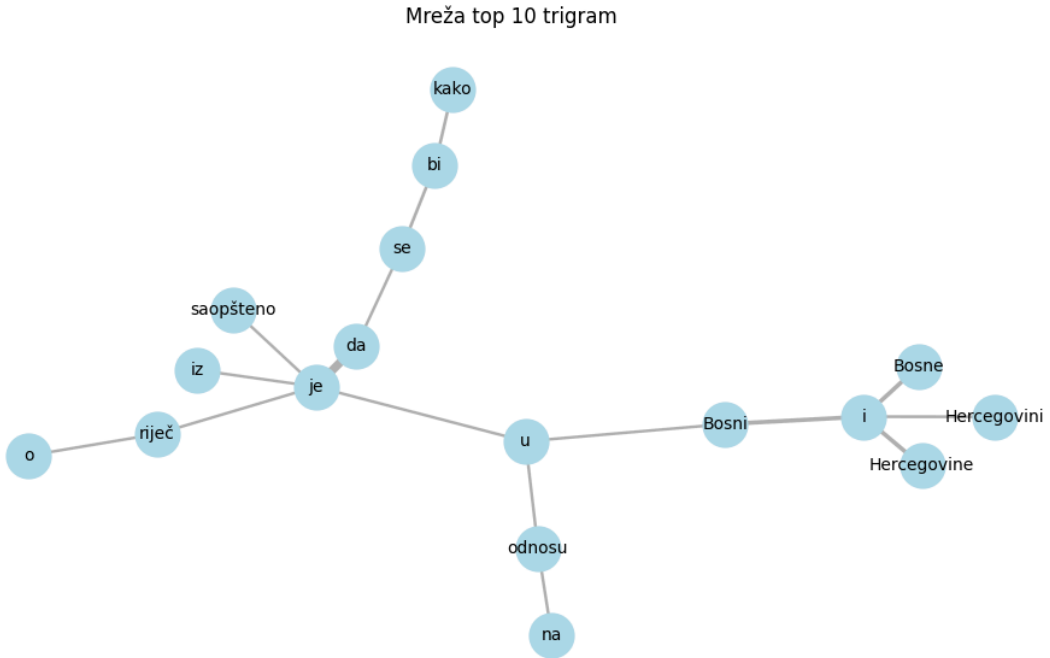
Slika 3.4.6: Raspodjela dužine riječi (broj karaktera)

Kao mjera jezične kompleksnosti, prikazana je distribucija dužine riječi. Većina riječi sadrži između do 20 karaktera, ali prisutni su i dugi izrazi koji mogu predstavljati složenice ili imenice s nastavcima. Outlier naravno uključuju anomalije u scrapeanim podacima kao što su tabele, grafikoni i sl.



Slika 3.4.7: Mreža top 10 bigram

Bigrami poput “da je”, “da se”, “što je” često se javljaju, pokazujući obrasce karakteristične za tematiku i stil dataseta. Ove informacije su ključne za eventualni dodatni jezički modeling poput tekstualne generacije.



Slika 3.4.8: Mreža top 10 trigram

Trigrami dodatno pojačavaju kontekstualno razumijevanje — fraze kao “kako bi se”, “u odnosu na” pomažu prepoznati gramatiku i sintaktičke obrasce koje model treba da nauči.

Faza 4: Odabir, formiranje, treniranje i testiranje modela

4.1 Odabrana metoda i proces treniranja modela

Za razvoj modela je korišten set komandi koji je dostupan out-of-the-box od strane **SpaCy**-a, što skriva detalje modela, ali pojednostavljuje njegovu konfiguraciju.

SpaCy ima podršku za definisanje konfiguracije modela pomoću `config.cfg` fajla unutar kojeg definišemo sve pojedinačnosti modela.

U nastavku slijedi detaljniji opis arhitekture korištene za treniranje modela. Sama arhitektura je bazirana na hrvatskom Spacy jezičkom modelu uz eventualne razlike u hiperparametrima. Sama arhitektura za treniranje je specifična za svaku komponentu pipeline-a uz neke zajedničke atribute.

Modeli komponenti koriste `HashEmbedCNN` arhitekturu, koja je efikasna i pogodna za srednje velike datasetove. Sve komponente koriste zajednički podmodel `tok2vec`, što omogućava dijeljenje parametara i efikasnije učenje.

Trening parametri:

- `batch_size = 128`
  `dropout = 0.1`
- `max_steps = 20000`
- `eval_frequency = 200`
  `optimizer = Adam`

Slijedi opis modela po komponenti:

`tok2vec` model koristi:

Convolutional neural network (CNN) slojeve:

- `depth = 2` (dva konvolucijska sloja)
  `width = 64` (dimenzija konvolucijskih filtera)
- `window_size = 1` (lokalni kontekst)
- `maxout_pieces = 2` (broj Maxout aktivacija po sloju)

`subword_features = true` omogućava učenje karakteristika iz n-grama znakova, što je korisno za slavenske jezike sa bogatom morfologijom.

`tagger` koristi:

- Ulazne reprezentacije iz `tok2vec` modela.
- Izlazne dimenzije (`nO`) se automatski određuju prema broju tagova.
- Trening se vrši pomoću negativnog log-verovatnoćnog gubitka.
- Aktivacija je softmax.
- Ključna metrika: `tag_acc`.

`morphologizer` - ova komponenta koristi istu arhitekturu kao i `tagger`, ali predviđa **morfološke oznake** (npr. `Case=Nom|Gender=Fem|Number=Sing`). Ove oznake se treniraju kao višeklasna klasifikacija.

- `overwrite = true` omogućava ovoj komponenti da prepise postojeće morfološke anotacije ako ih ima.

- Ključna metrika: `morph_acc`.

`parser` - SpaCy koristi **transition-based parser**, koji uči niz akcija za kreiranje zavisnosne strukture nad stablom rečenice. Karakteristike:

- `hidden_width = 128`: širina skrivenih slojeva
- `maxout_pieces = 3`: broj Maxout aktivacija po sloju
- Koristi `tok2vec` kao ulaz.
- `dep_uas` i `dep_las` su glavne metrike

`ner` - NER koristi istu osnovnu arhitekturu kao parser, ali sa `state_type = "ner"`.

- `hidden_width = 64`: manja širina zbog kompaktnosti
- Metrika: `ents_f` (F1 score za entitete)

`lemmatizer` - Za potrebe ovog modela koristi se jednostavni **leksički lematizator** koji radi na principu rječnika. Ovo je efikasno ako je unaprijed poznata lista lematiziranih oblika. Može se zamijeniti ML lematizatorom za složenije slučajeve.

- Metrika: `lemma_acc`

`senter` koristi istu arthitekturu kao `tagger` i `morphologizer.`

`attribute_ruler` je netrenabilna komponenta.

Za optimizaciju se koristi Adam optimizer s `warmup_linear` schedulerom:

- `initial_rate = 5e-5`
- `L2 = 0.01`
- `grad_clip = 1.0`
- `accumulate_gradient = 3`

Definisana je i strategija podešavanja stope učenja linear learning rate warmup and decay sa sljedećim parametrima:

- `warmup_steps = 250`
- `total_steps = 20000`
- `initial_rate = 0.00005`

Model se trenira 20.000 koraka (`max_steps = 20000`), s validacijom svakih 200. U `score_weights` sekciji, metrikama se dodjeljuju težine kako bi se osigurala ravnoteža između komponenti.

Skoriranje i evaluacija:
Metrički sistem vrednuje tačnost pojedinih komponenti modela uz različite težine, metrike označene sa 0 ne utiču na konačni score pa nisu ovdje spominjate:

- `tag_acc = 0.17` (POS tagging)
- `lemma_acc = 0.17` (lematizacija)
- `ents_f = 0.17` (NER F1 skor)
- `sents_f = 0.17` (segmentacija rečenica)
- `pos_acc = 0.08` – Alternativna POS tačnost za poređenje POS tagova u različitim standardima
- `morph_acc = 0.08` – Tačnost morfoloških atributa
- `dep_uas = 0.08` – Unlabeled Attachment Score, ocjenjuje koliko su tačno predviđene veze u sintaksičkom stablu
- `dep_las = 0.08` – Labeled Attachment Score, striktnija od prijašnje jer se mora poklopiti i glava (head) i tip odnosa (deprel).

4.2 Testiranje modela i izvještaj

S obzirom da je koristen 70-30 split, testiranje modela je izvršeno nad setom od 2.730.938 * 0.3 = 819.281 tokena. S obzirom da se koristi Spacy konfiguracija za treniranje, pregled rezultata je ograničen onim što Spacy vraća kao rezultat treniranja, a to su vrijednosti funkcija gubitka i metrike za svaku komponentu po epohi.
U nastavku slijedi izvještaj treniranja po svakoj metrici koja je uticala na ukupni score modela, te ukupni score:

TAG_ACC	86.84 %
POS_ACC	94.82 %
ENTS_F	67.60 %
MORPH_ACC	87.45 %
DEP_UAS	79.23 %
DEP_LAS	71.83 %
SENTS_F	91.94 %
LEMMA_ACC	80.91 %
SCORE	0.82

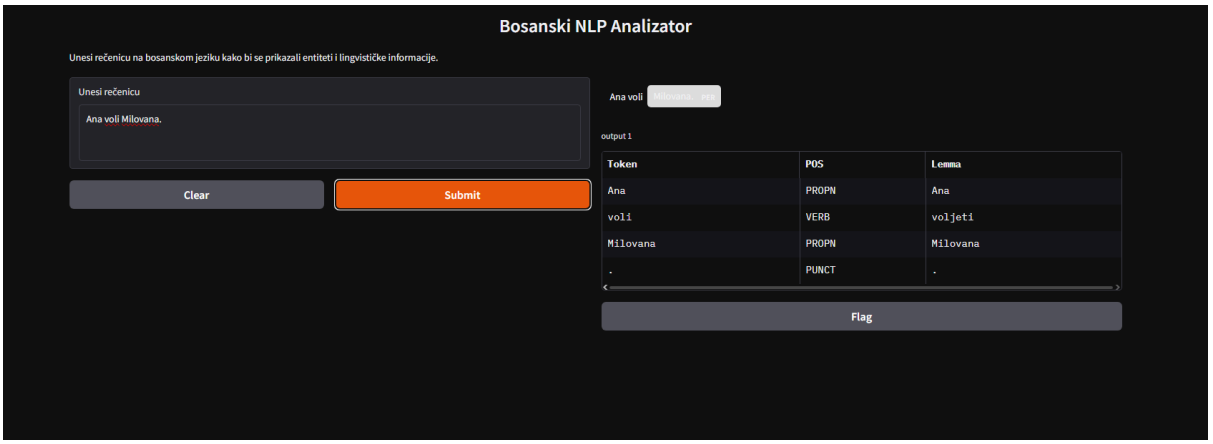
4.3 Diskusija dobijenih rješenja i identifikovani rizici

Dobijeni rezultati pokazuju da razvijeni jezički model za bosanski jezik na zadatku obuhvata više ključnih komponenti obrade prirodnog jezika, pri čemu ostvaruje uravnotežene performanse na većini metrika. Iako postoje jasne snage, identifikovane su i potencijalne slabosti i rizici koji mogu uticati na upotrebljivost modela u praksi.

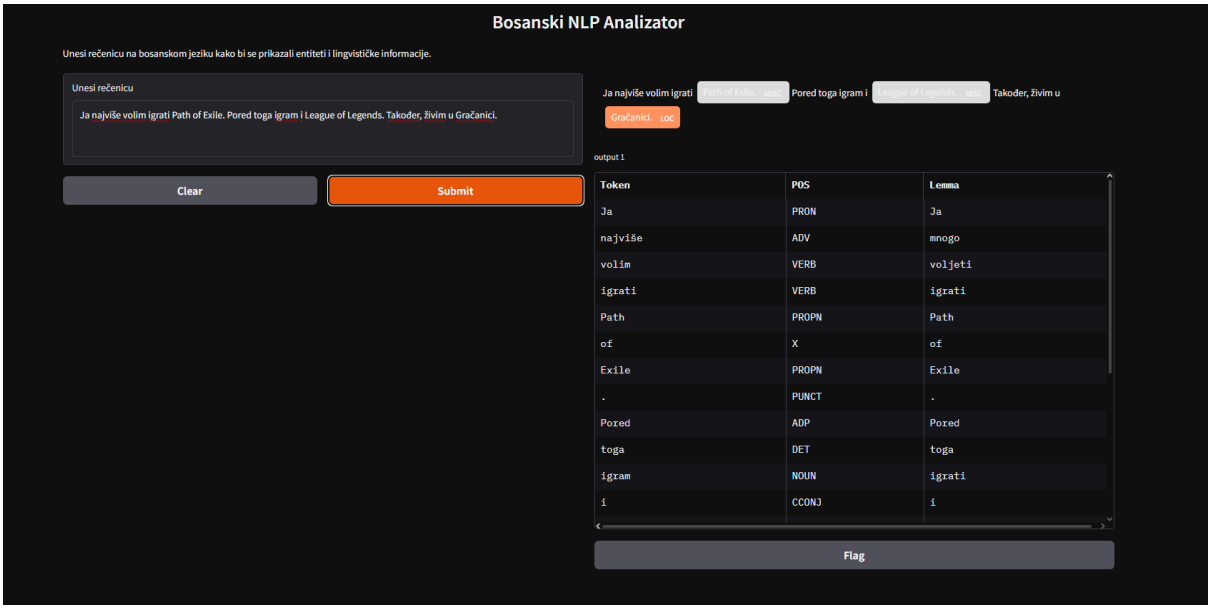
Naime, sve metrike uglavnom pokazuju zadovoljavajuće rezultate, jedino se ENTS_F metrika za NER ističe kao dosta manja. Ovo je najslabija komponenta, što je očekivano, jer NER obično zahtijeva veliku količinu precizno označenih podataka. Ipak, rezultat je upotrebljiv za mnoge osnovne aplikacije.

4.4 Inferencija modela

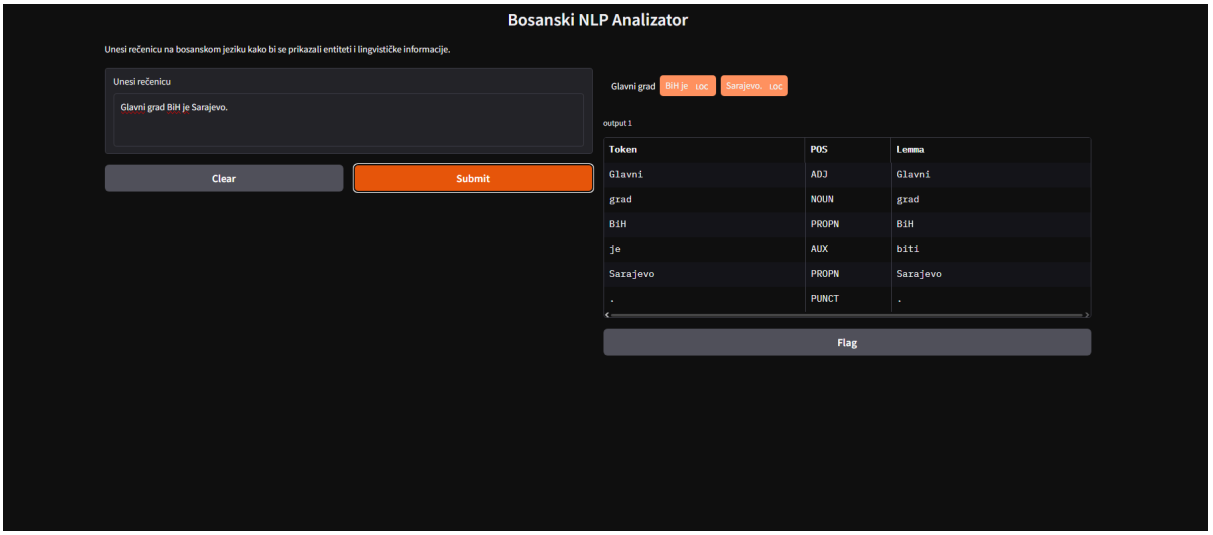
U svrhu testiranja inferencije modela napravljen je UI koji vizuelno prikazuje rezultate testiranja modela na neviđenim podacima.



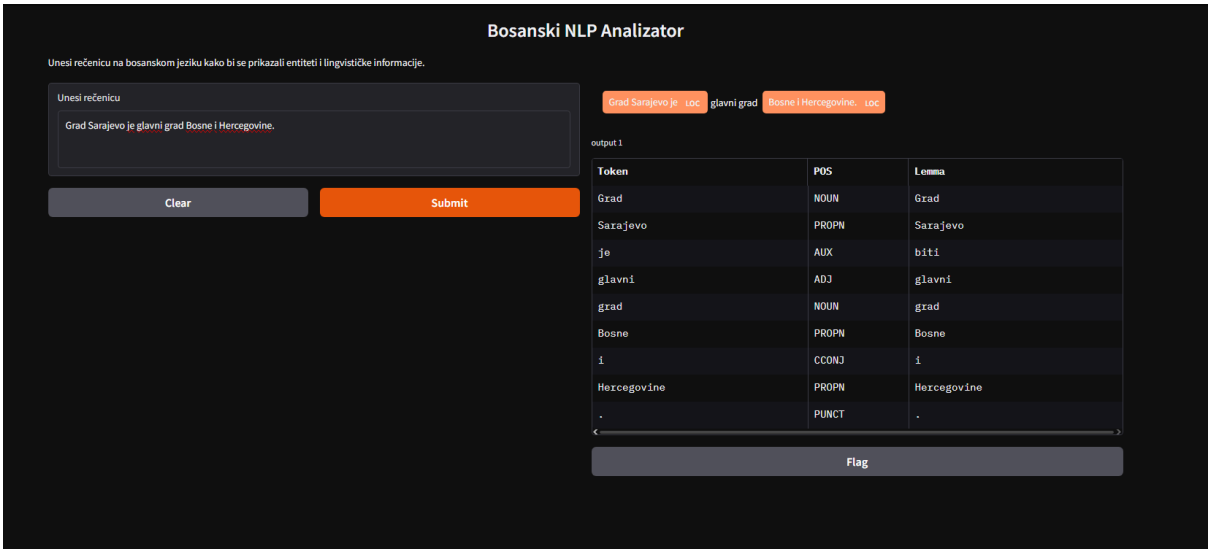
Slika 4.4.1: tipični lingvistički test



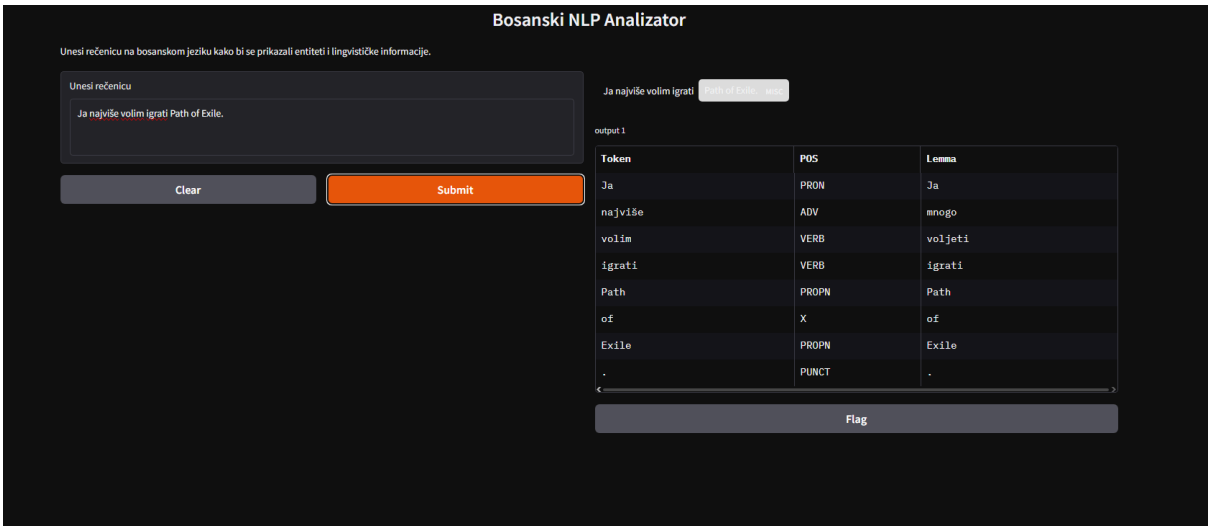
Slika 4.4.2: test na rečenici sa lokacijskim NER-om



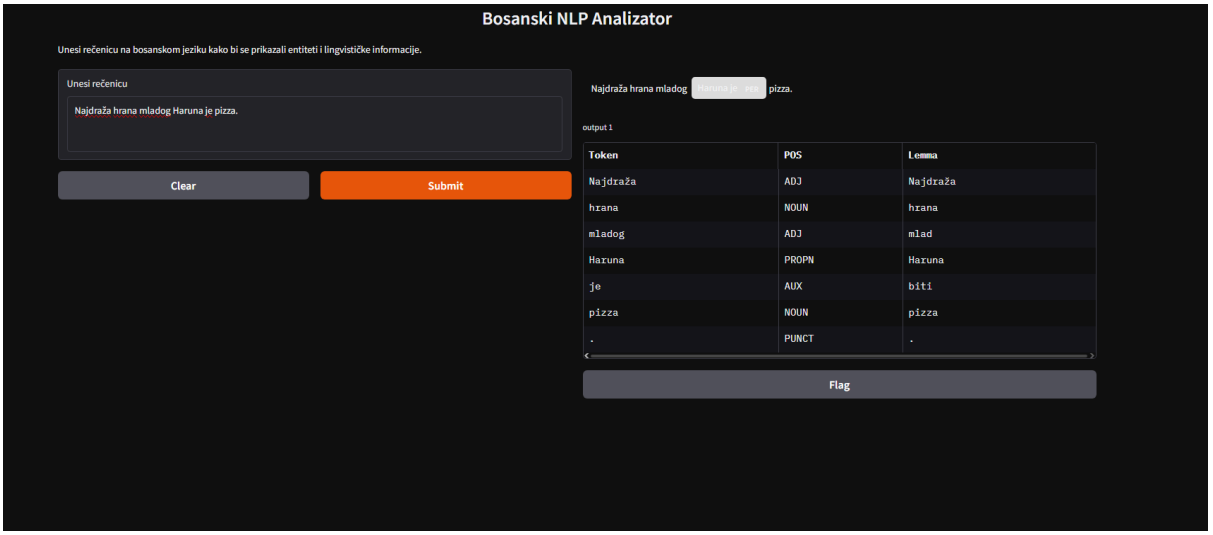
Slika 4.4.3: test kompleksnije NER inferencije



Slika 4.4.4: uočena neispravna inferencija NER komponente



Slika 4.4.5: inferencija na rečenici koja nije slična po tematici kao trening podaci



Slika 4.4.5: dodatna inferencija na rečenici koja nije slična po tematici kao trening podaci

Faza 5: Cjelokupni osvrt na problem i dobijeno rješenje

5.1 Postignuti rezultati

U okviru ovog rada uspješno je razvijen jezički model zasnovan na arhitekturi *spaCy v3*, prilagođen za potrebe bosanskog jezika, koristeći raspoloživi model hrvatskog jezika i domaće anotirane tekstove za treniranje. Rezultati koji su postignuti tokom procesa treniranja i evaluacije pokazuju da je model sposoban za preciznu obradu jezika u više ključnih dimenzija prirodne jezičke analize.
Najvažniji rezultati su prikazani u sekcijama 4.2 i 4.3 gdje je kroz više metrika prikazano koliko kvalitetno radi svaka komponenta jezičkog modela. Evaluacija na ručno unesenim rečenicama potvrdila je da model generalizuje i na stvarne, dosad neviđene sadržaje.

5.2 Prostor za napredak i dalji rad

Na osnovu metrika i ručnog testiranja se da zaključiti da bi NER komponenta trebala biti unaprijeđena dodatnim treniranjem sa većim podacima i eventualnim tuniranjem uticaja metrike ENTS_F na konačni score.
Dalji rad na ovome modelu nakon kursa obuhvata dodatno treniranje na BHNC korpusu, te ručna verifikacija rezultata od strane stručnih kadrova iz oblasti lingvistike sa Filozofskog fakulteta UNSA. Nakon toga procesa bi se moglo izvršiti dodatno treniranje sa eventualnim ručno primjećenim greškama, te bi se model mogao spremiti kao zvanični jezički model u sklopu Spacy biblioteke, na čemu autori planiraju nastaviti raditi nakon kursa.
Treniranje nad većim količinama podataka će zahtijevati više računarskih resursa, jer i dosadašnje treniranje na manjoj količini podataka bilo izuzetno izazovno na ličnom računaru i google colab okruženju zbog ograničenih resursa, pogotovo RAM-a.

5.3 Zaključak

Ovaj model predstavlja značajan korak ka stvaranju jezičke infrastrukture za bosanski jezik i postavlja stabilne temelje za dalji razvoj u oblasti obrade prirodnog jezika.

Reference

[1] Agić, Željko, and Nikola Ljubešić. "Universal Dependencies for Croatian (that work for Serbian, too)." *The 5th workshop on Balto-Slavic natural language processing*. 2015.

[2] Erjavec, Tomaž. "Multext-east." *Handbook of linguistic annotation* (2017): 441-462.

[3] Zupan, Katja, Nikola Ljubešić, and Tomaž Erjavec. *Annotation guidelines for Slovenian named entities: Janes-NER*. Technical report, Jožef Stefan Institute, September. Retrieved from https://www. clarin. si/repository/xmlui/bitstream/handle/11356/1123/SlovenianNER-eng-v1 . 1. pdf, 2017.

[4] Ljubešić, Nikola, et al. "New inflectional lexicons and training corpora for improved morphosyntactic annotation of Croatian and Serbian." *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 2016.