Chatbot za prepoznavanje korisničkih namjera zasnovan na modelima

Ferlatti, Aldo

Sveučilište Jurja Dobrile u Puli
Fakultet informatike
Diplomski studij informatike

ALDO FERLATTI

CHATBOT ZA PREPOZNAVANJE KORISNIČKIH NAMJERA ZASNOVAN NA MODELIMA

Diplomski rad

Pula, 2021.

Sveučilište Jurja Dobrile u Puli
Fakultet informatike
Diplomski studij informatike

CHATBOT ZA PREPOZNAVANJE KORISNIČKIH NAMJERA ZASNOVAN NA MODELIMA

Diplomski rad

JMBAG: 0303068849, redovan student

Studijski smjer: Informatika

Predmet: Izrada informatičkih projekata

Znanstveno područje: Društvene znanosti

Znanstveno polje: Informacijske i komunikacijske znanosti

Znanstvena grana: Informacijski sustavi i informatologija

Mentor: doc. dr. sc. Nikola Tanković

Pula, rujan 2021.

IZJAVA O AKADEMSKOJ ČESTITOSTI

Ja, dolje potpisani ALDO FERLATTI, kandidat za magistra INFORMATIKE ovime izjavljujem da je ovaj Diplomski rad rezultat isključivo mojega vlastitog rada, da se temelji na mojim istraživanjima te da se oslanja na objavljenu literaturu kao što to pokazuju korištene bilješke i bibliografija. Izjavljujem da niti jedan dio Diplomskog rada nije napisan na nedozvoljeni način, odnosno da je prepisan iz kojega necitiranog rada, te da ikoji dio rada krši bilo čija autorska prava. Izjavljujem, također, da nijedan dio rada nije iskorišten za koji drugi rad pri bilo kojoj drugoj visokoškolskoj, znanstvenoj ili radnoj ustanovi.

Student

Aldo Ferlatti

U Puli, 07.09.2021

IZJAVA O KORIŠTENJU AUTORSKOG DJELA

Ja, ALDO FERLATTI dajem odobrenje Sveučilištu Jurja Dobrile u Puli, kao nositelju prava iskorištavanja, da moj diplomski rad pod nazivom CHATBOT ZA PREPOZNAVANJE KORISNIČKIH NAMJERA ZASNOVAN NA MODELIMA

koristi na način da gore navedeno autorsko djelo, kao cjeloviti tekst trajno objavi u javnoj internetskoj bazi Sveučilišne knjižnice Sveučilišta Jurja Dobrile u Puli te kopira u javnu internetsku bazu završnih radova Nacionalne i sveučilišne knjižnice (stavljanje na raspolaganje javnosti), sve u skladu s Zakonom o autorskom pravu i drugim srodnim pravima i dobrom akademskom praksom, a radi promicanja otvorenoga, slobodnoga pristupa znanstvenim informacijama.

Za korištenje autorskog djela na gore navedeni način ne potražujem naknadu.

U Puli, 07.09.2021

Potpis

Aldo Ferlatti

Zahvala

Zahvaljujem mentoru doc. dr. sc. Nikoli Tankoviću na pomoći, vodstvu, slobodi te poticanju na rad izvannastavnih projekata koji su doprinijeli mom razvoju i uspjesima.
Zahvaljujem Ivanu Vuliću, Ph.D., na susretljivosti, pomoći i savjetima u ključnim trenucima tijekom pisanja ovog rada.
Hvala mojoj djevojci koja me podržava u svim trenucima, projektima te idejama u koje se uputim.
Te na posljetku, posebna zahvala mojoj obitelji na velikoj potpori, strpljenju i odricanju koje su mi pružali tijekom školovanja i budućem osobnom razvoju.

Sadržaj

Uvod **1**

1 Arhitektura chatbota **4**

2 BPMN **7**

3 NLP **10**

3.1 Predobrada tekstualnih podataka 11

3.2 Tehnika pozornosti i transformeri 13

3.2.1 Pozornost 13

3.2.2 Transformeri 14

3.3 BERT 17

4 Prepoznavanje namjere **19**

4.1 Kodiranje riječi vs kodiranje rečenica 20

4.2 LaBSE 24

5 Praktični problem **27**

5.1 Opis problema 27

5.2 Prikupljanje i analiza podataka 29

5.3 Modeliranje i treniranje 33

5.4 Rezultati testova 37

5.5 Zaključci provedenih testova 40

Zaključak **43**

Literatura **45**

Popis tablica **53**

Popis slika **54**

Sažetak **55**

Abstract **56**

Uvod

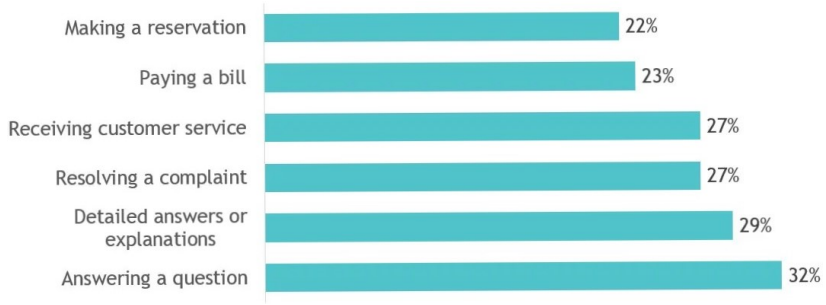
Verbalna komunikacija se smatra jednom od ključnih faktora za socio-psihološko zdravlje. Puno ljudi shvaćaju komunikaciju kao proces između sebe i još jedne osobe. Također, znanstvenici koji proučavaju komunikaciju smatraju je kao proces isključivo za ljude [1]. Biti u mogućnosti komunicirati te uspješno prenijeti svoje namjere putem konverzacije je kompleksan čin te u nekim oblicima moguć samo s međuljudskom interakcijom. Dijeljenje informacija ili razmjena informacija tijekom socijalne interakcije je puno kompleksniji proces od samo dijeljenja riječi s drugom osobom. Proces je višeslojni te obuhvaća stav i utjecaj govornika, koristi tjelesne pokazatelje te također fizičko okruženje gdje se konverzacija razvija. Da bi komunikacija bila uspješna potrebno je ispuniti sljedeće uvjete [2]:

- Sudionik razgovora mora jasno predstaviti svoje namjere za jednostavno prepoznavanje.
- Sadržaj, koncepti ili ideje, odnosno semantičke informacije koje govornik želi prenijeti moraju biti učinkovito predstavljene

Kroz niz godina istraživanja, strojevi i komunikacija smatrali su se kao odvojene domene koje nije moguće spojiti. Nedavni razvoji u domeni umjetne inteligencije (UI) omogućili su veliki napredak u konceptu komunikacije čovjeka i stroja. Istraživanja UI se koncentriraju primarno na reproduciranju raznih aspekata ljudske inteligencije, uključujući sposobnost komunikacije, sa strojnog stajališta. U kontrastu UI, teorije komunikacije gledaju strojeve samo kao tehnologiju koja služi kao mediji prijenosa poruka. UI i ljudska interakcija s njom ne pristaje konceptima teorije komunikacija koja se formirala na idejama komunikacije čovjeka s drugim ljudima. Kao odgovor razlikama teorije komunikacije i uporaba UI, razvilo se novo područje ”komunikacija čovjek-stroj” (*eng, Human-Machine Communication (HMC)*), koje se fokusira na stvaranju smisla između ljudi i stroja s naglaskom na interakciju s komunikacijskom tehnologijom kao što su roboti ili chatbotovi. Guzman je u svom radu [3] dokazao da ljudi percipiraju robote kao komunikacijske partnere, istodobno su svjesni da ne komuniciraju s čovjekom ali uistinu objektom koji posjeduje socijalne aspekte; također je pokazao da interakciju s virtualnim asistentima ljudi smatraju komunikacijom s tehnologijom što pokazuje da je komunikacijska UI je dizajnirana da imitira komunikatora ali je ljudi također percipiraju

kao takvu [4].

U domeni komunikacijske umjetne inteligencije, vjerojatno trenutno najpoznatija tehnologija su virtualni asistenti kao Appleov Siri, Microsoftova Cortana, Amazonova Alexa ili Googleov virtualni asistent. Napredovanje u domeni dubokog učenja omogućilo je drastičan utjecaj na dizajn virtualnih asistenta. Na primjer, razvoj WaveNet arhitekture omogućilo je stvaranje Google Duplexa [5], sustava za telefonske razgovore gdje se razgovara s umjetnom inteligencijom bez da je sugovornik svjestan toga. Iako je namijenjen za male zadatke kao što je stvaranje rezervacija kod frizera, uspjeh tehnologije je ogroman [6]. Kompleksnije zadatke rješava tehnologija razvijena od strane PolyAI [7], vodeća tvrtka za telefonske virtualne asistente koja ima sustave razvijene za razne domene te prihvaća višejezičnost. Kada se promijeni mediji komunikacije s verbalnog na tekstualni mediji, chatbotovi su najpoznatija tehnologija za komunikaciju sa strojem. Jednostavni alati dostupni razvojnim programerima za implementaciju chatbotova unutar svojih rješenja, popularizirali su chatbotove kao zamjena službe za korisnike koja služi za rješavanje čestih problema i upita. Alati kao što je konverzacijska platforma RASA, Flow.ai, Googleov Digitalflow ili Amazonova Lexa, imaju široki ekosustav koji omogućava implementaciju chatbotova i za ljude izvan domene umjetne inteligencije.



Slika 1: Globalna upotreba chatbotova [8]

Neovisno o tehnologiji, alatu ili cilju implementacije konverzacijskog sustava, prvi korak je prepoznati namjeru sudionika razgovora. U konverzacijskim sustavima, prepoznavanje namjere korisnika je neophodan ključ za uspješnu uspostavu interakcije [9]. Kako bi stroj mogao pravilno generirati odgovor, izvući relevantne entitete iz konverzacije te procijeniti sljedeću najbolju opciju, najprije mora razumjeti namjeru korisnika. Problem prepoznavanja namjere iz ulaznih podataka se može iskoristiti za razne aplikacije: sažimanje uobičajene i česte korisnikove ciljeve ili grupiranje funkcija asocirane

s poslovanjem ili proizvodom; može naglasiti i dodijeliti prioritete učestalim greškama i poteškoćama prijavljene od strane službe za korisnike ili javnih foruma; može prepoznati elemente unutar elektroničke pošte ili transkripcija sastanka koje zahtijevaju aktivnu radnju od strane korisnika [10]. Razmatrajući važnosti i prednosti pravilnog prepoznavanja namjera u komunikaciji, ovaj rad se koncentrira na tehnologije i metode klasificiranja rečenica u predodređene namjere.

Ovaj diplomski rad predstavlja dio većeg projekta koji ima cilj dizajniranja i implementiranja arhitekture komunikacijskog alata između čovjeka i računala temeljenim na BPMN modelima, odnosno chatbota temeljenog na BPMN modelima. Uvodnom svrhom, poglavlja 1 i 2 predstavljaju uvod u terminologiju i arhitekturu chatbotova i BPMN modelima. Ostatka rada podijeljen je na sljedeći način: u poglavlju 3 se objašnjavaju koncepti procesiranja teksta te važne tehnike kao što su tehnike pozornosti i transformeri; u poglavlju 4 su opisane tehnike i modeli za uspješno prepoznavanje namjera; zadnje poglavlje 5 je prikazan praktični problem s dobivenim rezultatima i zaključcima gdje se primjenjuje teorija opisana u ranijim poglavljima.

Cijeli projekt, podaci, testovi i arhitektura modela mogu se pronaći na javnom GitHub repozitoriju (link: *https://github.com/AldoF95/intent_recognition_masters_thesis*).

1 Arhitektura chatbota

Chatbot je česti primjer sustava umjetne inteligencije i jedan od jednostavnijih i rasprostranjenijih primjera interakcije između čovjeka i računala. Karakteristika i cilj chatbota je da interakcija izgleda kao interakcija s čovjekom, neovisno da li se izvodi tekstualnim ili govornim putem. Chatbot je temeljna primjena tehnologije procesiranja prirodnog jezika (eng. Natural Language Processing - NLP) [11, 12]. Iako su chatbot sustavi napredni u današnja vremena, još uvijek nisu potpunosti u stanju voditi svakodnevne prirodne konverzacije s ljudima. Postoje nekoliko većih izazova povezana sa stvaranjem naprednih konverzacijskih sustava kao što su chatbotovi: automatsko prepoznavanje konverzacijskog govora (*eng. Conversational Automatic Speech Recognition*) za mogućnost slobodnog govora, razumijevanje prirodnog govora (*eng. Natural Language Understanding*), konverzacijske baze podataka, kontekstualno modeliranje, planiranje dijaloga, generiranje odgovora, generiranje prirodnog jezika (*eng. Natural Language Generation*), detekcija osjećaja, filtriranje neprimjerenog izraza (npr. govor mržnje ili neprimjereni humor) itd. [13].

Povijest: osvrtom na povijest chatbotova, Alan Turing je bio prvi koji je postavio pitanje ”Može li stroj razmišljati?” [14], gdje razmišljanje je definirano kao sposobnost koju ljudi posjeduju. Analogno tome, Turing preporučuje test pod nazivom ”igra imitacije” (kasnije nazvanim Turingovim Testom) koji najviše podsjeća današnjim chatbotovima. Test je dizajniran tako da mjeri uspješnost sustava nalik chatbotovima [15]. Ako osoba nije u stanju ustanoviti da li je na drugoj strani komunikacije stroj ili druga osoba, onda se smatra da je sustav prošao test. Dugi niz godina je prošlo prije nego se pojavio prvi chatbot nakon utemeljenja koncepta od strane Turinga. 1966. godine, Joseph Weizenbaum s MIT-a je stvorio prvi stroj za kojeg se moglo smatrati da približno imitira čovjeka: ELIZA. ELIZA je funkcionirala na principu ključne riječi: iz ulazne rečenice bi identificirala ključne riječi koje bi služile kao ključevi za pretragu unaprijed definiranih pravila za generiranje odgovora [16]. Nakon ELIZA-e tehnologija je značajno napredovala te 1972. godine, Kenneth Colby sa Stanforda razvio je chatbot PERRY koji je imitirao ponašanje shizofreničara [17]. Tek 1995. godine je Richard Wallace napravio chatbot ALICE koji je tada bio najkompleksniji chatbot na tržištu. ALICE je funkcionirao na tehnici podudaranja uzoraka: iz ulaza bi izvukao uzorak te uspoređivao sa spašenim

dokumentom koji je sadržavao parove uzorak ulaza - predložak izlaza. Dokument je bio napisan s AIML (*Artificial Intelligence Markup Language*) jezikom koji se dan danas koristi za neka rješenja chatbotova. Napredak u tehnologijama za dohvaćanje i analizu podataka temeljenima na strojno i duboko učenje omogućilo je razvoj današnjih naprednih virtualnih asistenata (odnosno glasovni chatbotovi) kao što su Appleov Siri, Amazonova Alexa ili Microsoftova Cortana [18].

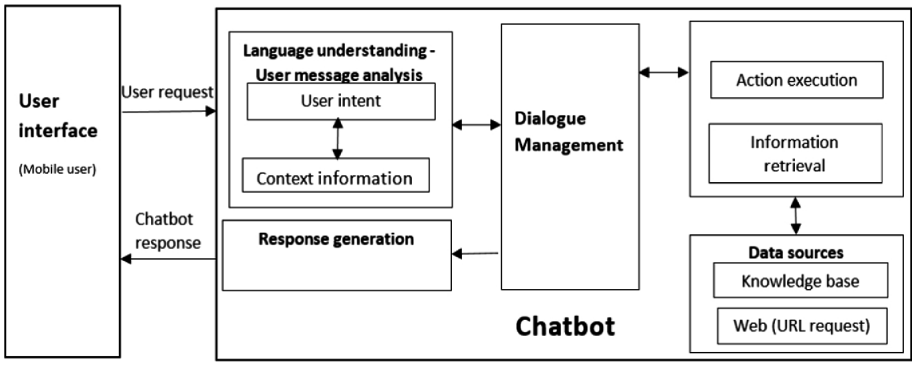
Većina današnjih chatbotova mogu se podijeliti u tri grupe, ovisno o arhitekturi korištene tehnologije za implementaciju: chatboti temeljeni na lingvističkim tehnikama(ili jednostavnije temeljenim unaprijed definiranim pravilima, *eng. rule-based chatbot*), chatbotovi temeljenom umjetnom inteligencijom (*eng. AI chatbot*), te hibridna verzija chatbotova koja koristi elemente obje spomenute vrste [19].

- **Rule-Based chatbot**: ova vrsta je trenutno među češćima u industriji zbog jednostavne logike i sigurnosti vođenja toka razgovora. Ova metoda koristi *if/then* logike za stvaranje konverzacijskog toka. Lingvistički uvjeti se stvaraju prema ključnim riječima, redoslijedom riječi, sinonimima, čestim rečenicama, itd. kako bi se kategoriziralo odgovore za rečenice s istim značenjem. Iako koristi određenu razinu procesiranja ljudskog govora, interakcija je strukturirana i podsjeća na automatizirane *FAQ* sustave.

- **AI chatbot**: temeljeni na tehnikama strojnog učenja, ova vrsta chatbotova su kompleksniji od lingvističkih te više liče na normalan ljudski razgovor. S vremenom uče od korisnika te mogu postati više personalizirani. Primarna mana je što im je potrebna velika količina podataka za treniranje modela te u slučaju kvara teško je intervenirati unutar modela. Zbog navedenih razloga, u industriji ova vrsta chatbotova je rjeđa od jednostavnijih lingvističkih.

- **Hybrid**: hibridna verzija je odgovor na nedostatke gornjih slučajeva, uzima sve prednosti oba sustava te usklađuju se u jedan sustav. Lingvistički pristup omogućava gradnju sustava s minimalnom količinom podataka te uvodi transparentnost u funkcioniranju sustava, dok pristup strojnog učenja omogućava proširenje lingvističkih pravila uvodeći kompleksna sučelja i mogućnost vođenja kompleksnijih razgovora koja više podsjećaju na prirodan ljudski razgovor.

Neovisno o domeni i vrsti implementiranog chatbota, svaki prati osnovnu arhitekturu (Slika 2) koja je sastavljena od elemenata potrebnh za potpuni i ispravan rad chatbota. Međutim, temeljni elementi, bez kojih chatbot ne može funkcionirati, su elementi za prepoznavanje ljudskog unosa koji predstavljaju temu ovog diplomskog rada.



Slika 2: Osnovna arhitektura chatbota [11]

2 BPMN

Prvog puta uvedena 2004. godine od strane BPMI (*Business Process Management Initiative*), *Business Process Management Notation* je od tada postao standard u industriji za dokumentiranje, shvaćanje, modeliranje, analiziranje, simuliranje, izvođenje te ažuriranje poslovnih procesa [20]. Cilj predlaganja BPMN-a je bio stvoriti notaciju koja je jednostavna za shvatiti svim sudionicima poslovnih procesa: od analitičara koji kreira početne nacrte procesa, do razvojnog programera odgovornog za implementaciju tehnologije koja će izvoditi procese, te na posljetku, poslovnim menadžerima koji moraju monitorirati i nadgledati napredak procesa [21]. Definiran je pomoću dijagrama temeljnima na tehnikama dijagramima toka pomoću kojih se stvaraju grafički modeli (Slika 3); početne kategorije elemenata dijagrama su: elementi protoka, spojni elementi, razdvojni elementi (*eng. Swimlines*) i artefakti [21, 22].

- Elementi protoka: ovi elementi predstavljanju srž dijagrama i sastavljeni su od tri jednostavnih oblika.

  Događaji: grafički predstavljeni kao krugovi, podijeljeni u 3 grupe: početak, kraj te posrednici; Označavaju događaj koji se desi tijekom procesa te su popraćeni značajkama okidača događaja i rezultata nakon događaja.

  Aktivnosti: grafički predstavljeni kao pravokutnik te predstavlja rad koji se odvija unutar procesa; postoje dvije bazne vrste, zadatak i pod-aktivnost.

  Prolazi: grafički predstavljeni kao rombovi, označavaju konvergenciju i divergenciju tijeka procesa.

- Spojni elementi: ovi elementi se koriste za spojiti elemente protoka u skladnu osnovno strukturu dijagrama.

  Sekvencijalni spoj: grafički prikazan kao solidna puna strijela te označuje smjer kretanja tijeka procesa, odnosno kojim redoslijedom će se aktivnosti izvoditi.

  Poruka: grafički prikazane kao isprekidana strijela te označuje tijek poruka između aktera unutar procesa.

  Asocijacija: grafički označene kao točkaste strijele, označuju srodnost raznih elemenata, kao što su podaci ili tekst, s elementima protoka.

- Razdvojni elementi: tehnika za organizaciju aktivnosti u različite vizualne kategorije kako bi se prikazalo razne aktere i odgovornosti. Komunikacija između aktera se označuje s elementima poruka.

  Participa (*eng. Pool*): predstavlja aktera unutar procesa te grafički označava aktivnosti za koje je taj akter odgovoran. Inače akteri su u domeni B2B procesa.

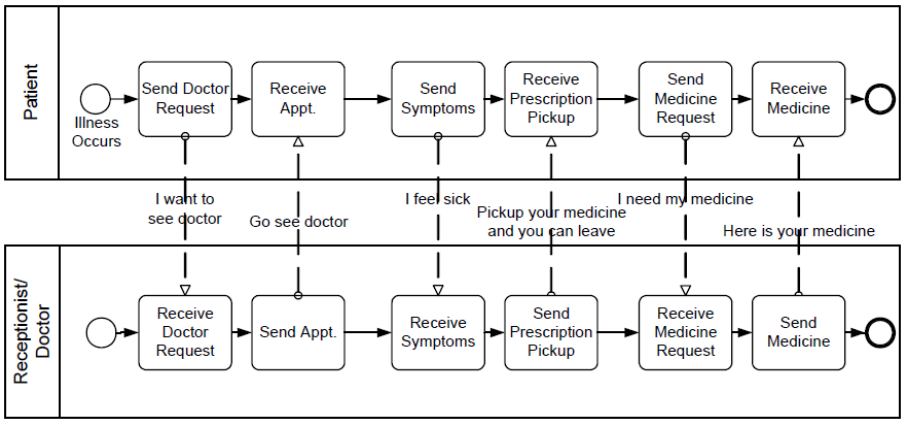
  Traka (*eng. Lane*) participa unutar participe; također se koristi za podijeliti aktivnosti.

- Artefakti: proširuje baznu notaciju s kontekstualnim elementima specifične za situaciju koju se modelira. U dijagram se može dodati proizvoljan broj artefakta sve dok imaju kontekstualnoga smisla. Postoje tri temeljne vrste koje se proširuju u BPMN 2.0 notaciji.

  Podatkovni objekti: spojeni na aktivnosti pomoću asocijacija, označuju potrebu za podacima ili da je rezultat aktivnosti podatkovni dokument.

  Grupe: grafički prikazan kao isprekidan pravokutnik; koristi se kao dokumentacijski alat ili za analizu tijeka ali pritom ne utječe na tijek izvođenja procesa.

  Notacije: komentari koji daju dodatne informacije čitatelju dijagrama.



Slika 3: Jednostavan primjer BPMN dijagrama [21]

Publikacijom BPMN 2.0, standardu se dodalo nove funkcionalnosti kao nativno izvođenje BPMN modela i standardizirani serijski format. Cilj objave nove verzije je bio smanjenje poteškoća koja se dese kada se želi modelirani proces provesti kroz softversku

implementaciju [23]. Unatoč očitim prednostima, istraživanje provedeno 2012.godine, kratko nakon objave verzije BPMN 2.0, pokazalo je da 52% ispitanika koristi notaciju samo za dokumentacijske svrhe, dok tek 37% su koristili za izvođenje poslovnih procesa. Dodatna informacija je da 70% ispitanika izvršavaju dijagrame samostalno kroz nekih od komercijalno dostupnih softverskih rješenja te 36% kažu da koriste notaciju samo za repetitivne poslove [22]. Međutim, iako je uporaba relativno niska, istraživanje provedeno sa strane Sedick i Seymour [24] potvrdilo je da je kod svih ispitanika je uporaba BPMN-a poboljšala odnose između menadžmenta i informatičkog tima: ispitanici su potvrdili da notacija zaista utječe na softversku realizaciju željenih specifikacija menadžmenta na način da informatički tim dobije zahtjeve njima razumljivim jezikom.

3 NLP

Sve što izrazimo, bilo to verbalno ili tekstualno, sadržava veliku količinu informacija, odabir teme, naš ton izražavanja, redoslijed riječi itd. nadodaje jednu vrstu informacija koja se može interpretirati i izvući kvantitativnu vrijednost. U teoriji, s tolikom količinom informacija može se predvidjeti ljudsko ponašanje te profilirati pojedinu osobu [25].

Navedena vrsta informacija predstavlja nestrukturirane podatke. Nestrukturirani podaci su najčešći u stvarnome svijetu jer proizlaze iz entropijskog okruženja te takvi podaci su teški za procesiranje. Nova rješenja u strojnom učenju i dostupnosti velikim količinama podataka, nestrukturirani podaci, kao što je to ljudski govor, postaju sve pristupačnijima. Sa statične analize teksta (pronalaženje ključnih riječi) se napredovalo na kognitivne tehnike gdje se pokušava razumjeti značenje tih riječi. To omogućava procesiranje teksta tako da se može prepoznati ironija u rečenici ili dobiti razinu osjećaja skrivenu u rečenici [25].

Područje koje se bavi kognitivnom analizom teksta se zove *Natural Language Processing* (NLP, ili prevedeno na hrvatski procesiranje prirodnog jezika). NLP je grana umjetne inteligencije koja kao ulaz uzima nestrukturirane podatke govora ili teksta te izlaz je strukturirani oblik teksta prema kojem se izvlače značenja prirodnog jezika (*eng. Natural Language Understanding*)[18].

NLP ima široku primjenu: ranije spomenuti chatbotovi, detekcija spamova, analiza teksta, ispravljanje i provjera pravopisa, stvaranje sažetka teksta, generiranje tekstualnih datoteka itd. Neovisno o primjerni, svaki model u suštini prati korake (Slika 4) predobrade teksta, izvlačenje značajki te gradnju modela [26]:

- **Predobrada teksta**: niz postupka s kojima se nestrukturirani izvorni tekst čisti, normalizira te transformira kako bi bio spreman i kompatibilan za izvlačenje značajki.
- **Izvlačenje značajki**: korak nakon obrade teksta; iz tekstualnih podataka se izvlače značajke koje su primjerene za željeni cilj te dobivaju oblik spreman za pohranjivanje modelu.

- **Modeliranje**: korak zajednički svim ML modelima; gradnja modela, treniranje nad podacima te stvaranje predikcija prema treniranom modelu.

Slika 4: Osnovni postupci NLP procesiranja [26]

3.1 Predobrada tekstualnih podataka

Razina kvalitete izlaznih podataka je jednaka ili proporcionalna razini kvalitete ulaznih podataka [27], ili u znanstvenom svijetu više poznato kao GIGO (*eng. Garbage In Garbage Out* [28]). Izreka je proizašla iz razmišljanja da neovisno o kvaliteti modela ili algoritma kojeg se koristi, ako se pohranjuju loši podaci, nije moguće proizvesti dobre rezultate. Iz navedenoga se može zaključiti da predobrada podataka predstavlja jedan od važnijih koraka u procesu, ne samo NLP-a nego i drugih procesa strojnog učenja.
U NLP-u postoje niz operacija za pretprocesiranje podataka koje se preporučuje izvesti neovisno o svrsi modela.

1. **Čišćenje teksta**: oblik ove operacije ovisi o izvoru podataka. obuhvaća operacije uklanjanja HTML tagova za podatke s online izvora, micanje specifičnih simbola kao što su emotikoni ili interpunkcijski znakovi [29]. Cilj ovog koraka je dobiti čiste podatke sastavljene samo od riječi i znakova koje imaju leksičku važnost. Proces se razlikuje ovisno o jeziku zbog razlika u abecedama i korištenom pismu, npr. kinesko pismo je sastavljeno samo od posebnih simbola.

2. **Tokenizacija**: osnovna strategija za većinu NLP sustava. To je proces razdvajanja rečenica u svoje osnovne dijelove zvani tokeni. Praksa je razdvajati prema razmacima, što rezultira da pojedina riječ predstavlja jedan token [30]. Postoje razni alati s rječnicima tokena za razne jezike. Također razni enkoderi dolaze spremni s vlastitim tokenizacijskim alatima.

3. **Brisanje čestica**: čestice su dijelovi rečenice koje ne nose nikakvu korisnu informaciju. U drugim riječima, ako se eliminiraju čestice iz rečenice, ta rečenica će održati potpuno kontekstualno značenje. Svaki jezik ima svoje čestice, te uklanjanje se izvodi tako da se rečenica provodi kroz listu svih postojećih čestica za potreban jezik [31].

4. **Prepoznavanje entiteta imenica**: riječi ili rečenice koje se kategoriziraju prema određenoj temi. Sadrže ključne informacije unutar rečenice koje služe kao važne oznake za većinu sustava za procesiranje jezika. Pokazalo se da dobro kategoriziranje entiteta poboljšava sustave kao što su QA, strojno prevađanje, automatizirano prikupljanje informacija, itd. Inače pripadaju kategorijama kao što su osobe (<PER>), lokacija (<LOC>), organizacije (<ORG>) te dodatno se označuju ne definirane kategorije [32].

5. **Svođenje na korijen (*eng. stemming*)**: pretvorba morfološkog oblika riječi na svoje korijene. Korijen ne mora postojati u rječniku, ali sve varijacije moraju sadržavati i mapirati na korijen. Važnost ovog koraka ovisi o morfološkoj kompleksnosti jezika, ako je jezik jednostavan onda ovaj korak manje dolazi do izražaja [33].

6. **Lematizacija**: proces svođenja riječi na korijene, ali za razliku od steminga, korijen mora postojati u rječniku. U morfološkim kompleksnim jezicima, većinu riječi su derivat korijena te to predstavlja prepreku za NLP sustave [34].

Razlika između lematizacije i stemanja je suptilna. Kod stemanja se korijen dobije nakon što se primjene niz transformacijskih pravila (npr. uklanjanje sufiksa) ali se ne obazire na vrstu riječi i kontekstualnu važnost. U kontrastu, lematizacija svodi riječ na korijen tek nakon shvaćanja koja je vrsta riječi i kontekstualnu važnost u rečenici. Za razliku od alata za lematizaciju, alati za stemanje su jednostavniji za implementiranje pošto nije potrebno raditi operacije za shvaćanje vrste riječi i konteksta [33].

Text	Stemming	Lemmatization
information	inform	information
informative	inform	informative
computers	comput	computer
feet	feet	foot

Tablica 1: Usporedba lematizacije i stemanja

3.2 Tehnika pozornosti i transformeri

Kada se govori o NLP-u mora se spomenuti tehnika pozornosti (*eng. attention*) i modele temeljenje na transformerima koji su doveli do velikog napretka u polju procesiranja prirodnog jezika.

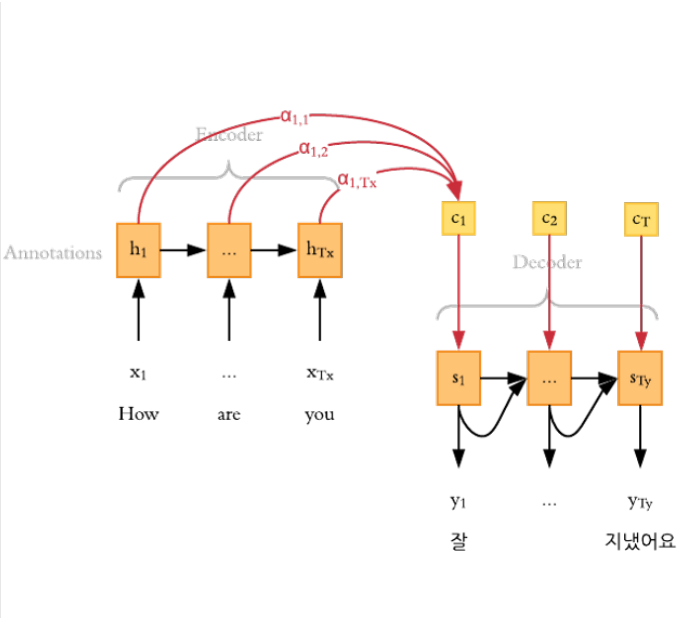
3.2.1 Pozornost

Bahdanau je u svom radu [35] prvog puta uveo tehniku pozornosti u polju procesiranja prirodnog jezika za strojno prevađanje. Slična tehnika je već bila predstavljena u području računalnog vida nakon opservacije da biološka očna mrena se koncentrira na relevantne dijelove optičke slike te pritom gubi rezoluciju na okolnim dijelovima slike. Osim što tehnika pozornosti poboljšava performanse modela, također se može koristiti za interpretaciju ponašanja neuralnih mreža za koje je poznato da dominiraju pojedinost *crne kutije* [36].

Tehnika pozornosti je predloženo rješenje za granice koje posjeduju modeli koji kodiraju ulazne sekvence na fiksiranu vektorsku dužinu iz koje potom dekodiraju izlazne sekvence na svakom koraku. Problem nastaje kada se želi dekodirati duge sekvence, pogotovo sekvence koje su duže od treniranih. Kod pozornosti, kada model pokušava predvidjeti sljedeću riječ, traži pozicije u izvornoj rečenici gdje su koncentrirane najrelevantnije informacije [37].

Temeljna ideja tehnike pozornosti je da svaki put kada model pokuša predvidjeti izlazni podatak, koristi samo dijelove ulazne rečenice gdje su koncentrirane najrelevantnije informacije umjesto cijele rečenice, odnosno pokušava dati veću važnost nekoliko

ulaznim riječima umjesto cijelu rečenicu tretirati kao jednako važnu [37].



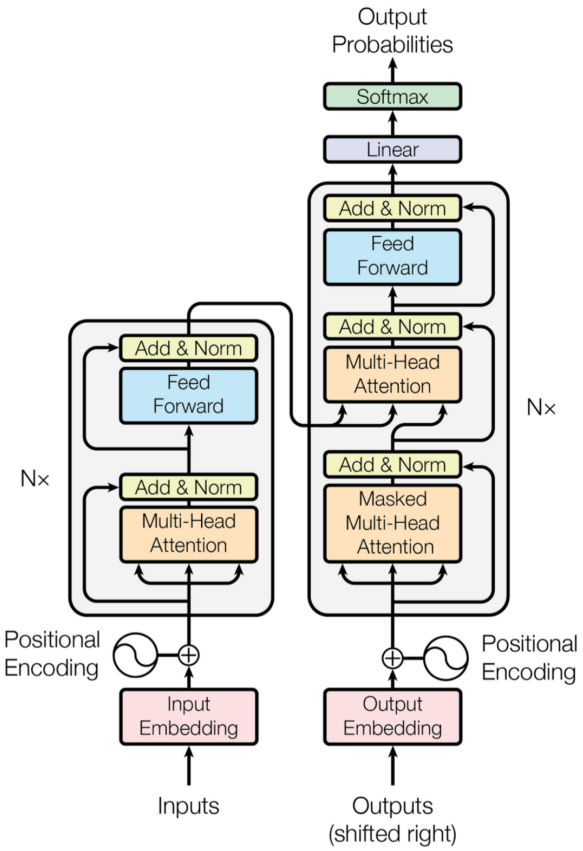
Slika 5: Tehnika pozornosti. Dekoder se računa sa kontekstualnim vektorom, s prethodnim izlazom, prethodnim skrivenim stanjem te posebnim kontekstualnim vektorom za svaku ciljanu riječ. Ti vektori se računaju kao težinski zbroj za aktivacijsko stanje te također predstavljaju koliku pozornost će dobiti za generiranje izlazne riječi [37]

3.2.2 Transformeri

Transformeri su prvog puta uvedeni od tima iz Googlea 2017. godine kao logička nadogradnja za NLP modele u radu pod naslovom ”Attention is all you need” [38]. RNN modeli su bili standard u industriji za procesiranje prirodnog jezika, ali zbog svoje prirode su spori i teški za treniranje te posjeduju problem nestajućeg gradijenta kada se trenira duže tekstove. Taj problem rješava LSTM modeli (Long-Short-Term-Memory): u suštini su jednaki RNN modelima ali posjeduju dodatne sklopke koje olakšavaju pamćenje prijašnjih podataka. Međutim, LSTM modeli su još teži i sporiji za treniranje te nije moguće koristiti ih za prenošeno učenje (*eng. transfer learning*). Kod RNN i LSTM modela, učenje se izvodi sekvencijalno, odnosno riječ po riječ, te zbog tog svojstva nije ih moguće trenirati paralelno što pridonosi svojstvu teškog učenja mreže [39, 40]. Transformeri omogućuju sljedeće prednosti:

- Omogućuju paralelizaciju

- Omogućuju prenošeno učenje
- Olakšava rad s većim ulaznim tekstovima
- U potpunosti se oslanja na tehnike pažnje



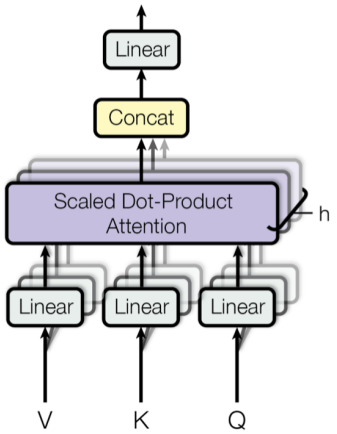
Slika 6: Arhitektura modela s transformersima [38]

Prema izvornom radu: ”Transformeri su prvi seq2seq modeli koji se u potpunosti oslanjaju na tehnici samo-pažnje (*eng. self-attention*) da bi izračunali reprezentaciju ulaznog i izlaznog podataka bez da koristi usklađene sljedove kao RNN ili konvolucije.” [38] Da bi model bio uspješan u navedenom, autori rada uvode tri značajna dijela (Slika 6):

1. **Pozicijski enkoder**: ovo svojstvo omogućava da modelu nisu potrebne povratne veze (ponavljajući faktor) kao u RNN modelima. Model koristi pozicijski enkoder kako bi slojevima enkodera dao informacije o relativnim ili apsolutnim pozicijama riječi unutar rečenice [38].
2. **Višestruka pažnja**: umjesto računanja razine pažnje za svaku riječ sekvencijalno, model koristi tzv. višestruku pažnju (*eng. Multi-headed attention*). Pažnja

se računa paralelno za svaku riječ te se na kraju spajaju u jedinstvenu vrijednost [38]. To omogućava paralelni rad modela (Slika 7). Transformeri koriste posebnu tehniku pažnje (*eng. self-attention*) koja omogućava povezivanje riječi na različitim pozicijama od jedne rečenice kako bi izračunao reprezentaciju sekvencije.

3. **Izlaz enkodera se pohranjuje svakom sloju dekodera**: prema slici 6 se može primijetiti da enkoderski dio je sekvencijalno spojen, naslagani slojevi jedni iznad drugih, međutim izlazni podatak enkodera se pohranjuje svakom sloju dekodera [39].



Slika 7: Reprezentacija višestruke pažnje [38]

3.3 BERT

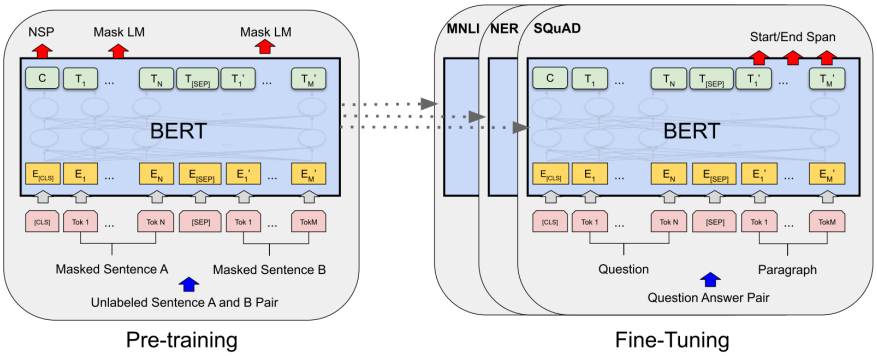
U svim znanstvenim radovima ili člancima s tematikom NLP-a, uvijek se spominje BERT model. BERT je spojio prednosti tehnike pažnje i trensformera rezultirajući u model koji je nadmašio do tada razvijene jednosmjerne modele. BERT, akronim s engleskog naziva *Bidirectional Encoder Representations from Transformers*, uvodi inovativnu tehniku dvosmjernosti, odnosno ne čita dokument s lijeve na desnu, ili s desne na lijevu, nego gleda u oba smjera relativno prema trenutnom tokenu. Drugim riječima, gleda tekstualni ulaz kao cjelinu [41, 42]. Razvijen od istraživačkog tima s Googlea, ima cilj poboljšati metode dotjerivanja dotadašnjih modela, tvrdeći da jednosmjerni modeli limitiraju prednosti pretreniranja, pogotovo za pristup dotjerivanja. Jednosmjerni modeli limitiraju odabir arhitekture koje se mogu koristiti za proces pretreniranja [42]. Kako bi dostigli tvrdnju, autori koriste dvije tehnike tijekom treniranja:

1. *Maskirani jezični modeli*: u jednosmjernim modelima se inače koristi tehnika predviđanja sljedeće riječi, ali zbog razloga navedenih ranije, to limitira model na odabir arhitekture. BERT koristi tehniku maskiranja, odnosno sakrije 15% riječi i dodijeli cilj mreži da predvidi skrivene riječi ovisno o kontekstu dobivenoga od ne maskiranih riječi [41, 42].

2. *Predviđanje sljedeće rečenice*: ova tehnika služi kako bi model naučio analizirati odnose među rečenicama. Tijekom treniranja model dobije kao ulaz par dviju rečenica te mora predvidjeti ako druga rečenica slijedi prvu rečenicu. Kod odabira rečenice za svaki korak, 50% slučaja druga rečenica je istinita rečenica koja slijedi prvu, dok ostalih 50% je nasumična rečenica izvučena iz korpusa [42].

Da bi BERT bio primjenjiv za razne NLP probleme, reprezentacija ulaza je modificirana kako bi lakoćom mogao nedvosmisleno prikazati jednu rečenicu ili slijed parova rečenica. Za to koristi WordPiece [43] kodiranje koja sadrži 30 000 tokena: prvi token svakog slijeda je specijalni token [CLS], na kraju svake rečenice se postavlja [SEP] token koji također služi za raspoznati početak sljedeće rečenice te za maskirane riječi se postavi token [MASK] [42].

Autori BERT-a su zamislili korištenje modela u dvije faze: faza predtreniranja te faza dotjerivanja specifična domeni željenog problema (Slika 8). Faza predtreniranja je objaš-



Slika 8: BERT konceptualna arhitektura [42]

njena ranije u ovom odlomku te odrađena od strane autora rada. Faza dotjerivanja je poprilično jednostavna pošto tehnika pažnje u transformerima omogućava BERT-u prilagođavanje velikim količinama zadataka, neovisno ako zahtijevaju rečenicu ili parove rečenica. Za svaki zadatak dovoljno je prilagoditi ulazne i izlazne podatke te dotjerati parametre cijelog modela. Na izlazu se postavi sloj ovisno o problemu, npr. klasifikacijski sloj za analizu osjećaja. Uspoređujući fazu dotjerivanja s fazom predtreniranja, završna faza je relativno računalski razumna: jedan sat na TPU ili par sati na GPU [42].

4 Prepoznavanje namjere

Chatbot usmjeren rješavanju zadataka omogućava interakciju korisnika i računala kroz prirodan razgovor s ciljem rješavanja specifičnog zadataka kao što je rezervacija hotela, kupnja avionske karte, rezervacija stola u restoranu, automatska podrška korisnicima, itd. [44].

Pojedinost koju karakterizira chatbotove je konverzacijski tijek razgovora. Konverzacija se razlikuje od drugih vrsta razgovora na puno načina [45]. Sudionici konverzacije su fizički prisutni te ne-verbalna i paralingvistička ponašanja imaju veliku ulogu u tijeku razgovora i razumijevanja teme. Također, prisutnost sudionika lice-u-lice daje dodatne predrasude koje utječu na slobodu misli i samim time zaključcima [46]. Za razliku od pisanoga i promišljenog teksta, konverzacijske izjave su tipično loše iskazane, sadržavaju lažne početke rečenica, oklijevanja, nepredvidljivost, ironiju, itd. [45], te ako se želi takve oblike razgovora integrirati u automatizirane sustave kao što su chatbotovi, sve zajedno predstavlja još kompleksniji problem.
Ironija predstavlja specifičan problem za prepoznavanje: oblik izjave je precizan i pravilan ali namjera je drugačija od doslovne. Razina kojom sudionik razgovora može odrediti ako je izjava ironična ili ne ovisi o brojnim faktorima koje čovjek nauči prema iskustvu te osobnim kognitivnim i socijalnim sposobnostima [47]. Kognitivna sposobnost koja je rijetka i teška za naučiti kod ljudi, za računalo predstavlja dodatni problem. Usporedivši računalo s ljudima, računalo ima kognitivnu sposobnost djeteta. Kod djece je pragmatika važno područje razvoja, odnosno sposobnost prepoznavanja konteksta i znanja iz specifičnih konverzacijskih uzorka kako bi pravilno interpretirao jezik i razgovor. Jedno područje pragmatike je također prepoznavanje ne-doslovne izraze kao što je ironija [47]. Konverzacijski ironijski izrazi predstavljaju specifičan problem kod prepoznavanja namjera zato što gledajući izraz doslovce, predviđena namjera će biti točna, međutim kontekstualna namjera će biti kriva, što se s konverzacijskim sustavima cilja postići. Zbog navedenih problema, potrebno je kvalitetno procesiranje ulaznog teksta tijekom razgovora [45].

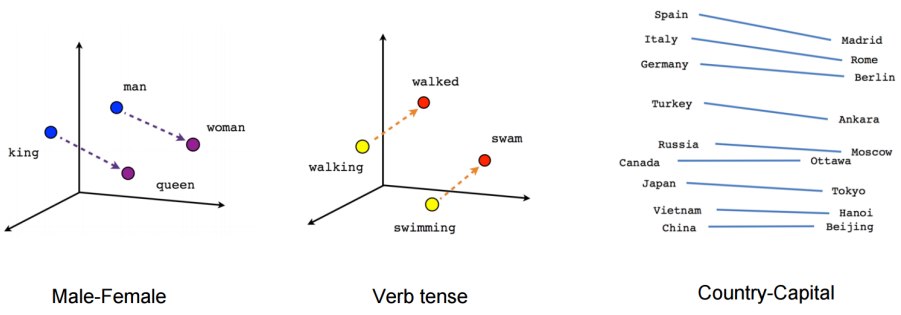
Govoreći o sustavima usmjerenim rješavanju specifičnih zadataka (za razliku od općih sustava koje nemaju specifični zadatak nego su dizajnirani samo za vođenje raz-

govora), prepoznavanje namjere je vitalna komponenta za pravilno prepoznati željeni zadatak [44].

Da bi sustav shvatio korisnikov trenutni cilj, mora manipulirati s detektorom namjera kako bi klasificirao korisnikovu izjavu (dobivenu u nekom obliku prirodnog jezika) u predodređene klase, odnosno namjere. Na primjer, u domeni hotelijerstva, korisnik može zahtijevati rezervaciju soba ili otkazivanje soba, iako oba zahtjeva imaju veze s istom domenom, imaju suprotne namjere. Također, važnost točno klasificirane namjere prikazuje činjenica da krivo klasificirana namjera je prvi faktor za neodrživi razgovor [44].

4.1 Kodiranje riječi vs kodiranje rečenica

Kodiranje umetanjem (*eng. embedding*) je niskodimenzionalno prikazivanje točke u višedimenzionalnom vektorskom prostoru. Stoga, kodiranje riječi ili cijelih rečenica je prikaz gustih višedimenzionalnih vektora u nižedimenzionalne vektorske prostore. Prvi riječni enkoder koji koristi neuralne mreže je objavljen 2013. godine od Googleovih znanstvenika [48], te od tada su prisutni u skoro svim NLP modelima zbog njihove učinkovitosti [49]. Prednost kodiranja je što riječi ili rečenice sa sličnim značenjem će se nalaziti na približno ili istim pozicijama u vektorskom prostoru (Slika 9), što s drugim reprezentacijama, kao npr. one-hot-encoding ili bag-of-words reprezentacijama, nije moguće. Vektorski prikaz riječi ili rečenica može obuhvatiti, osim kontekstualnog značenja, također semantičko značenje teksta [50].



Slika 9: Reprezentacija riječi u vektorskom prostoru [51]

Za svrhu klasificiranja namjere potrebno je kontekstualno kodiranje jer kodiranje riječi

zanemaruje cjelovito značenje rečenice, već gleda samo trenutnu riječ te kodira značenje pojedinačne riječi. Kako bi se dobilo značenje cijele rečenice, a ne samo pojedinačnih riječi unutar rečenice, potrebno je kodiranje cijelih rečenica (*eng. sentence embedding*). Kodiranje rečenica je slično kodiranju riječi: oba imaju svrhu prikazati višedimenzionalne vektore u nižedimenzionalne vektorske prostore. Algoritmi su također slični; u većini slučajeva su algoritmi za kodiranje riječi prilagođeni za kodiranje rečenica, ali da pritom zadrže isti cilj: slične rečenice imaju slično vektorsko kodiranje [52].

Kao u većini slučajeva kod drugih ML/DL problema, treniranje i odabir modela se često svodi na biranje nekog od već istreniranih velikih modela, kao ranije spomenuti BERT. Masivni modeli imaju veliku doprinos u raznolikim NLP problemima i aplikacijama zahvaljujući njihovom svojstvu treniranja na velikim količinama generalnih tekstualnih korpusa [53]. Adaptacija modela za specifične domene se inače izvodi na sljedeći način: na masivni model, koji je prijašnje treniran na velikim količinama generalnih podataka, nadoda se izlazni sloj koji je specifičan za odabrani zadatak te se izvodi proces dotjerivanja (*eng. fine-tuning*) cijelog modela. Međutim, kao što se može očekivati, to zahtijeva veliku količinu procesorske snage [44].

U nastavku su nabrojeni poznatiji te opširnije korišteni rečenični enkoderi:

Doc2Vec: *Document to Vector* [54], također poznat kao *Paragraph Vector*, predložen 2014 godine od Googlea te je temeljen prema algoritmu Word2Vec. Algoritam prati pretpostavku da okruženje jedne riječi ujedno definira značenje [52].

LASER: *Language Agnostic Sentence Representation* [55], razvijan od istraživačkog tima s Facebooka; treniran na 233 milijuna rečenica na 93 jezika s enkoder-dekoder arhitekturom. Enkoder je sastavljen od višeslojnog dvosmjernog LSTM-a te dekoder je jednoslojni jednosmjerni LSTM. Stvara kodirani izlaz od 1024 dimenzije nad zadnjem slojem enkodera koji služi kao ulazni podatak dekoderu na svakom koraku [56].

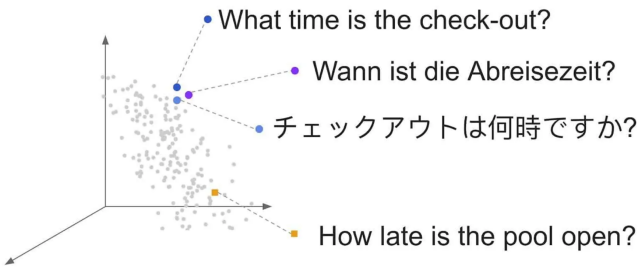
mUSE: *Multilingual Universal Sentence Encoder* [57], treniran na 16 jezika, oslanja se na arhitekturu dualnog enkodera (ulaz su dvije rečenice). Proširena verzija univerzalnog enkodera na više jezika. Treniran istovremeno na više jezika te ma-

pira rečenice od različitih jezika na jednake koordinate u vektorskom prostoru.

SBERT: *Sentence-BERT* [58], za izračunati rečenične srodnosti za 10.000 rečenica sa izvornim BERT-om bilo bi potrebno 65 sati računanja na modernom GPU. Autori SBERT-a predlažu kao rješenje sijamsku arhitekturu koja omogućava derivacije rečenica fiksnih dužina. Autori koriste metriku srodnosti *"kosinska srodnost"* kako bi pronašli semantičke srodne rečenice. Takva arhitektura smanjuje proces za računanje 10.000 rečenica na 5 sekundi [58].

LaBSE: *Language Agnostic BERT Sentence Embedding* [59], jezično neovisan rečenični enkoder za 109 jezika. Model kombinira jezično maskirane modele s modelima za prevođenje. Model je treniran s metodom rangiranja prijevoda koristeći dvosmjerne dualne enkodere [59].

Dodatni problem kojem treba obratiti pozornost je potreban jezik za određeni zadatak. Većina modela su trenirana i namijenjena za korpuse na engleskom jeziku. Rješenje tome je treniranje na više jezika: mBERT (*Multilingual BERT*) je treniran na 104 jezika [42]. Modeli trenirani na 100+ jezika kao što je mBERT postali su standard za procesiranje i prikaz više jezičnih NLP problema [56]. Međutim, iako je model prilagođen za više jezika, također dobiva i sve ranije navedene mane što donose veliki modeli trenirani na masivnim korporama. Neovisno o modelu, rješenja se oslanjaju na istom konceptu: smatra se da neovisno o korištenom jeziku, rečenice bi trebale imati isto semantičko značenje, odnosno u vektorskom prostoru njihove kodirane pozicije bi trebale biti približno iste (Slika 10)



Slika 10: Reprezentacija višejezičnih rečenica u vektorskom prostoru: neovisno o jeziku, kodirano značenje rečenice se mora nalaziti na sličnim pozicijama [60]

Za svrhu ovog rada, potrebno je da model zadovolji sljedeće stavke:

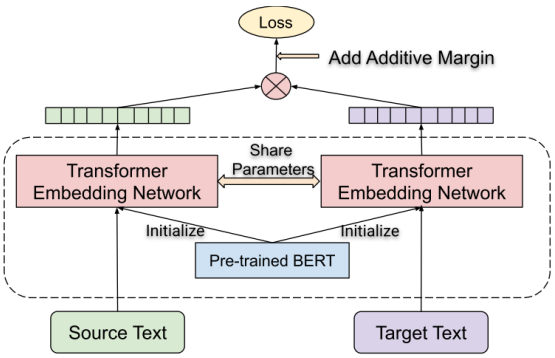
1. Mora prihvaćati hrvatski jezik: da se među jezicima na kojima je model treniran također nalazi hrvatski jezik
2. Mora biti enkoder rečenica bez potrebe dotjerivanja cijelog modela: moglo se vidjeti da postoje puno modela za kodiranje rečenica, ali potreba za dotjerivanjem masivnih modela predstavlja problem zbog dostupnih resursa.

Model koji zadovoljava navedene kriterije je LaBSE te je detaljno opisan u sljedećem poglavlju. Također, razlog kriterijima i konačnom odabiru LaBSE modela će biti obrazloženi u praktičnom djelu rada.

4.2 LaBSE

Kao što se može zaključiti iz imena modela, LaBSE predstavlja prilagođenu verziju BERT-a kako bi bila kompatibilna za više jezične namjene, preciznije za 109 jezika. LaBSE kombinira maskirane jezične modele s modelima za prevođenje koje se treniralo s dvosmjernim dualnim enkoderom [61].

Predtreniranje modela s maskiranim jezicima te dotjerivanje istih za specifične zadatke pokazalo se kao snažan alat za NLP probleme [42], ali također se pokazalo da taj pristup ne donosi dobre rezultate za kodiranje rečenica. Ranije navedeni SBERT, dotjerani monojezični BERT, postiže odlične performanse na mjerenjima za kvalitetu kodiranja rečenica, ali funkcionira samo na engleskom jeziku [59]. Većina višejezičnih modela ne koristi maskirane jezike, već se oslanjaju na direktno prevedene parove što rezultira za potrebom velike količine paralelnih podataka za treniranje.



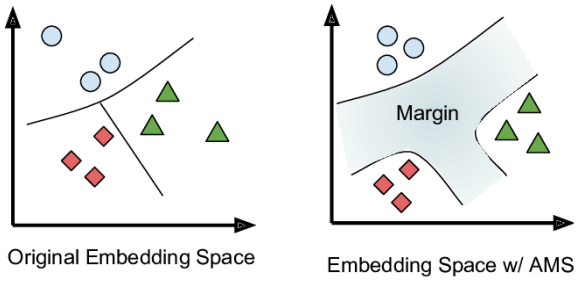
Slika 11: LaBSE arhitektura [59]

LaBSE uvodi maskirane jezične modele te koristi arhitekturu dualnih enkodera. Prema slici 11 može se vidjeti da su parovi rečenica kodirani odvojeno, s enkoderom temeljenim na BERT-u. Srodnost rečenica se izračuna sa kosinom kodiranih rečenica [59]. Model je treniran sa softmax aditivnom marginom (AM-Softmax). Tijekom klasifikacije se stvaraju granice odluke za odvojiti pojedine klase. Međutim, stvara se problem kada izlazna klasifikacijska vrijednost se nalazi blizu granica. AM-Softmax rješava taj problem na način da doda margine nad klasifikacijskim granicama (Slika 12) kako bi naglasio odvojivosti klasa [62]:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{\infty}\frac{e^{\phi(x_i,y_i)-m}}{e^{\phi(x_i,y_i)-m}+\sum_{n=1,n\neq i}^{N}e^{\phi(x_n,y_n)-m}}$$

Kodirani vektorski prostor se dobiva iz funkcije $\phi(x, y)$ koju prateći izvorni rad od [63] se postavi na $cos(x, y)$. Funkcija pokuša procijeniti rečenicu y_i, odnosno prijevod od rečenice x_i među $N-1$ alternativa. Procjena se računa također kada $\phi(x_i, y_i)$ se smanji za marginu m. Rezultat funkcije je asimetričan pošto je različit ovisno ako se računa za izvornu rečenicu ili ciljnu rečenicu od parova ulaza. Kako bi se dobio konačni gubitak, potrebno je zbrojiti gubitak dobiven iz izvora prema ciljnoj rečenici te gubitak od ciljne rečenice prema izvornoj [59]:

$$\overline{\mathcal{L}} = \mathcal{L} + \mathcal{L}'$$



Slika 12: Rezultat apliciranja funkcije softmax aditivne margine [63]

Model je testiran u svrhu evaluacije performansi nad *Tatoeba* korpusu koji je sastavljen od 1000 rečenica na engleskom jeziku usklađenim s prijevodima 112 jezika. Cilj je pronaći najbližu rečenicu koristeći kosinsku srodnost [64].

Model	**14Jezika**	**36Jezika**	**82Jezika**	**112Jezika**
mUSE	93.9	-	-	-
LASER	95.3	84.4	75.9	65.5
LaBSE	95.3	95.0	87.4	83.7

Tablica 2: Usporedba performansi modela nad Tatoeba [64] korpusu [59]

Kako se može primijetiti iz Tablice 2, model postiže jednake performanse s LASER modelom na razini gdje oba modela imaju velike količine podataka (LASER treniran nad 82 jezika, dok LaBSE sa 109 jezika). Model značajno postiže bolji rezultat nad cijelom

setu podataka, vjerojatno zbog treniranja s većim brojem jezika.

Za ovaj rad se odabralo LaBSE model za kodiranje rečenica zato što prihvaća oba potrebna jezika, engleski i hrvatski (sljedeće poglavlje objašnjeno u detalje) te kao što se moglo vidjeti (Tablica 2) model nadmašuje druge višejezične modele u problematici sintaktičke srodnosti. Dodatan razlog njegovom odabiru je jednostavnost implementacije: dostupan je na TensorFlow Hub-u (https://tfhub.dev/google/LaBSE/2) te dovoljno je nadodati klasifikacijski sloj za potrebne namjene.

5 Praktični problem

5.1 Opis problema

Transakcija s učeničkog sustava na studentski sustav, te također tijekom cijelog studentskoga razdoblja, često donosi veliku količinu pitanja i dvojba za ispunjavanje i prikupljanje velike količine dokumentacije ili za jednostavno odgovoriti na informativna pitanja. Sveučilišta najčešće postave referadu za rješavanje kompliciranijih problema te za jednostavnije upite služe web mjesta i e-mail pretinci. Problem je što takvi sustavi rade na individualnoj razini: maksimalan broj upita koja referada može rješavati je direktno povezana s brojem zaposlenih ljudi te mailovi su u većini slučajeva upućeni individualnim studentima [65]. Posljedica tome je vrijeme čekanja na odgovore i rješenja se proporcionalno povećaju s brojem upita.

Cilj projekta, od kojeg se u ovom radu razrađuje NLP prepoznavanje namjere te služi kao dokaz funkcioniranja koncepta cijelog sustava s BPMN modelima, je implementirati konverzacijski sustav kako bi se automatizirali ciklični i repetitivni upiti. Drugim riječima, implementirati pametan chatbot koji bi preko sučelja i pomoću NLP tehnologije mogao razumjeti namjere korisnika (studenta) iz konverzacije vođene prirodnim jezikom te pravilno formirati odgovore i pokrenuti potrebne procese.

Konverzacijsko sučelje nije nova metodologija ali trenutni chatbotovi i virtualni asistenti su sve više popularni za pristup podacima i uslugama na prirodniji ljudski način. Vračanje na tekstualno sučelje je potaknuto od strane većeg dostupnog broja platforma za komunikaciju sa porukama (chatova) te jednostavnost implementiranja istih u postojeće sustave: npr. Mastercard ima uslugu za jednostavnu integraciju plaćanja tijekom razgovora unutar Facebook chata [65]. U radu objavljenom 2017. godine, autori su dokazali da implementacija virtualnog asistenta u obliku chatbota za rješavanje vremensko zahtjevne procese, kao što su upisi na sveučilište, drastično skraćuju vrijeme obavljanja procesa [66].

Za svrhu ovog diplomskog rada i za dokazivanje učinkovitosti algoritma odabralo se dva poslovna procesa, odnosno namjere, specifična za sveučilište: prijava teme završnog

ili diplomskog rada i upisi koji obuhvaćaju prve upise na sveučilište ili upise na više godine studija:

- **Prijava teme završnog ili diplomskog rada**: proces prijave teme je dinamičan proces koji zahtijeva sudjelovanje više aktera. Student mora istražiti dostupne teme, provjeriti slobodne mentore ili prijaviti vlastitu temu; dok mentori (profesori) moraju provjeriti predloženu temu te prihvaćati ili odbiti studente za mentoriranje. Cijeli proces za profesore je jako redundantan pošto se ponavlja isti koraci za svakoga studenta, te takav proces predstavlja dobrog kandidata za zamijeniti s chatbot asistentom.

- **Upisi**: upisi predstavljaju stresno i kaotično razdoblje za studente koje dovodi mnoštvo pitanja popraćen s velikom količinom potrebne dokumentacije. Iako puno odgovora se može pronaći na službenim web stranicama sveučilišta, sve-jedno se referade popune s pitanjima. Kao i u prvom slučaju, na ponavljajuća pitanja može odgovarati virtualni asistent te kada je potrebno također pokrenuti proces samog upisa ili uključiti u razgovor potrebnog zaposlenika.

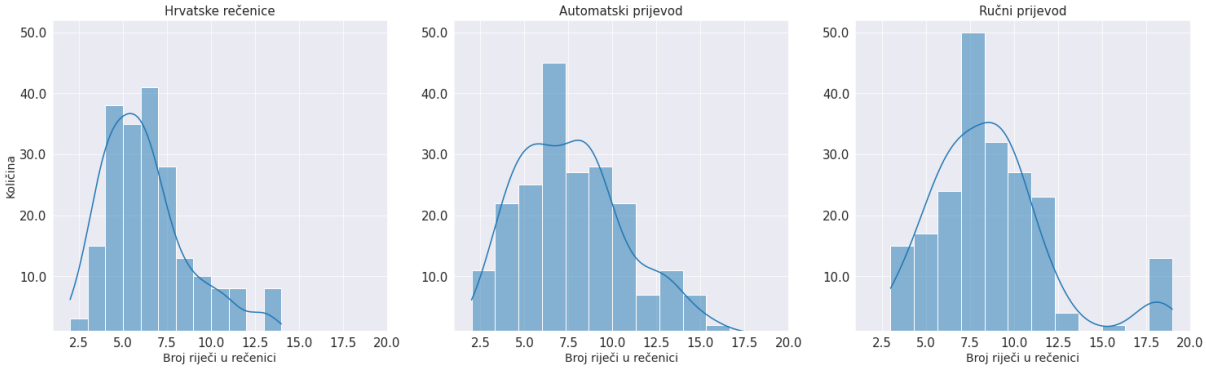
Cilj je sagraditi model koji s visokom točnosti klasificira željenu namjeru. U sljedećem poglavlju se opisuje način prikupljanja i analiza podataka s domenom gore navedenih namjera.

5.2 Prikupljanje i analiza podataka

Količina i kvaliteta podataka imaju veliku ulogu na krajnje rezultate modela. Održavajući pravilo *"Garbage in - garbage out"*, korak prikupljanja podataka je jedan od važnijih, ako ne i najvažniji, za uspješno treniranje modela. Za ovaj projekt se želi prikupiti podatke u obliku rečenica-oznaka, odnosno za svaku rečenicu ili pitanje koja predstavlja dio konverzacije, pridružiti odgovarajuću oznaku namjere. Za te svrhe razvojni programeri i eksperti domene smatraju korisnim povijesne konverzacijske zapise između čovjeka-čovjeka ili čovjeka-računala te naknadno izvode ručno dodjeljivanje oznaka. Dodjeljivanje oznaka je vremensko zahtjevni proces i u većini slučajeva odrađen od strane eksperata ciljane domene [67]. Postoje automatizirani alati za označavanje zapisa, međutim nisu prikladni za domenu ovog projekta. Specifičnost ovog projekta je hrvatski jezik te nema dostupnih povijesnih zapisa konverzacije između studenata i studentske službe ili studenta i profesora s ciljanom domenom ovog projekta. Iz tog razloga stvaranje skupa podataka se izvodilo samostalno ručnim upisom i parafraziranjem željene rečenice. Na taj način se osiguralo preciznost označavanja kategorija namjera prema ulaznoj rečenici. Primarni jezik sveučilišta je hrvatski jezik, međutim sveučilište prihvaća također studente iz inozemstva (ERASMUS studenti) za koje se pretpostavlja da razumiju i da mogu komunicirati na engleskom jeziku. Iz tog razloga, model mora prihvaćati ulaze na primarnom hrvatskom jeziku i također engleskom jeziku.

Baza podataka je sastavljena od 208 rečenica koja su naknadno prevedena na engleski jezik. Prevođenje se izvelo na dva načina: 1) automatsko prevođenje s Google Prevoditeljem te 2) ručno prevađanje. Prevođenje se izvelo na dva načina zbog kvalitete prijevoda. Google Prevoditelj, iako dobar prevoditelj, u nekim slučajevima nije u mogućnosti pravilno prevesti; s ručnim prevođenjem gledalo se ispraviti greške napravljene od strane automatskog prevoditelja te je u nekim slučajevima potpuno razmijenilo prijevod.

Prijevod je također imao utjecaj na dužinu ulaznih rečenica: dok najduža rečenica na hrvatskom jeziku ima 14 riječi te najkraća 2 riječi, automatski prijevod je povećao dužinu rečenice na 18 riječi a ručni prijevod dodatno na 19 riječi; najkraće rečenice na prijevodima je nepromijenjena na automatskom prijevodu ali za ručni prijevod se po-



Slika 13: Promjena dužine rečenice prijei nakon prevađanja na engleski jezik

većala na 3 riječi. Prevođenje s hrvatskog jezika na engleski jezik automatskim putem pridonosi povećanjem prosječne dužine rečenice za 24%, odnosno za 40% za ručno prevađanje. Važnost tog podataka stoji u činjenici da odabrani model LaBSE je baziran na BERTu koji ima ograničenja na ulazu od 512 tokena (uključujući i specijalne tokene [CLS] i [SEP]): nakon ograničenja, potrebno je rezati ulaz što može utjecati na rezultate [68]. Iako ovaj set podataka nema rečenice duže od 19 riječi ali treba voditi računa da kod prijevoda se dužina rečenica povećava te treba pripaziti na limite predtreniranog modela.

Jezik	**Tekst**
Hrvatski	kako prijaviti zavrsni
Google prijevod	How to Report Final
Ručni prijevod	how to register my final paper
Hrvatski	Rečeno mi je da se javim za provedbu upisa
Google prijevod	I was told to report to enforcement
Ručni prijevod	I was told to come here for the enrollment
Hrvatski	treba mi potvrda da sam student
Google prijevod	I need a confirmation that I am a student
Ručni prijevod	i need a verification document that i am a student

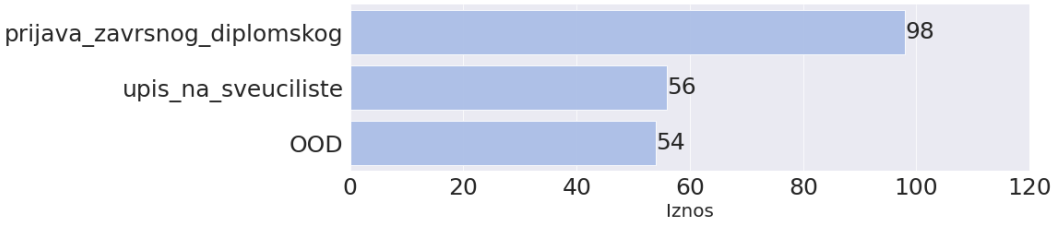
Tablica 3: Primjer zapisa u bazi podataka

U tablici 3 se mogu vidjeti primjeri zapisa u bazi podataka. Može se primijetiti da automatsko prevođenje dovodi dosta grešaka koje se isprave s ručnim prevođenjem. Također se može primijetiti da se ne prati gramatički pravopis. Razlog tome je što ko-

nverzacija s chatbotom se razlikuje od konverzacije uživo ili putem e-maila: početak rečenice ne počinje velikim slovom, pitanja ne završavaju s pravilnim interpunkcijama, imena u većini slučajevima ne počinju s velikim slovom, koristi se žargon, ne pazi se na specijalne znakove (npr. 'š' je zamijenjen sa slovom 's'), rečenice su kratke itd.

Rečenice su označene u dvije grupe namjera: Grupa_1 koja sadrži 3 klase namjera te Grupa_2 koja sadrži 5 klasa namjera.

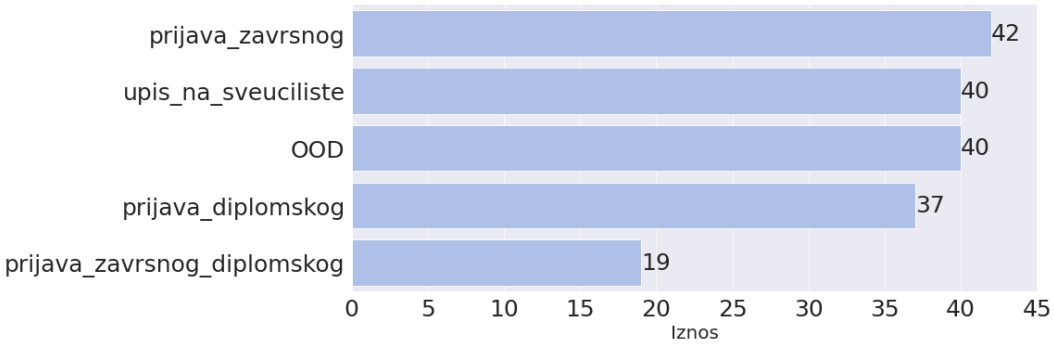
Grupa_1: prva grupa sadrži 3 klase (Slika 14) namjere koje odgovaraju odabranim procesima u prijašnjem poglavlju: *prijava zavrsnog diplomskog*, *upis na sveuciliste* te kontrolna klasa *'OOD'* (*eng. Out Of Domain*) koja služi za klasificiranje svih ulaza koji ne pripadaju niti jednoj namjeri, odnosno da model prepozna kada nema odgovor na određeni ulaz.



Slika 14: Distribucija namjera Grupa_1

Grupa_2: druga grupa (Slika 15) je proširena verzija druge grupe. Namjera *prijava zavrsnog diplomskog* je podijeljena u 3 potklase s većim kontekstualnim značenjem: *prijava zavrsnog*, *prijava diplomskog* te namjera za općenite teme u vezi prijava *prijava zavrsnog diplomskog*. Dodatne namjere su kao i u prvoj grupi, *upis na sveuciliste* i *OOD*. Razlog razdvajanja namjere za prijavu na 3 dijela je povećanje ravnoteže podataka; iz prvog grafa se može primijetiti da namjera *prijava zavrsnog diplomskog* je znatno brojnija od ostalih namjera.

U Tablici 4 se mogu vidjeti primjeri dodjeljivanja namjera ulaznim rečenicama. Iz rečenice 1. se može primijetiti kako namjera iz Grupe_1 daje manje informacija od namjere iz Grupe_2, na taj način informacija dobivena iz predviđene namjere iz Grupe_2 je preciznija i uža što olakšava obradu pravilnog odgovora i odabiru sljedećeg koraka chatbota. Međutim, u slučajevima kao što je rečenica 4., nije moguće preciznije odrediti



Slika 15: Distribucija namjera Grupa_2

namjeru u skupini namjera Grupe_2 pa se dodjeljuje općenita namjera prijave. Rečenica 3. je označena kao OOD, čime se primjećuje da OOD mogu biti zahtjevi koji su svejedno vezani s procesima sveučilišta ali nisu podržani od modela: cilj razvoja sustava je da s vremenom namjera OOD ima što manje zahtjeva vezana za sveučilište. U budućem razvoju moguće je preciznije odrediti namjere za dobiti bolji model: npr. smanjenjem količine namjera vezane s procesima sveučilišta iz OOD klase, preciznije definiranje namjera upisa (rečenica 4. može se definirati kao namjera prijave teme, neovisno ako je završni ili diplomski u pitanju, pritom da se ne izlazi iz domene opće prijave).

n.	Tekstualni ulaz	Grupa_1 namjera	Grupa_2 namjera
1.	kako prijaviti zavrsni	prijava zavrsnog diplomskog	prijava zavrsnog
2.	Rečeno mi je da se javim za provedbu upisa	upis na sveuciliste	upis na sveuciliste
3.	treba mi potvrda da sam student	OOD	OOD
4.	mogu prijaviti temu razvoja web aplikacije	prijava zavrsnog diplomskog	prijava zavrsnog diplomskog

Tablica 4: Primjer klasifikacije rečenica prema pripadajućim namjerama Grupe_1 i Grupe_2

5.3 Modeliranje i treniranje

Tehnologije i paketi korišteni za modeliranje i pisanje koda su Tensorflow, Keras, pandas, nktl, seaborn, text-hr. Cijeli projekt je pisan u jeziku Python te koristilo se Google-ovo programsko okruženje CoLab:

- *Tensorflow*: razvijena od Google tima te objavljen 2015. godine, Tensorflow je platforma otvorenog izvora za razvoj strojnog učenja od koncepta do produkcije. Ima jednostavan i fleksibilni ekosustav alata, paketa te široku zajednicu s dostupnim resursima. Razvijen za razvoj znanstvena istraživanja i za jednostavnu implementaciju razvojnim programerima svojim aplikacijama [69].

- *Keras*: također API za razvoj strojnog učenja te od verzije 2.0 je dio ekosustava tensorflowa. Prema riječima službenih stranica Keras je API dizajniran za ljudska bića a ne strojeve. Nudi dosljedan i jednostavan API, minimizira potrebnu interakciju korisnika za česte slučajeva razvoja te nudi jasne i djelotvorne poruke za greške [70].

- *Pandas*: Python paket za jednostavnu manipulaciju podataka temeljen na podatkovnim okvirima. Nudi alate za čitanje i pisanje podataka raznih formata, integrirano rukovanje podacima koji nedostaju, jednostavna manipulacija oblika podataka, automatsko indeksiranje te sve optimizirano za vrhunske performanse [71].

- *Seaborn*: Python paket za vizualizaciju podataka baziranim na matplotlib paketu. Nudi jednostavno sučelje za crtanje atraktivnih i informativnih statističkih grafova [72].

- *Natural Language Toolkit*: NLTK je platforma za stvaranje Python programa za rad s podacima ljudskog jezika. Pruža jednostavno sučelje za preko 50 korpora i leksičkih resursa poput WordNet. Također pruža alate za procesiranje teksta za klasifikaciju, tokenizaciju, lematizaciju, semantičko zaključivanje i dr. [73].

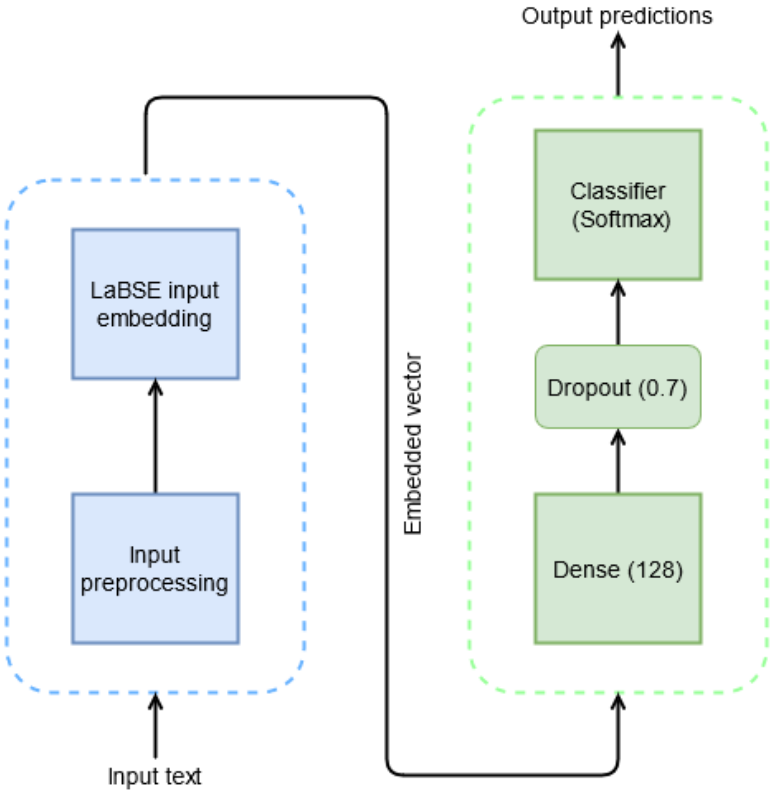
- *text-hr*: iako još uvijek u alfa fazi, paket za manipulaciju teksta na hrvatskom jeziku. Pruža resurse kao listu čestica te rješenja za označavanje dijelove govora [74].

Za treniranje modela iz početnog seta podataka odvojilo se podatke u set za treniranje i set za validaciju s odnosnom 80:20. Primarni problem predstavlja jezik. Za engleski jezik dostupno je veliki broj resursa s kojima bez problema se može predobraditi tekstualne podatke te mnoštvo opcija za odabir modela. Problem nastaje kod hrvatskog jezika; resursi za predobradu su minimalni ili ne postojeći te mali broj modela prihvaća hrvatski jezik. Problem odabira modela se lako rješava zahvaljujući višejezičnim rečeničnim enkoderima, preciznije modela LaBSE opisanog u poglavlju 4.2: treniran je sa 109 jezika te među njima, osim engleskog, je također hrvatski jezik. Istraživanjem NLP dostupnih modela, prihvatljiv odabir bi bio odabrati BERT model (ili mBERT za višejezičnost), međutim BERT model je dobar za kodiranje rečenica samo u slučaju ako se dotjera cijeli model, ali dostupne hardverske resurse i raspoloživo vrijeme ne omogućavaju taj odabir. Iz tog razloga odabralo se LaBSE koji dio predtreniranja za rečenično kodiranje je već odrađeno. Nakon što se odabrao bazni model, potrebno je samo nadodati klasifikacijske slojeve koji su specifični za problem klasifikacije namjere. Slijedno tome, pratilo se rad [44] gdje klasifikacijski sloj je jedan skriveni sloj s 512 perceptrona te postavljenim značajnom regulacijom od 70% otpada (*eng. Dropout*. Analogno tome klasifikacijski sloj je sastavljen od jednoga skrivenoga sloja s 128 perceptrona s 'relu' aktivacijskom funkcijom te također 70% otpada prije izlaznog sloja. Izlazni sloj, odnosno klasifikator, je jednostavan gusti sloj (*eng. Dense*) s brojem perceptrona jednakim broju namjera za predviđanje te sa 'softmax' aktivacijskom funkcijom. Arhitektura cijelog modela se može vidjeti na slici 16. Takav model rezultira s 471 milijuna parametra, od kojih 470,926,849 su predtrenirani parametri od LaBSE modela, dok 98,819 su klasifikacijski parametri za treniranje za 3 namjere, odnosno 99,077 za 5 namjera. Sveukupno treniranje modela zahtijeva nešto manje od minute nad CPUom za svaku epohu, što čini treniranjem brzim i učinkovitim.

Problem jezika se odlučilo riješiti na sljedeći način: stvorilo se 5 pod-setova koji se razlikuju prema korištenom jeziku (Tablica 5):

1. Hrvatski jezik (HR)
2. Automatski prevođen engleski (AUTO-ENG)
3. Ručno prevođen engleski (TRANS-ENG)

4. Kombinacija hrvatskog i automatski prevođen engleski (HR + AUTO-ENG)
5. Kombinacija hrvatskoga i ručno prevođen engleski (HR + TRANS-ENG)



Slika 16: Arhitektura modela

Cilj takve raspodjele je kako bi se moglo pratiti utjecaj odabranoga jezika na rezultate treniranja modela: da li će podaci sastavljeni od samo engleskih rečenica davati bolje rezultate s baznim modelom (LaBSE) kojem je primarni jezik engleski ili jezik neće imati utjecaj na rezultate. Dodatna značajka je da podaci će se pohranjivati modelu bez predobrade: razlog tome je gore navedeni problem rijetkih resursa te naknadno usporediti rezultate s predobrađenim podacima. Treba napomenuti da model ima ugrađen predobradni sloj (Slika 16) koji služi za pripremiti podatke za LaBSE enkoder: koristi predobradni proces za univerzalni rečenični enkoder koji podijeli rečenice u tokene te postavi specijalne tokene [CLS] i [SEP].

Pošto količina podataka nije velika i brzina treniranja je mala, moguće je trenirati više modela za evaluaciju (za svaki set podataka jedan model). Za treniranje modela se

Set podataka	Uzorak 3 namjere	Uzorak 5 namjera
HR	156	143
AUTO-ENG	156	143
TRANS-ENG	156	143
HR + AUTO-ENG	312	286
HR + TRANS-ENG	312	286

Tablica 5: Uzorci podataka na raspolaganju ovisno o korištenom jeziku

koristila stopa učenja od 0.001, funkcija gubitka je kategorična unakrsna entropija te s 5 epoha što prema autorima članka [75] je preporučeno od modela te u većini slučajeva je i previše.

5.4 Rezultati testova

Nakon treniranja svih modela bez predobrade dobije se 10 modela, jedan za svaki jezik i broj klasifikacijskih namjera. Dobivene rezultate se može vidjeti u Tablici 6. Za treniranje modela s 5 namjera se postavilo veći broj epoha zbog veće količine klasifikacijskih parametra koji model mora naučiti.

Iz Tablice 6 se može vidjeti da bez predobrade podataka model se može trenirati s preko 90% točnosti. Najbolji model s 3 namjere je model koji je treniran sa setom podataka sastavljenim od hrvatskih rečenica i automatski prevođenih rečenica na engleski jezik. Kod modela s većim brojem namjera međutim najveću točnost nad validacijom ima model treniran s ručno prevođenim rečenicama, ukazavši na to da za veći broj klasa bolje rezultate se dobije s nativnim jezikom modela.

Set podataka	br. namjera	epohe	treniranje	validacija
HR	3	5	0.91	0.85
AUTO-ENG	3	5	0.94	0.86
TRANS-ENG	3	5	0.93	0.92
HR + AUTO-ENG	3	5	**0.95**	**0.94**
HR + TRANS-ENG	3	5	0.93	0.93
HR	5	8	0.90	0.91
AUTO-ENG	5	8	**0.97**	0.91
TRANS-ENG	5	8	0.96	**0.94**
HR + AUTO-ENG	5	10	0.95	0.91
HR + TRANS-ENG	5	10	0.96	0.93

Tablica 6: Rezultati treniranja bez predobrade podataka

Slijedno treniranju bez predobrade teksta, radi usporedbe i evaluacije trenira se s predobrađenim podacima. Metode predobrade su opisani u poglavlju 3.1. Koraci predobrade su sljedeći:

1. Micanje interpunkcijskih znakova kao što su upitnici, navodnici, uskličnici itd.
2. Pretvaranje teksta u mala slova.

3. Tokenizacija rečenica odnosno odvajanje rečenica na riječi od koje je sastavljena: dobije se lista riječi.

4. Micanje čestica kao što su zamjenice.

5. Lematizacija odnosno svođenje riječi na tvorbeni korijen

Preskočio se korak stemanja zato što svodi riječi na korijene koje u nekim slučajevima nemaju sintaktičkog značenja te utječu na sveukupno značenje rečenice. Dodatna notacija u gore navedenim koracima je što su izvedeni samo za rečenice na engleskom jeziku, neovisno da li su prevedene ručno ili automatski. Za hrvatske rečenice, zbog navedene manjkavosti resursa, odrađeni su koraci 1., 2., 3. i 4. Zadnji korak je spojiti listu riječi u jednu rečenicu kako bi se je moglo pohraniti kao ulaz modelu.

Metoda predobrade	**Izlaz**
Izvorno	**where can I pay enrollment fees**
Interpunkcije	where can I pay enrollment fees
Mala slova	where can i pay enrollment fees
Tokenizacija	[where, can, i, pay, enrollment, fees]
Micanje čestica	[pay, enrollment, fees]
Lematizacija	[pay, enrollment, fee]

Tablica 7: Primjer rečenice nakon svakog koraka predobrade

S novim predobrađenim setom podataka se ponovno treniralo prijašnjih 10 modela. Dobiveni rezultati se mogu vidjeti u Tablici 8. U slučaju treniranih modela s predobrađenim podacima, model s kombinacijom hrvatskog jezika i ručnim prijevodom daje najveću točnost za klasifikaciju namjera iz Grupe_1. Kod modela za klasifikaciju namjera iz Grupe_2 ručni prijevod daje najbolje rezultate s čak 98% točnosti na validaciji. Iako model s kombinacijom hrvatskog jezika i ručnog prevođenja nema najvišu točnost, uzimajući u obzir razinu treniranja je također među boljima navodeći da pravilan prijevod igra ulogu u krajnjoj točnosti i performansama modela.

U Tablici 9 se uspoređuju rezultati modela prije predobrade i nakon predobrade. Iako predobrada ulaza nije donesla značajna povećanja u točnosti, svejedno daje bolje rezultate u većini slučajeva: za modele iz Grupe_1, 4 od 5 modela imaju bolje rezultate s

predobrađenim podacima, dok modeli iz Grupe_2, 3 od 5 modela imaju bolje rezultate s predobrađenim podacima te jedan model ima iste rezultate s objema vrste podataka.

Set podataka	br. namjera	epohe	treniranje	validacija
HR	3	5	0.93	**0.94**
AUTO-ENG	3	5	0.92	0.88
TRANS-ENG	3	5	0.94	**0.94**
HR + AUTO-ENG	3	5	**0.95**	0.93
HR + TRANS-ENG	3	5	**0.95**	**0.94**
HR	5	8	0.90	0.95
AUTO-ENG	5	8	0.88	0.89
TRANS-ENG	5	8	0.90	**0.98**
HR + AUTO-ENG	5	10	**0.94**	0.91
HR + TRANS-ENG	5	10	**0.94**	0.97

Tablica 8: Rezultati treniranja nakon predobrade podataka

Set podataka	br. namjera	Bez predobrade	Predobradom
HR	3	0.85	**0.94**
AUTO-ENG	3	0.86	**0.88**
TRANS-ENG	3	0.92	**0.94**
HR + AUTO-ENG	3	**0.94**	0.93
HR + TRANS-ENG	3	0.93	**0.94**
HR	5	0.91	**0.95**
AUTO-ENG	5	**0.91**	0.89
TRANS-ENG	5	0.94	**0.98**
HR + AUTO-ENG	5	0.91	0.91
HR + TRANS-ENG	5	0.93	**0.97**

Tablica 9: Usporedba rezultata modela bez predobrade i sa predobradom podataka

5.5 Zaključci provedenih testova

Iz gore navedenih rezultata može se zaključiti sljedeće stavke:

1. Korištenje rečenice na engleskom jeziku daje veću točnost nego kada se trenira s rečenicama na hrvatskom jeziku. U svim slučajevima modela bez predobrade podataka, korištenje engleskog jezika, čak u kombinaciji s hrvatskim jezikom, daje bolje rezultate; dok modeli s predobrađenim podacima pokazuju da pravilan ručni prijevod daje znatno bolje rezultate od ostalih modela. U ovom slučaju automatski prijevod daje lošije rezultate od svih modela.

2. U 8 od 10 modela predobrađeni podaci daju bolje rezultate od modela koji su trenirani bez predobrade podataka. Predobrada zahtijeva pravilnije rečenice zbog vrsta algoritama koje se koriste; posljedično tome, automatski prijevod koji sadržava primjetan broj grešaka daje najlošije rezultate od svih modela.

3. U svim testovima treniranja broj namjera za klasificiranje nije igrao ulogu u dostignuću razine točnosti modela. Jedina razlika stoji u vremenu učenja: s više namjera za klasifikaciju, klasifikacijski sloj je veći, pa je potrebno više epoha da model nauči parametre.

4. Poznato je da podaci igraju veliku ulogu u svijetu strojnog učenja i dubokog učenja odnosno što više to bolje. Podaci imaju toliki utjecaj da ako se ima dovoljno podataka može se slobodnije manipulirati s arhitekturom modela. Iz tog razloga, količina podataka za ovaj projekt je predstavljala jedan od primarnih mogućih zastoja razvoja. Iz rezultata se vidi da je moguće trenirati model s visokom točnosti s manje od 50 uzorka po namjeri. Razlog tome je korištenje predtreniranog rečeničnog enkodera koji mapira rečenice u vektorski prostor bez potrebe za dotjerivanja cijelog modela.

5. Već se utvrdilo da engleski jezik daje bolje rezultate od modela treniranih na hrvatskim podacima. Ako se uspoređuje modeli samo na engleskim jezikom ali s različitom vrstom prijevoda, ručni prijevod daje bolje rezultate u svim slučajevima. Značajnost pravilnog prijevoda se primjetno vidi u modelima s predobrađenim podacima: algoritmi za predobradu su sagrađeni za riječi na engleskom jeziku. Automatski prevoditelji, u slučajevima kada nisu sigurni za prijevod neke riječi

ili jednostavno nemaju prijevod, samo kopiraju izvornu riječ iz početne rečenice, što rezultira u prijevodu koji sadrži ciljani i izvorni jezik u jednoj rečenici. Za takve slučajeve, predobradni alati nisu dizajnirani.

Za kraj je preostalo odabrati adekvatan model za objašnjeni problem. Već se odredilo da engleski jezik daje bolje rezultate te kod usporedbe automatskog prevođenja i ručnog, ručni prijevod daje veću točnost. Međutim kod odabira modela za korištenje, također treba razmatrati dostupne resurse i potrebno vrijeme izvedbe. Za ručni prijevod je potrebno uložiti vrijeme kako bi se prevelo rečenice. Pošto u procesu prepoznavanja namjera nema mjesta za čekanje ručnog prijevoda, potrebno je implementirati alat za automatsko prevođenje koje je specifično namijenjeno za hrvatski-engleski prijevod te tako dobivati bolje prijevode od Googleovog prevoditelja. Druga stavka koju treba razmotriti je dostupnost resursa za predobradu podataka. Hrvatski jezik daje bolje rezultate od automatskog prevođenja, ali treba uočiti da raspoloživi alati za hrvatski jezik su jako rijetki ili ne postojeći: paket korišten u ovom radu za procesiranje hrvatskih rečenica je zadnjeg puta ažuriran 2020. godine a prije toga 2012. godine. Na suprot tome, NLTK platforma je redovito ažurirana te ima veliki ekosustav podrške. Zadnja točka koju treba primijetiti je da iako postojeće, razlike u rezultatima su malene: najgori rezultat je 0.85 točnosti koji pripada modelu s isključivo hrvatskim jezikom bez procesiranja.

Nakon razmatranja kriterija, može se zaključiti da postoje dvije opcije za odabir modela:

1. Odabir modela s *ručno prevedenim rečenicama*: odabirom ove vrste potrebno je najprije implementirati automatskog prevoditelja koji će raditi bolje prijevode od Googleovog prevoditelja. Nakon toga prednosti su očite. Pravilan prijevod i engleski jezik daju bolje rezultate te je jednostavnije implementirati buduće radove na projektu.

2. Odabir modela s *automatskim prijevodom*: automatizacija prijevoda ubrzava proces ali daje lošije rezultate od ostalih modela. Međutim, uzimajući u obzir dostupnost resursa za engleski jezik naprotiv dostupnosti resursa za hrvatski jezik, lošiji rezultat postaje beznačajan.

Sljedeći koraci i budući rad na ovom projektu imaju široki raspon koji vode cilju stva-

ranja cjelovitog funkcionalnog chatbota. U domeni prepoznavanja namjere testiranje s manjim rečeničnim enkoderima te dodavanjem većeg broja namjera. Sljedeći korak bi bio izvlačenje entiteta iz ulazne rečenice kao što su imena i brojčanih podataka: ispunjavanje potrebnih varijabli iz BPMN modela s izvučenim podacima iz upita (*eng. slot filling models* [76]). Stvaranje procese odgovora te sastavljanje ugovora dijaloga za pravilno odgovaranje i predviđanje najboljeg sljedećeg koraka chatbota.

Zaključak

U ovom radu se razmotrilo tehnologije, metode, razmišljanja potrebna za uspješnu detekciju namjere korisnika tijekom komunikacije sa strojem. Prepoznavanje namjere je prvi korak za održivu interakciju. Nestrukturirani podaci su najčešći u prirodnom okruženju te govor ili tekstualni prirodni jezik je jedan od češćih primjera takvog oblika. Konstantan razvoj u domeni NLP-a je omogućilo takve nestrukturirane podatke pretvoriti u nešto korisno i shvatljivo za računala. U ovom radu se uspješno sagradio model za prepoznavanje namjera s točnosti preko 90%. Izazov je bio rad s hrvatskim jezikom zbog malo raspoloživih resursa za rad s rečenicama na hrvatskom. Međutim izazov je bio predvidljiv pošto hrvatski jezik ima mali broj govornika: procjenjuje se da globalno postoje oko 7 milijuna govornika [77] (za usporedbu, govornika engleskog jezika ima 1.348 milijardi [78]). Posljedično tome, interes za razvoj alata za procesiranje jezika s niskim brojem govornika je mala. Dodatni izazov je bio količina dostupnih podataka. Praksa kaže da za treniranje modela dubokog učenja je potrebna velika količina podataka. Iako u ovom radu se koristi predtrenirani model za kodiranje rečenica, klasifikacijski sloj je treniran od nule. Prema rezultatima ovog rada potvrdile su se tvrdnje autora drugih radova [44] da je moguće trenirati model dubokog učenja i s malo dostupnih podataka.

Prepoznavanje namjere je jedna od osnovnih primjena NLP-a te predstavlja vitalnu točku za bilo koji konverzacijski sustav. Zahvaljujući napretku NLP-a moguće je pomaknuti se od starih metoda prepoznavanja namjera kao što je izvlačenje ključnih riječi za mapiranje odgovora. Moderan NLP omogućava kontekstualno i sintaktičko shvaćanje rečenica. Nove tehnike približavaju kognitivnu sposobnost stroja korak bliže sposobnostima čovjeka. Da bi računalo mogao napredovati s performansama shvaćanja ulaznog upita, postoje nekoliko ključnih izazova za riješiti:

- Direktno povezano s prepoznavanje namjere, chatbotovi često znaju krivo protumačiti rečenicu. Razlog tome je što NLP radi s jako nestrukturiranim podacima. Rečenica može sadržavati više jezika, žargone, spomenute gramatičke greške itd.
- Izvedba nepreciznih naredba. Korisnici često imaju dvojbene namjere unutar

jedne izjave.

- U slučaju verbalnog procesiranja, modeli imaju poteškoće procesirati dijalekte i razne naglaske kako bi ispravno mogao prepoznati namjere

Prema istraživanju [79], 80% tvrtka namjerava imati neki oblik chatbot automatizacije, dok 50% klijenata očekuje da korisnička služba bude konstantno dostupna. Implementacija chatbota može smanjiti do 30% troškova korisničke službe. Udio tržišta automatiziranih konverzacijskih sustava, bilo to chatbot ili neki oblik virtualnog asistenta, će konstantno rasti te njegov rast je direktno povezan s napretkom tehnologije umjetne inteligencije, preciznije u domeni NLP-a.

Literatura

[1] Andrea L. Guzman. Human-Machine Communication: Rethinking Communication, Technology, and Ourselves. *link: https://andrealguzman.net/new-hmc-book*, 2018.

[2] Izidor Mlakar, Darinka Verdonik, Simona Majhenič, and Matej Rojc. Towards Pragmatic Understanding of Conversational Intent: A Multimodal Annotation Approach to Multiparty Informal Interaction – The EVA Corpus. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2019.

[3] Andrea L. Guzman. Voices in and of the machine: source orientation toward mobile virtual assistants. *Computers in Human Behavior*, 2019.

[4] Andrea L. Guzman and Seth C. Lewis. Artificial intelligence and communication: A Human–Machine Communication research agenda. *New Media and Society*, 2020.

[5] Google. Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone . *link: https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html*, 2018.

[6] Even Lokøy. Voice it! Will virtual assistants become a vital part of the future? *link: https://www.tietoevry.com/en/blog/2020/04/will-virtual-assistants-become-a-vital-part-of-the-future/*, 2020.

[7] PolyAI. PolyAI. *link: https://www.polyai.com/*, 2021.

[8] Drift.com. Conversational Systems Market - Growth, Trends, COVID-19 Impact, and Forecasts (2021 - 2026) . *link: https://www.mordorintelligence.com/industry-reports/conversational-systems-market*, 2021.

[9] Joyce Chai, Chen Zhang, and Tyler Baldwin. Towards conversational QA: automatic identification of problematic situations and user intent. *Proceedings of the COLING/ACL on Main conference poster sessions*, 2006.

[10] Nikhita Vedula, Nedim Lipka, Pranav Maneriker, and Srinivasan Parthasarathy. Towards Open Intent Discovery for Conversational Text. *Proceedings of ACM Conference*, 2019.

[11] Eleni Adamopoulou and Lefteris Moussiades. An Overview of Chatbot Technology. *IFIP Advances in Information and Communication Technology*, 2020.

[12] Anirudh Khanna, Bishwajeet Pandey, Kushagra Vashishta, Kartik Kalia, Bhale Pradeepkumar, and Teerath Das. A Study of Today's A.I. through Chatbots and Rediscovery of Machine. *International Journal of u- and e-Service, Science and Technology*, 2015.

[13] Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, Eric King, Kate Bland, Amanda Wartick, Yi Pan, Han Song, Sk Jayadevan, Gene Hwang, and Art Pettigrue. Conversational AI: The Science Behind the Alexa Prize. *http://arxiv.org/abs/1801.03604*, 2018.

[14] Alan M. Turing. *Computing machinery and intelligence*. 1950.

[15] Abdul Kader A. Sameera and John Woods. Survey on Chatbot Design Techniques in Speech Conversation Systems. *International Journal of Advanced Computer Science and Applications*, 2015.

[16] Joseph Weizenbaum. ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 1966.

[17] Kenneth Mark Colby. Modeling a paranoid mind. *Behavioral and Brain Sciences*, 1981.

[18] Jack Cahn. CHATBOT: Architecture, Design, Development. *University of Pennsylvania, School of Engineering and Applied Science*, 2017.

[19] Artificial Solutions. Types of Chatbot Technology. *link: https://medium.com/voice-tech-podcast/types-of-chatbot-technology-72d095df2540*, 2019.

[20] Jan Recker, Marta Indulska, and Peter Green. How good is BPMN really? Insights from theory and practice. *European Conference on Information Systems*, 2006.

[21] Vural Taner Yilmaz, F. Fevzi Ersoy, Huseyin Kocak, Gulsen Yakupoglu, and Gultekin Suleymanlar. Introduction to BPMN. *Turkish Nephrology Dialysis Transplantation*, 2012.

[22] Michele Chinosi and Alberto Trombetta. BPMN: An introduction to the standard. *Computer Standards and Interfaces*, 2012.

[23] Matthias Geiger, Simon Harrer, Jörg Lenhard, and Guido Wirtz. BPMN 2.0: The state of support and implementation. *Future Generation Computer Systems*, 2018.

[24] Faizel Sedick and Lisa F. Seymour. BPMN usage: An analysis of influencing factors. *Advances in Enterprise Information Systems II*, 2012.

[25] Diego Lopez Yse. Your Guide to Natural Language Processing (NLP). *link: https://towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e1*, 2019.

[26] Chaitanya Krishna Kasaraneni. Understanding NLP Pipeline. *link: https://medium.com/analytics-vidhya/understanding-nlp-pipeline-9af8cba78a56*, 2020.

[27] Wikipedia. Garbage in, garbage out. *link: https://en.wikipedia.org/wiki/Garbage_in,_garbage_out*, 2021.

[28] Rob Stenson. Is This the First Time Anyone Printed, 'Garbage In, Garbage Out'? *link: https://www.atlasobscura.com/articles/is-this-the-first-time-anyone-printed-garbage-in-garbage-out*, 2016.

[29] C. S. Pavan Kumar and L. D. Dhinesh Babu. Novel Text Preprocessing Framework for Sentiment Analysis. *Smart Innovation, Systems and Technologies*, 2018.

[30] Y. A. Solangi, Z. A. Solangi, S. Aarain, A. Abro, G. A. Mallah, and A. Shah. Review on Natural Language Processing (NLP) and Its Toolkits for Opinion Mining and Sentiment Analysis. *IEEE 5th International Conference on Engineering Technologies and Applied Sciences (ICETAS)*, 2018.

[31] Kavita Ganesan. What are Stop Words? *link: https://www.opinosis-analytics.com/knowledge-base/stop-words-explained*, 2019.

[32] Behrang Mohit. Named Entity Recognition. *Theory and Applications of Natural Language Processing*, 2014.

[33] Anjali Ganesh Jivani. A Comparative Study of Stemming Algorithms. *IJCTA*, 2011.

[34] A. Chakrabarty, A. Chaturvedi, and U. Garain. CNN-based Context Sensitive Lemmatization. *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data - CoDS-COMAD '19*, 2019.

[35] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

[36] Andrea Galassi, Marco Lippi, and Paolo Torroni. Attention in Natural Language Processing. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[37] Nikhil Agrawal. Understanding Attention Mechanism: Natural Language Processing. *link: https://medium.com/analytics-vidhya/https-medium-com-understanding-attention-mechanism-natural-language-processing-9744ab6aed6a*, 2020.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[39] Prateek Joshi. How do Transformers Work in NLP? A Guide to the Latest State-of-the-Art Models. *link: https://www.analyticsvidhya.com/blog/2019/06/understanding-transformers-nlp-state-of-the-art-models/*, 2019.

[40] Aditi Mittal. Understanding RNN and LSTM. *link: https://aditi-mittal.medium.com/understanding-rnn-and-lstm-f7cdf6dfc14e*, 2019.

[41] Rani Horev. BERT Explained: State of the art language model for NLP. *link: https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270*, 2018.

[42] Jacob Devlin. mBERT. *link: https://github.com/google-research/bert/blob/master/multilingual.md*, 2019.

[43] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *google*, 2016.

[44] Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient Intent Detection with Dual Sentence Encoders. *link: http://dx.doi.org/10.18653/v1/2020.nlp4convai-1.5*, 2020.

[45] Thomas Holtgraves. Automatic intention recognition in conversation processing. *Journal of Memory and Language*, 2008.

[46] P. Brown and S. Levinson. Politeness: Some universals in language use. *Cambridge University Press*, 1987.

[47] Maria Zajaeczkowska, Kirsten Abbot-Smith, and Christina S. Kim. Using shared knowledge to determine ironic intent; A conversational response paradigm. *Journal of Child Language*, 2020.

[48] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations ofwords and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 2013.

[49] Andreas Pogiatzis. NLP: Contextualized word embeddings from BERT. *link: https://towardsdatascience.com/nlp-extract-contextualized-word-embeddings-from-bert-keras-tf-67ef29f60a7b*, 2019.

[50] Hongmin Li, Xukun Li, Doina Caragea, and Cornelia Caragea. Comparison of Word Embeddings and Sentence Encodings as Generalized Representations for

Crisis Tweet Classification Tasks. *Proceedings of the ISCRAM Asian Pacific 2018 Conference*, 2018.

[51] Purva Huilgol. Top 4 Sentence Embedding Techniques using Python! *link: https://www.analyticsvidhya.com/blog/2020/08/top-4-sentence-embedding-techniques-using-python/*, 2020.

[52] Diogo Ferreira. From Word Embeddings to Sentence Embeddings. *link: https://medium.datadriveninvestor.com/from-word-embeddings-to-sentence-embeddings-part-2-3-21a5b03592a1*, 2020.

[53] Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2020.

[54] Quoc Le and Tomas Mikolov. Distributed Representations of Sentences and Documents. *Proceedings of the 31 st International Conference on Machine Learning*, 2015.

[55] Mikel Artetxe and Holger Schwenk. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 2019.

[56] Robert Litschko, Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. Evaluating Multilingual Text Encoders for Unsupervised Cross-Lingual Retrieval. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2021.

[57] Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. Multilingual Universal Sentence Encoder for Semantic Retrieval. *ACL 2020*, 2020.

[58] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical*

Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference, 2020.

[59] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT Sentence Embedding. *ArXiv*, 2020.

[60] Megagon Labs. Emu: Enhancing Multilingual Sentence Embeddings with Semantic Similarity. *link: https://megagon.ai/blog/emu-enhancing-multilingual-sentence-embeddings-with-semantic-similarity/*, 2020.

[61] Bijula Ratheesh. Word Embeddings, WordPiece and Language-Agnostic BERT (LaBSE). *link: https://medium.com/mlearning-ai/word-embeddings-wordpiece-and-language-agnostic-bert-labse-98c7626878c7*, 2021.

[62] Fathy Rashad. Additive Margin Softmax Loss (AM-Softmax). *link: https://towardsdatascience.com/additive-margin-softmax-loss-am-softmax-912e11ce1c6b*, 2020.

[63] Yinfei Yang, Gustavo Hernandez Abrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun Hsuan Sung, Brian Strope, and Ray Kurzweil. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax). *IJCAI International Joint Conference on Artificial Intelligence*, 2019.

[64] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. *Association for Computational Linguistics*, 2017.

[65] Massimiliano Dibitonto, Katarzyna Leszczynska, Federica Tazzi, and Carlo M. Medaglia. Chatbot in a campus environment: Design of lisa, a virtual assistant to help students in their university life. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018.

[66] Lindsay C. Page and Hunter Gehlbach. How an Artificially Intelligent Virtual Assistant Helps Students Navigate the Road to College. *AERA Open*, 2017.

[67] Ajay Chatterjee and Shubhashis Sengupta. Intent Mining from past conversations for Conversational Agent. *10.18653/v1/2020.coling-main.366*, 2021.

[68] Ajit Rajasekharan. Quantitative evaluation of a pre-trained BERT model. *link: https://towardsdatascience.com/quantitative-evaluation-of-a-pre-trained-bert-model-73d56719539e*, 2021.

[69] Google. Tensorflow. *link: https://www.tensorflow.org/*, 2021.

[70] François Chollet. Keras. *link: https://keras.io/*, 2021.

[71] AQR Capital Management. Pandas library. *link: https://pandas.pydata.org/*, 2021.

[72] AQR Capital Management. Seaborn library. *link: https://seaborn.pydata.org/*, 2021.

[73] Steven Bird, Edward Loper, and Ewan Klein. Natural Language Toolkit. *link: https://www.nltk.org/*, 2021.

[74] Robert Lujo. Text-hr. *link: https://pypi.org/project/text-hr/*, 2020.

[75] Vinura Dhananjaya. Fine-Tuning “LaBSE” for a Sentiment Classification Task. *link: https://towardsdatascience.com/fine-tuning-labse-for-a-sentiment-classification-task-56e34b74e655*, 2021.

[76] Sebastian Ruder. Intent Detection and Slot Filling. *link: http://nlpprogress.com/english/intent_detection_slot_filling.html*, 2021.

[77] Wikipedia. Hrvatski jezik. *link: https://hr.wikipedia.org/wiki/Hrvatski_jezik*, 2021.

[78] Wikipedia. List of languages by total number of speakers. *link: https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers*, 2021.

[79] Snigdha Patel. Top 12 Chatbots Trends and Statistics to Follow in 2021. *link: https://www.revechat.com/blog/chatbots-trends-stats/*, 2021.

Popis tablica

1 Usporedba lematizacije i stemanja 13
2 Usporedba performansi modela nad Tatoeba [64] korpusu [59] 25
3 Primjer zapisa u bazi podataka 30
4 Primjer klasifikacije rečenica prema pripadajućim namjerama Grupe_1 i Grupe_2 32
5 Uzorci podataka na raspolaganju ovisno o korištenom jeziku 36
6 Rezultati treniranja bez predobrade podataka 37
7 Primjer rečenice nakon svakog koraka predobrade 38
8 Rezultati treniranja nakon predobrade podataka 39
9 Usporedba rezultata modela bez predobrade i sa predobradom podataka 39

Popis slika

1 Globalna upotreba chatbotova [8] 2
2 Osnovna arhitektura chatbota [11] 6
3 Jednostavan primjer BPMN dijagrama [21] 8
4 Osnovni postupci NLP procesiranja [26] 11
5 Tehnika pozornosti. Dekoder se računa sa kontekstualnim vektorom, s prethodnim izlazom, prethodnim skrivenim stanjem te posebnim kontekstualnim vektorom za svaku ciljanu riječ. Ti vektori se računaju kao težinski zbroj za aktivacijsko stanje te također predstavljaju koliku pozornost će dobiti za generiranje izlazne riječi [37] 14
6 Arhitektura modela s transformersima [38] 15
7 Reprezentacija višestruke pažnje [38] 16
8 BERT konceptualna arhitektura [42] 18
9 Reprezentacija riječi u vektorskom prostoru [51] 20
10 Reprezentacija višejezičnih rečenica u vektorskom prostoru: neovisno o jeziku, kodirano značenje rečenice se mora nalaziti na sličnim pozicijama [60] 22
11 LaBSE arhitektura [59] 24
12 Rezultat apliciranja funkcije softmax aditivne margine [63] 25
13 Promjena dužine rečenice prijei nakon prevađanja na engleski jezik . . 30
14 Distribucija namjera Grupa_1 31
15 Distribucija namjera Grupa_2 32
16 Arhitektura modela 35

Sažetak

Razvoj u području obrade prirodnog jezika (NLP) u zadnjih je godina unaprijedilo mogućnosti rješenja u domeni komunikacije sa strojem. Sve popularniji chatbotovi i virtualni asistenti zamjenjuju ljude u ponavljajućim i statičnim procesima kao odgovaranje na pitanja ili rješavanje jednostavnih problema. Do nedavno su se domene komunikacije i strojeva smatrale kao odvojena područja istraživanja ali uvodom novih tehnologija kao što su tehnike pozornosti i transformeri smanjuju udaljenost domena do razine da osim davanja točnog odgovora, sada se također pazi na osjećaje, skrivene namjere, kulturalne razlike itd. sugovornika sa strojem.

U ovome radu se razmatra prvi korak održive komunikacije s čovjekom, odnosno prepoznavanje namjere. Istražuju se postojeće metode i tehnologije u području NLP-a za krajnje prepoznavanje namjere. Za svrhu dokazivanja funkcionalnosti razvio se set modela s preko 90% točnosti za prepoznavanje namjere za sveučilišne procese kao što su upisi na sveučilište te prijava završnog ili diplomskoga rada. Za kraj, dodatan doprinos rada je razmatranje utjecaja korištenog jezika za prepoznavanje namjere u višejezičnim uvjetima.

Ključne riječi: *NLP, prepoznavanje namjere, chatbot, višejezičnost, konverzacijski sustavi, umjetna inteligencija*

Abstract

Recent developments in the field of natural language processing (NLP) have increased capabilities in the domain of human-machine communication. Employees for repetitive and static processes like question answering or resolving simple problems are being replaced by the increasingly present chatbots and virtual assistants. Until recently, communication and machines were considered two separate research domains. However, new technologies like attention mechanism and transformers architecture are closing the gap between the two domains to the level that, besides giving the correct answer to a question, machines are aware of user's sentiments, hidden intentions, cultural differences etc.

This thesis elaborates on the first step of sustainable communication with a human being: intent recognition. The paper research on existing methods and technologies in the field of NLP for the end objective to recognize the user's intent. For the purpose of showing functionalities of the methods, a set of models were developed with accuracy over 90% for classifying intents in the domain of university processes like course enrollments and thesis registration. As an additional contribution of this paper is the research on the influence of language for intent recognition in a multilingual environment.

Key words: *NLP, intent recognition, chatbot, multilingual, communication systems, artificial intelligence*