Programski modul za odgovore na pitanja o procesima

Krstačić, Rafael

Sveučilište Jurja Dobrile

Fakultet informatike u Puli

Rafael Krstačić

Programski modul za odgovore na pitanja o procesima

Završni rad

Sveučilište Jurja Dobrile

Fakultet informatike u Puli

Rafael Krstačić

Programski modul za odgovore na pitanja o procesima

Završni rad

Ime Prezime studenta, JMBAG: Rafael Krstačić, 0303092283

Studijski smjer: preddiplomski sveučilišni studij informatika

Znanstveno područje: Društvene znanosti

Znanstveno polje: Informacijske i komunikacijske znanosti

Znanstvena grana: Informacijski sustavi i informatologija

Kolegij: Programsko Inženjerstvo

Mentor: doc. dr. sc. Nikola Tanković

IZJAVA O AKADEMSKOJ ČESTITOSTI

Ja, dolje potpisani Rafael Krstačić, kandidat za prvostupnika Informatike ovime izjavljujem da je ovaj Završni 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 Završnog 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

Rk

U Puli, 14.09.2022.

IZJAVA O KORIŠTENJU AUTORSKOG DJELA

Ja, Rafael Krstačić dajem odobrenje Sveučilištu Jurja Dobrile u Puli, kao nositelju prava iskorištavanja, da moj Završni rad pod nazivom Programski modul za odgovaranje na pitanja o procesima

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, 14.09.2022.

Potpis

Rk

Sadržaj

Sažetak

Odgovaranje na pitanja o poslovnim procesima je posao koji najčešće obavlja nekolicina ljudi, ponavlja se i izvodi na sličan način. U cilju vremenske optimizacije i smanjenje troškovnih aspekata tog posla, organizacije mogu u svoje poslovanje integrirati umjetnu inteligenciju, posebice tehnike strojnog učenja kako bi se automatizirao taj posao. Ovaj rad predlaže i provodi prvi korak prema implementaciji takvih automatizacija. Rad se sastoji od programskog rješenja konverzacijskog sučelja otvorenog koda i dokumentacije istog. Sučelje se temelji na integraciji postojećih rješenja za rješavanje potproblema cilja. Ovo konverzacijsko sučelje ima mogućnost odgovarati na pitanja o jednostavnim poslovnim procesima nad kojima je trenirano. Kroz ovu dokumentaciju daje se uvod u poslovne procese i konverzacijska sučelja te se opisuje arhitektura i metodologija programskog djela rada.

Ključne riječi

Chatbot, konverzacijsko sučelje, strojno učenje, često postavljena pitanja, poslovni procesi, programski modul, transformeri

Abstract

Answering questions about business processes is a job that is usually done by a few people, is repetitive and is performed in a similar way. To optimize the time component and reduce the cost aspects of this work, organizations can integrate artificial intelligence into their operations, specifically machine learning techniques, to automate the work. This paper proposes and implements the first step towards the implementation of such automations. The work consists of the software solution of the open-source conversational interface and its documentation. The interface is based on integration with already existing solutions for solving the sub-problems of the goal. This conversational interface can answer questions about simple business processes it has been trained on. Through this documentation, an introduction to business processes and conversational interfaces is given, and the architecture and methodology of the programming part of the work is described.

Keywords

Chatbot, conversational interface, machine learning, frequently asked questions, business processes, program module, transformers

1. Uvod

Sam postupak odgovaranja na pitanja o poslovnim procesima je štoviše repetitivan zadatak ako se radi o pitanjima koja se često postavljaju. Za ovaj postupak najčešće je zadužena nekolicina ljudi koja odgovara na pitanja putem elektroničke pošte, aplikacije za izravnu razmjenu poruka ili nekim drugim medijem. S rastućim trendom korištenja umjetne inteligencije, organizacije mogu automatizirati ovaj proces integracijom konverzacijskih sučelja (eng. chatbot).

Ova dokumentacija opisuje izrađeni programski modul za odgovaranje na pitanja o poslovnim procesima. Preciznije, opisuje arhitekturu samog modula, korištene modele u samom razvoju te fazu treniranja i rezultate modula.

Rad je podijeljen u nekoliko poglavlja. Prvo poglavlje se odnosi na sam uvod u dokumentaciju. Drugo se poglavlje odnosi na opis važnosti kvalitetne izvedbe poslovnih procesa i BPMN notaciju. U trećem poglavlju se dotiče domena umjetne inteligencije i konverzacijskih sučelja. Četvrto je poglavlje namijenjeno opisu korištenih modela u svrhu izrade modula. Kroz peto poglavlje se opisuje arhitektura samog modela. Postupak treniranja je opisan u šestom poglavlju, a rezultati u sedmom poglavlju. U osmom poglavlju se nalazi zaključak o modulu. Korištena literatura nalazi se u devetom poglavlju.

Poveznica na GitHub repozitoriji: https://github.com/rkrstacic/Software-module-for-answering-questions-on-processes

2. Poslovni procesi i BPMN

Kako se ovaj modul temelji na odgovaranje na pitanja o procesima, bitno je odrediti što su procesi i kakvu notaciju koristimo za prikaz istih.

Poslovni proces je kombinacija skupa aktivnosti unutar poduzeća sa strukturom koja opisuje njihov logičan poredak i ovisnost čiji je cilj proizvesti željeni rezultat. Modeliranje poslovnih procesa omogućuje zajedničko razumijevanje i analizu poslovnog procesa, a model procesa može pružiti sveobuhvatno razumijevanje procesa [1].

Poslovni procesi su od iznimne važnosti jer su vodič koji opisuje kako se stvari rade na najbolji mogući način i olakšava fokusiranje na poboljšanje poslovnih procesa. Poslovni procesi imaju ključnu ulogu u učinkovitom i djelotvornom funkcioniranju organizacije i strukture [2]. Dobro planirani i organizirani poslovni procesi pomažu u:

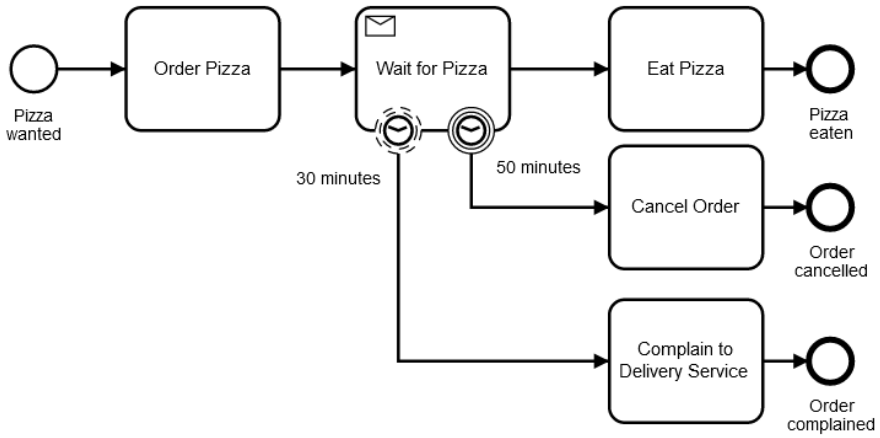
1. Smanjenju rizika i troškova
2. Smanjenju ljudskih pogrešaka
3. Poboljšanju učinkovitosti
4. Poboljšanu usredotočenosti na kupca
5. Učinkovitoj komunikaciji
6. Poboljšanju upravljanja vremenom
7. Prilagodbi novih tehnologija

Inicijativa za upravljanje poslovnim procesima (BPMI) razvila je standardnu notaciju za modeliranje poslovnih procesa (BPMN). Ova specifikacija predstavlja više od dvije godine truda *Notation Working Group* skupine za notaciju BPMI. Primarni cilj nastojanja BPMN-a bio je pružiti notaciju koja je lako razumljiva svim poslovnim korisnicima, od poslovnih analitičara koji stvaraju početne nacrte procesa, do tehničkih programera odgovornih za implementaciju tehnologije koja će izvoditi te procese, konačno, gospodarstvenicima koji će upravljati i pratiti te procese [3].

Glavne komponente BPMN modela su objekti toka (eng. flow objects), objekti za povezivanje, „grupe“ (eng. swimlanes) i podaci (eng. data/artefacts). Objekti toka su dijelovi koji tvore cjelokupni tijek rada. Tri glavna objekta toka poznata su kao događaji (eng. events), aktivnosti (eng. activities) i pristupnici (eng. gateways). Događaji su okidači procesa, a oni mogu biti u obliku standardnog okidača, poruke kao okidača, greške kao okidača, vrijeme kao okidač i ostali. Aktivnost je generički pojam za posao koji tvrtka obavlja u procesu. Aktivnost može biti

atomska ili složena (spoj). Vrste aktivnosti koje su dio modela procesa su: potproces i zadatak. Zadatak je atomska aktivnost koja se koristi kada rad u procesu nije raščlanjen na finiju razinu detalja procesa, u suprotnom radi se o potprocesu. Pristupnici određuju kojim se putem prolazi kroz proces. Objekti za povezivanje povezuju objekte toka, predstavljajući odnose i ovisnosti među njima. Od objekata za povezivanje najčešće se koriste tijek sekvence (prikaz reda izvršavanja aktivnosti), tijek poruke i pridruživanje (spajanje artefakata i informacija s nekim BPMN elementom). Swimlanes su grafička reprezentacija sudionika procesa. Podatci (eng. data/artefacts) mogu biti: bilješka, grupa te podatkovni objekti. Bilješka je običan komentar unutar samog programa za modeliranje, grupa predstavlja grupiranje grafičkih BPMN elemenata (ne utječe na tijek sekvence). Podatkovni objekti pružaju informacije o tome koje aktivnosti zahtijevaju da se izvrše i što proizvode.

Na slici ispod nalazi se jedan primjer BPMN modela procesa naručivanja pizze. Proces započinje događajem „Želja za pizzom“. Nakon događaja slijedi aktivnost naručivanje pizze. Kada se pizza naruči, onda slijedi čekanje pizze. U ovom koraku moguća su 3 ishoda. Prvi ishod je pojesti pizzu kada stigne te proces završava događajem „Pizza pojedena“. Drugi ishod se izvršava ako prođe 50 minuta prekida se narudžba, aktivnost „Čekanje pizze“ se više ne izvršava te proces završava događajem „Narudžba prekinuta“. Posljednji ishod je ako prođe 30 minuta od čekanja pizze, kupac šalje pritužbu na uslugu i okine se događaj „Tužba na uslugu“, no aktivnost „Čekanje pizze“ još uvijek traje dok se ne dogodi ishod 1 ili 2.

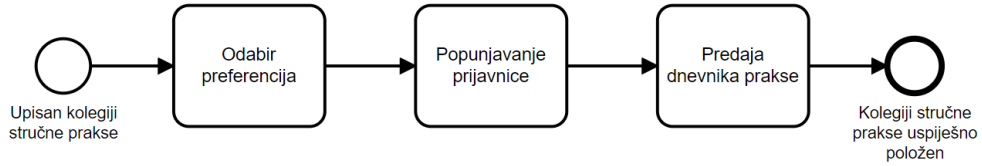


Slika 1. Primjer naručivanja pizze [13]

U svrhu ovog modula linearni se poslovni procesi (sadrži početni i završni događaj te zadatke spojene slijedom sekvence) zapisani u BPMN notaciji mapiraju u jedan JSON objekt koji se koristi prilikom donašanja odgovora za pojedini upit korisnika. Poslovni procesi koji su se

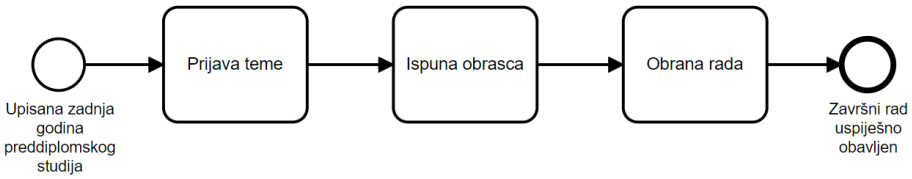
koristili u svrhu ovog rada su poslovni proces izvršavanja stručne prakse studenata na fakultetu te poslovni proces izrade i obrane završnog rada na fakultetu.

Modelirani proces izvršavanja stručne prakse je pojednostavljen model izvršavanja stručne prakse, a sastoji se od sljedećih zadataka: odabir preferencija, popunjavanja prijavnice i predaje dnevnika prakse. Za svaki zadatak su dostupne informacije o nazivu, sličnim izrazima, opisu te vremenu trajanja, a za sam proces su dostupni naziv, zadatci i vrijeme trajanja cjelokupnog procesa. Na slici ispod prikazan je proces u notaciji BPMN.



Slika 2. Proces izvršavanje stručne prakse

Modelirani proces izrade i obrane završnog rada je također pojednostavljeni model realnog procesa izrade i obrane završnog rada, a sastoji se od prijave teme, ispunjavanja obrasca te obrane rada. Kao i u prethodnom procesu, za svaki zadatak su dostupne informacije o nazivu, sličnim izrazima, opisu te vremenu trajanja, a za sam proces su dostupni naziv, zadatci i vrijeme trajanja cjelokupnog procesa. Slika ispod prikazuje model kroz BPMN.



Slika 3. Proces izrade i obrane završnog rada

3. Konverzacijska sučelja

Postoje razne definicije konverzacijskih sučelja (eng. chatbot), a neke od njih su sljedeće:

- Chatbot se može definirati kao „Računalni program dizajniran za simulaciju razgovora s ljudskim korisnicima, posebno putem interneta“. Takva sučelja su također poznata kao pametni botovi, interaktivni agenti, digitalni pomoćnici ili umjetni entiteti razgovora [4].
- Chatbot je račun za razmjenu trenutnih poruka koji može pružati usluge pomoću okvira za razmjenu trenutnih poruka s ciljem pružanja usluga razgovora korisnicima na učinkovit način [5].
- Chatbot je računalni program ili umjetna inteligencija koji vodi razgovore putem zvuka ili teksta te komunicira s korisnicima u određenoj domeni ili temi dajući inteligentne odgovore na prirodnom jeziku [6].

Može se reći kako je chatbot računalni program koji je izrađen s ciljem interakcije čovjeka i računala putem teksta ili govora, a bazira se na tehnologiji obrade prirodnog jezika (eng. Natural Language Processing). Obrada prirodnog jezika je potpodručje lingvistike, računalnih znanosti i umjetne inteligencije koje se bavi interakcijama između računala i ljudskog jezika, posebno kako programirati računala za obradu i analizu velikih količina podataka prirodnog jezika [7].

Namjena chatbotova je široka. Chatbot se može koristiti u korisničkim službama, kao društvena i emocionalna podrška, u svrhu prikupljanja informacija, kao izvor zabave ili kao način povezivanja korisnika s drugim ljudima ili strojevima [8].

Arhitektura chatbotova je raznolika i teško je pronaći dva chatbota koji implementiraju potpuno istu arhitekturu, ali različiti chatbotovi imaju neke zajedničke dijelove. Opće komponente u dizajnu chatbota su grafičko korisničko sučelje (GUI), pozadina i kernel, koji se obično brine za NLP, i baza podataka. Ovo odvajanje čini GUI (ili frontend), backend i kernel apstraktnim i omogućuje kombinaciju različitih programskih jezika ili tehnologija [9].

Postoje razne vrste chatbotova, ali se ih po načinu izvođenja mogu se podijeliti na tri glavne kategorije:

U prvu kategoriju spadaju chatbot-ovi temeljeni na pravilima tj. botovi temeljeni na stablu odlučivanja. Kao što ime sugerira, koriste niz definiranih pravila. Ova su pravila osnova za

vrste problema s kojima je chatbot upoznat i za koje može ponuditi rješenja. Ovi chatbot-ovi mapiraju razgovore u iščekivanju onoga što bi korisnik mogao pitati i kako bi chatbot trebao odgovoriti. Chatbot-ovi temeljeni na pravilima mogu koristiti vrlo jednostavna ili komplicirana pravila. Međutim, oni ne mogu odgovoriti na pitanja izvan definiranih pravila. Ovi chatbot-ovi ne uče kroz interakcije. Također, izvode i rade samo sa scenarijima na kojima su trenirani.

U drugu kategoriju spadaju chatbot-ovi koji donose odluke vođene umjetnom inteligencijom. Ova vrste chatbota koriste strojno učenje (ML) i umjetnu inteligenciju (AI) za pamćenje razgovora s određenim korisnicima kako bi s vremenom učili i napredovali. Oni su dovoljno pametni da se sami poboljšaju na temelju onoga što korisnici traže i kako to traže. Na primjer, chatbot može pitati korisnika želi li određenu vrstu hrane. Zatim korisnik unosi vrstu hrane ili traži prijedlog. U oba slučaja podaci u prvom unosu tj. vrsta restorana spremaju se kao kontekst koji će služiti za donošenje budućih odluka.

Zadnja kategorija je chatbot s interakcijom agenata uživo. Live chat je vrsta sustava za chat koji se nalazi na web stranici ili mobilnoj aplikaciji na kojoj korisnici mogu komunicirati s timom za podršku i kontakt centru. Koristeći ovaj mehanizam, chatbotovi uključuju mogućnosti usmjeravanja za dodjelu rasprava u stvarnom vremenu. Kada korisnik treba komunicirati s predstavnikom nekog tima, chatbot skenira dostupnost agenta i usmjerava zahtjev za raspravu u skladu s tim. Povezat će kupca s nekim tko im može pomoći s njihovim problemom – tj. agentom s pravim vještinama i znanjem. Chatbot također upozorava agenta kada postoji upit kupca i obavještava kupca o detaljima agenta kao što su njegovo ime, vrijeme čekanja itd.

Modul predložen ovim radom je jedna vrsta hibrida chatbota koji donosi odluke vođene umjetnom inteligencijom i onog temeljenog na pravilima. Korišteni okviri u sklopu ovog modula su TensorFlow, PyTorch te razne HuggingFace biblioteke.

Za olakšanu izradu konverzacijskih sučelja gdje nije potrebno poznavati arhitekturu korištenih modela umjetne inteligencije postoje razni komercijalni okviri, koji najčešće niti ne zahtijevaju pisanje programskog koda, kao što su: Microsoft Bot Framework, Wit.ai, Dialogflow, IBM Watson, Pandorabots, RASA itd.

Ako se izrađuje konverzacijsko sučelje koje nije bazirano na jednom od komercijalnih rješenja, to se sučelje može razviti koristeći okvire u sklopu nekog programskog jezika.

3.1. Komercijalni okviri za izradu konverzacijskih sučelja

Microsoft Bot Framework jedan je od komercijalnih sveobuhvatnih okvira za izgradnju konverzacijskih AI sučelja. Zbirka su biblioteka, alata i usluga koje omogućuju izgradnju, testiranje, implementaciju i upravljanje inteligentnim botovima. Bot Framework uključuje modularni i proširivi SDK (komplet za razvoj softvera) za izradu botova i povezivanje s AI uslugama. Pomoću ovog okvira programeri mogu stvoriti sučelja koji koriste govor, razumiju prirodni jezik, odgovaraju na pitanja i još mnogo toga. Ovim se okvirom sučelja mogu razvijati u C#, JavaScript, Python i Java programskim jezicima. Microsoft Bot Framework pruža CLI (sučelje naredbenog retka) alate za pomoć pri razvoju chatbot-a od početne faze do krajnje faze postavljanja na poslužitelja spremnog za produkcijske svrhe. Ovaj okvir nudi i dodatnu integraciju s Azure Bot Service za usluge hostinga, Azure resursa za upravljanje i konfiguraciju botova i slično.

Wit.ai je web platforma za izdvajanje strukturiranih podataka iz nestrukturiranih upita korisnika. Ova platforma ne pruža komplet za izradu konverzacijskih sučelja, već samo dio dohvaćanja podataka s kojima se može upravljati. Iako nije samostalna platforma, u integraciji s primjerice Facebook Developer Portalom, može se u potpunosti razviti konverzacijsko sučelje za društvenu mrežu Facebook. Namijenjena je prvobitno za platforme u svrhu razmjena poruka. Pruža mogućnosti interakcije tekstom i glasom te upravljanje pametnim zvučnicima, uređajima, svijetlima i ostalo. Također pruža podršku za nosive uređaje kao što su Smart Watch.

Dialogflow je platforma za razumijevanje prirodnog jezika koja olakšava dizajn i integraciju konverzacijskog korisničkog sučelja u mobilnu aplikaciju, web aplikaciju, uređaj, interaktivni sustav glasovnog odgovora i slično. Dialogflow može analizirati više vrsta unosa korisnika, uključujući tekstualne ili audio unose (primjerice s telefona ili glasovne snimke). Također može odgovoriti korisnicima na nekoliko načina, putem teksta ili sintetičkog govora. Dialogflow pruža usluge virtualnog agenta za jednostavna i komplicirana sučelja. Obije usluge pružaju usluge virtualnog agenta za chatbotove i kontakt centre. Kontakt centar koji zapošljava ljudske agente, može koristiti Agent Assist da pomognete svojim ljudskim agentima. Agent Assist daje prijedloge u stvarnom vremenu za ljudske agente dok su u razgovoru s krajnjim korisnicima.

IBM Watson pruža uslugu Watson Assistent koja služi kao chatbot koji se integrira u CRM sustave kao što su Salesforce, Cisco, Avaya i drugi. Vrsta chatbota koju pruža Watson Assistent je hibrid chatbota s interakcijom agenata uživo te chatbot koji donosi odluke vođene umjetnom

inteligencijom. Ovo okruženje nudi tekstualnu i glasovnu interakciju sa sučeljem. Cijela procedura izrade chatbota se odvija online preko web sučelja, dakle bez potrebe za programiranjem. Chatbot može biti integriran kao SMS, WhatsApp, Facebook Messanger i Slack chatbot.

Pandorabots je okruženje za izradu chatbota preko online web sučelja. Višejezičan je, pruža izradu chatbota bez pisanja koda ili uz pisanje koda, ovisno o preferencijama. Pandorabots se može integrirati na razne platforme za razmjenu poruka, na vlastitu web stranicu ili mobilnu aplikaciju. Zanimljiva mogućnost je integracija s vanjskim servisima kao što su prognoze ili tražilice primjerice Wolfram Alpha. Kroz ovaj okvir, konverzacijsko sučelje može učiti i unaprjeđivati se dok razgovara s korisnicima.

RASA je open-source chatbot okvir koji se temelji na strojnom učenju. Integracija s vlastitim web mjestom, telegramom, Facebookom, WhatsAppom i ostalima je dostupna kao i potpora za integracijom na razne jezike. Uglavnom nije potrebno posjedovati vještine programiranja. Naglasak RASA daje na sigurnosti podataka, pošto je cijela arhitektura na vlastitoj infrastrukturi te klijenti imaju potpunu kontrolu nad njihovim modelima.

4. Korišteni modeli

Kako modul predložen ovim radom ne rješava trivijalni zadatak, modul ima takvu arhitekturu da koristi više različitih modela koji zajedno djeluju kao cjelina koja ostvaruje željene rezultate.

Prvi korišteni model u cilju zadatka prepoznavanja korisničke namjere jest LaBSE, jezično-agnostički BERT model ugrađivanja rečenice, koji ima sposobnost za maskiranu riječ u rečenici ponuditi najbolje sugestije o kojoj se riječi radi.

Drugi korišteni model u sklopu modula za ekstrakciju podataka je XLM-RoBERTa model koji je unaprijed treniran model nad 2,5 TB filtriranim podatcima CommonCrawl-a koji sadrže 100 različitih jezika. XLM-RoBERTa je višejezična verzija RoBERTa, transformatorski model unaprijed obučen na velikom korpusu engleskih podataka na samonadgledani način. To znači da je prethodno obučen samo na neobrađenim tekstovima, bez ljudi koji bi ih označavali na bilo koji način (zbog čega može koristiti mnogo javno dostupnih podataka) s automatskim postupkom za generiranje unosa i oznaka iz tih tekstova [10]. Obučen je s ciljem modeliranja maskiranog jezika.

Posljednji model koji se koristi u sklopu ovog modula je all-MiniLM-L6-v2 model transformacije rečenica koji preslikava rečenice i odlomke u 384-dimenzionalni gusti vektorski prostor i može se koristiti za zadatke poput klasteriranja ili semantičkog pretraživanja [11].

4.1. Transformeri

Transformer je model dubokog učenja koji usvaja mehanizam samopažnje, različito ponderirajući značaj svakog dijela ulaznih podataka. Koristi se prvenstveno u područjima obrade prirodnog jezika (natural language processing) i računalnog vida (computer vision). Poput rekurentnih neuronskih mreža (Recurrent Neural Networks), transformeri su dizajnirani za obradu sekvencijalnih ulaznih podataka, kao što je prirodni jezik, s ciljem rješavanja zadataka kao što su prijevod i sažimanje teksta. Međutim, za razliku od RNN-ova, transformeri obrađuju cijeli ulaz odjednom. Mehanizam pažnje osigurava kontekst za bilo koju poziciju u ulaznom nizu. Na primjer, ako je ulazni podatak rečenica prirodnog jezika, transformator ne mora obrađivati jednu po jednu riječ. To omogućuje veću paralelizaciju od RNN-ova i stoga smanjuje vrijeme treniranja.

Transformeri su sve više model izbora za NLP probleme, zamjenjujući RNN modele kao što je „long short-term memory RNN“ (LSTM). Dodatna paralelizacija treniranja omogućuje treniranje na većim skupovima podataka. To je dovelo do razvoja unaprijed obučenih sustava kao što su BERT (Bidirectional Encoder Representations from Transformers) i GPT (Generative Pre-trained Transformer), koji su obučeni s velikim skupovima podataka jezika, kao što su Wikipedia Corpus i Common Crawl, i mogu biti fino podešeni za specifične zadatke.

Arhitektura transformera sastoji se od kodera i dekodera. Funkcija svakog sloja kodera je generiranje kodiranja koja sadrže informacije o tome koji su dijelovi ulaza relevantni jedni drugima. Svoja kodiranja prosljeđuje sljedećem sloju kodera kao ulaze. Svaki sloj dekodera radi suprotno, uzima sva kodiranja i koristi njihove ugrađene kontekstualne informacije za generiranje izlazne sekvence. Kako bi se to postiglo, svaki sloj kodera i dekodera koristi mehanizam pažnje. Za svaki ulaz, pažnja važe relevantnost svih ostalih ulaza i iz njih da proizvodi izlaz. Svaki sloj dekodera ima dodatni mehanizam pažnje koji crpi informacije iz izlaza prethodnih dekodera, prije nego sloj dekodera crpi informacije iz kodiranja [14].

4.2. LaBSE (Language-agnostic BERT Sentence Embedding)

LaBSE model je jezično neovisan BERT rečenični „embedding“ publiciran u arXiv arhivi akademskih članaka 2020 godine. Rad opisuje kako se išlo s ciljem istraživanja učinka korištenja prethodno treniranih modela dvostrukih kodera (pretrained dual-encoder model). Modeli dvostrukog kodera učinkovit su pristup za učenje međujezičnih ugrađivanja (eng. embedding). Takvi modeli sastoje se od uparenih modela kodiranja koji daju funkciju bodovanja (eng. scoring function). Izvorna i ciljna rečenica kodirane su zasebno. Ugradnje rečenica izdvajaju se iz svakog kodera. Početni bazni transformer se nasumično inicijalizira kako bi se model odmah u početku razlikovao od drugih.

Za pretreniravanje, koristi se kombinacija modeliranja maskiranog jezika (MLM) i modeliranja jezika prijevoda (TLM), a trenirani korpus su jednojezični i dvojezični podaci. Jednojezični podaci prikupljeni su sa CommonCrawl-a i Wikipedije (uz dodatne filtere za isključivanja šuma u tekstu to jest beznačajnog teksta i uklanjanje prevelikih ili premalih rečenica).

Dvojezični korpus sastavljen je od web stranica korištenjem sustava bitekstnog rudarenja. Izdvojeni rečenični parovi filtrirani su unaprijed obučenim modelom bodovanja odabira kontrastivnih podataka (CDS). Ljudski anotatori ručno su procijenili parove rečenica iz malog podskupa prikupljenih parova i označavaju parove kao DOBRE ili LOŠE prijevode. Prag

modela bodovanja odabira podataka odabran je tako da je 80% zadržanih parova iz ručne evaluacije ocijenjeno kao DOBRO. Dodatno ograničenje odnosi se na maksimalan broj parova rečenica na 100 milijuna za svaki jezik kako bi se uravnotežila distribucija podataka. Mnogi jezici još uvijek imaju puno manje od 100 milijuna rečenica. Konačni korpus sadrži 6 milijardi parova prijevoda. Ovaj dvojezični korpus korišten je i za obuku dvostrukog kodera i za prilagođeno prethodno treniranje (customized pre-training).

4.3. XLM-RoBERTa (Cross-Lingual Language Model, Robustly Optimized BERT Pretraining Approach Model)

Cilj rada u kojem je objavljen ovaj model je poboljšati međujezično razumijevanje jezika (XLU), pomnim proučavanjem učinaka treniranja nenadziranih međujezičnih reprezentacija na vrlo velikoj razini. Radi se o repliciranju postojećeg pristupa XLM (Cross-Language Masked) jezičnog modela, na većem korpusu primjenjujući ga i na jezike s limitirajućim resursima. Razlike od postojećeg pristupa su na onim dijelova koji poboljšavaju performanse modela. Ovaj se model trenirao nad 100 različitih jezika, je treniran nad podacima Wikipedije i CommonCrawl-a.

Jezici s manjak resursa beneficiraju od modela koji inkorporira više jezika do određene točke, nakon čega cjelokupna performansa opada. Rad također pokazuje kako uključivši korpus CommonCrawl-a model poveća performanse nad jezicima niske rezolucije (mali broj resursa). Slični pokazatelji ukazuju na benefite treniranja modela na više jezika jer djeluje efekt „prijenosa smetnje“ (eng. transfer-interference trade-off).

Za evaluaciju modela koristile su se razne tehnike kao što su međujezično zaključivanje prirodnog jezika, prepoznavanje imenovanog entiteta, odgovaranje na međujezična pitanja. Od setova podataka za mjerenje (eng. benchmark datasets) su se koristili CoNLL-2002, CoNLL-2003, MLQA i GLUE.

4.4. all-MiniLM-L6-v2 (All Mini Language Model, 6 Layers, version 2)

Ovaj jezični model je fino podešena verzija „nreimers/MiniLM-L6-H384-uncased“ modela s jednom milijardom parova rečenica. Nreimers' MiniLM-L6-H384-uncased model je verzija „microsoft/MiniLM-L12-H384-uncased“ modela u kojoj je uzet svaki drugi sloj kodera (BERT encoder) tako da je od originalnih 12 slojeva ostalo samo 6 slojeva, što ujedno i smanjuje

veličinu modela. Microsoftov MiniLM-L12-H384-uncased model je uncased (ne razlikuje velika i mala slova), 12 slojni (BERT encoder), veličine skrivenih slojeva 384 Mini Language Model. Ključna ideja je duboko oponašanje modula samopažnje koji su temeljno važne komponente u modelima nastavnika i učenika koji se temelje na Transformeru.

Projekt all-MiniLM-L6-v2 ima za cilj trenirati modele ugrađivanja rečenica (sentence embedding models) na vrlo velikim skupovima podataka na razini rečenice koristeći samonadzirani kontrastivni cilj učenja (self-supervised contrastive learning objective). S obzirom na rečenicu iz para, model bi trebao predvidjeti koja je iz skupa nasumično odabranih drugih rečenica zapravo uparena s njom u našem skupu podataka. Ovaj model je razvijen tijekom tjedna zajednice koristeći JAX/Flax za NLP i CV (computer vision), koji je organizirao HuggingFace. Model se razvio kao dio projekta: „Uvježbajte najbolji model ugrađivanja rečenica ikada s jednom milijardom parova za uvježbavanje“. Iskorištena je učinkovita hardverska infrastruktura za izvođenje projekta: 7 TPU-ova (Tensor Processing Unit) verzija 3.8, s intervencijom Google Flaxa, JAX-a i člana tima za Cloud o učinkovitim okvirima dubinskog učenja.

5. Arhitektura rješenja

Arhitektura modula je linearna, a kreće s modulom za prepoznavanja namjere korisnika. Ako je korisnička namjera takve prirode da zahtjeva ekstrakciju podataka, slijedi modul za ekstrakciju podataka. Interakcija s modulom odvija se preko ugrađene konzole.

Modul ovoga rada temelji svoje odluke na prvenstveno JSON (JavaScript object notation) reprezentaciji linearnog procesa te na treniranim podatcima za sve spomenute module iz arhitekture.

5.1. Modul za prepoznavanja namjere korisnika

U svrhu ovog rada koristio se postojeći fino podešen modul za prepoznavanja namjere korisnika (Ferlatti, 2021). Radi se o jezično-agnostičkom BERT modelu ugrađivanja rečenice koji na temelju treniranih podataka pruža mogućnost kategoriziranja tekstualnog upita u jednu od domena pitanja (6 domena u svrhu modula ovoga rada).

Modul se sastoji od dohvaćanja podataka, pretprocesiranja ulaznog teksta i treniranja LaBSE modela. Pretprocesiranje podataka izvodi se sljedećim redoslijedom:

1. Uklanjanje interpunkcijskih znakova
2. Pretvorba teksta u tekst malih slova
3. Primjena tokenizacije
4. Uklanjanje zaustavnih riječi
5. Primjena lematizatora

5.2. Modul za ekstrakciju podataka

Sljedeća je linearna arhitektura modula za ekstrakciju podataka: dvostruko fino podešen model za zadatke ispunjavanja maske i model transformacije rečenica u svrhu zadatka rečenične sličnosti.

Modul ima sposobnost iz danog teksta zaključiti o kojem se zadatku procesa radi, a s tom informacijom se naknadno definira logika izvođenja cjelokupnog modula ovog rada.

Prvi korak k implementaciji ovog modula je fino podesiti model koji je namijenjen za zadatke ispunjavanja maske. Cilj takvog fino podešenog modela je izgraditi NER (Named Entity

Recognition) model. U svrhu prvog finog podešavanja modela koristi se WikiAnn hrvatski skup podataka za prepoznavanje imenovanih entiteta koji se sastoji od članaka na Wikipediji [12].

Nakon dobivenog NER modela isti se ponovno fino podešava kako bi prepoznao pojedine nazive zadataka. Cilj ovog pristupa je naučiti model da svrsta nazive treniranih zadataka pod entitet „organizacije“.

Kao zadnji korak u implementaciji, izdvojeni nazivi zadataka se uspoređuju sa svim dostupnim nazivima zadataka iz JSON-a kako bi se utvrdili kojem nazivu najviše sliči onaj na ulazu modela. Za ovaj se korak koristi metoda sličnosti kosinusa koja se bazira na „all-MiniLM-L6-v2“ rečeničnom transformatoru koji preslikava rečenice u 384 dimenzionalni gusti vektorski prostor.

6. Postupak treniranja

Modul se trenira u tri faze: treniranje modula za prepoznavanja namjere korisnika, treniranje modela koji je namijenjen za zadatke ispunjavanja maske te treniranje modela za prepoznavanje imenovanih entiteta.

Treniranje modula za prepoznavanje namjere korisnika se odvija u 10 epoha, a podatci za treniranje su spremljeni u Excel tablici. Pod stupcem „Text“ se nalaze ulazni tekstualni podatci, pod stupcem „Question“ se nalaze podatci o oznaci domene pitanja, a pod stupcem „Process“ se nalaze podatci o kojem se procesu radi. Podatci se prethodno filtriraju po željenom nazivu procesa te se primjenjuju tehnike pretprocesiranja nad ulaznim tekstom. Poželjno je imati što više raznih trening podataka za svaku od domena pitanja

U svrhu finog podešavanja modela koji je namijenjen za zadatke ispunjavanja maske koristi se WikiAnn hrvatski skup podataka za prepoznavanje imenovanih entiteta. Treniranje se izvodi u 2 epohe s prvih 1000 trening podataka iz skupa podataka. Limitirajuća brojka na trening podatke je razlog dugačkom vremenu treniranja za nesignifikantno povećan stupanj preciznosti.

Modela za prepoznavanje imenovanih entiteta mora prepoznati zadatke JSON procesa, radilo se o spominjanju naziva zadatka po njegovom nazivu ili njegovim sličnim izrazima (koji moraju biti spremljeni u JSON objektu). Za takvo fino podešavanje modela za prepoznavanje imenovanih entiteta treniranje se izvodi se nad Excel podatcima gdje se pod stupac „ner_tags“ nalaze očekivane oznake imenovanih entiteta (0 – ostalo, 1 – početak osobe, 2 – osoba, 3 – početak organizacije, 4 – organizacija, 5 – početak lokacije, 6 – lokacija) za svaku riječ iz stupca „tokens“. Stupac „tokens“ sadrži tekstualne podatke, stupac „spans“ sadrži podatke o očekivanim izdvojenim entitetima te stupac „process“ sadrži informaciju o kojem je procesu riječ. Kod ovog finog podešavanja, bitno je pokriti svaki definirani naziv nekog zadatka koji se nalazi u JSON objektu pod nazivom „alias“. Ovaj se model trenira u 10 epoha.

7. Rezultati

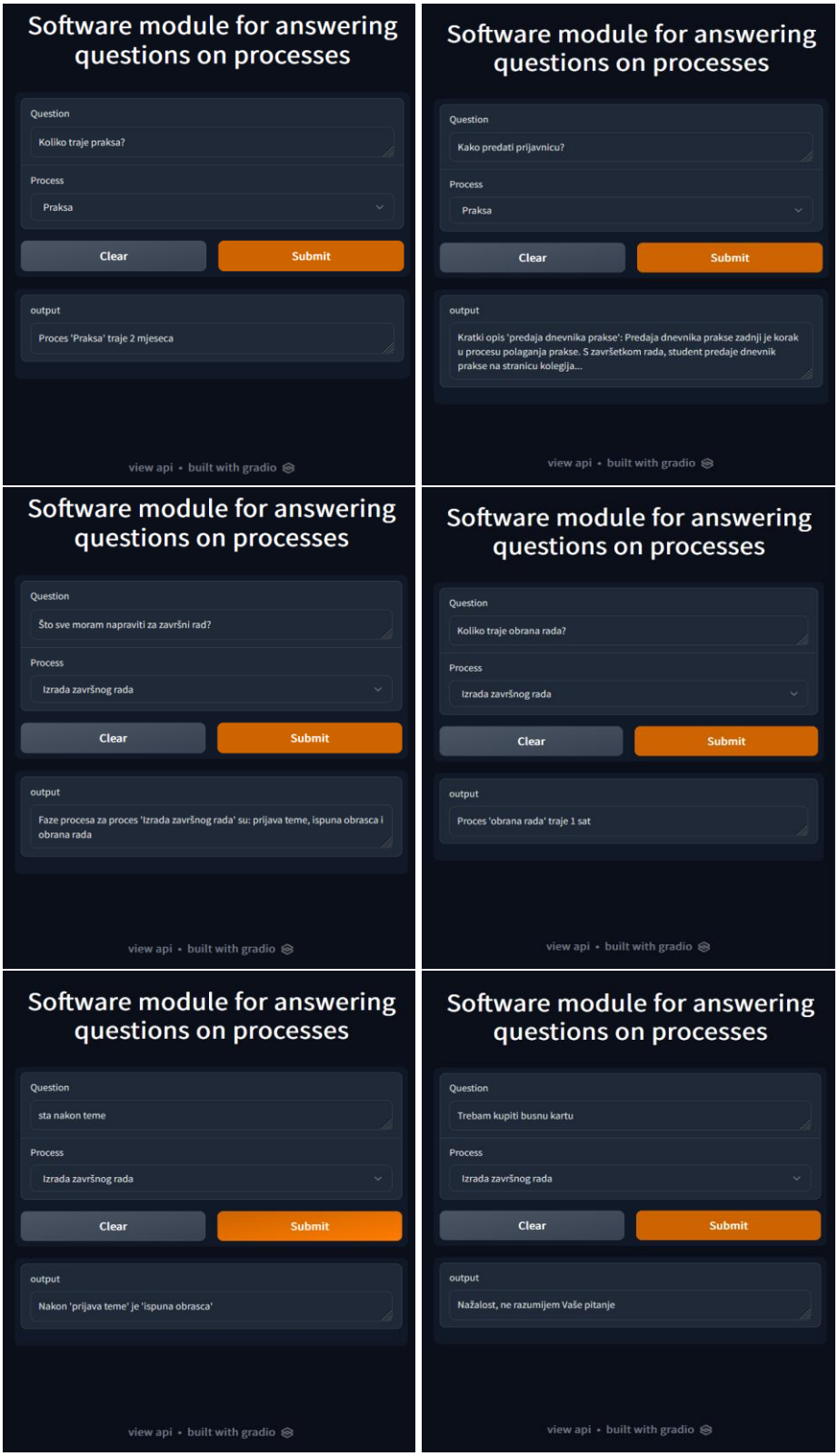
Sa završenim treniranjem modul za prepoznavanja korisničke namjere je postigao točnost od 66%, modul za prepoznavanje imenovanih entiteta točnost od 89%, a modul za prepoznavanje naziva zadataka 86%.

Cijeli postupak pokretanja koda i treniranja na Google-ovoj infrastrukturi visokih performansi za računalstvo u oblaku izvršava u 5 minuta i 30 sekundi ako se namjesti hardverski akcelerator na GPU. Kad se radi sa CPU akceleratorom onda vrijeme izvođenja traje 35 minuta. Prvi model za prepoznavanje korisničke namjere se trenira 15-ak sekundi na GPU, a 3 minute na CPU. Model za prepoznavanje imenovanih entiteta se trenira 40-ak sekundi na GPU, a 23 minute na CPU. Model za prepoznavanje naziva zadataka se trenira 1 sekundu na GPU, a 1 minutu na CPU.

Pokretanjem funkcije chatbot() koja kao parametar prima naziv procesa u JSON objektu, pokreće se sučelje koje je spremno za odgovarati na pitanja o procesima. Na slikama ispod može se vidjeti jedan primjer konverzacije s chatbotom.

Prvo postavljeno pitanje je „Koliko traje praksa?“ na što modul odgovara sa „Proces 'praksa' traje 2 mjeseca“. U JSON objektu se nalazi informacija o trajanju procesa prakse. Modul prepoznaje da se radi o pitanju P2 koji označava pitanje trajanja samog procesa. Na temelju te informacije ispisuje se unaprijed određen odgovor o trajanju procesa. Drugo postavljeno pitanje je „Kako predati prijavnicu?“ na što modul odgovara s opisom „Kratki opis 'predaja…“. U JSON objektu se nalazi informacija o opisu svakog zadatka. Modul prepoznaje da se radi o pitanju P3 koji označava pitanje opisa zadatka iz procesa. Na temelju te informacije ispisuje se unaprijed određen odgovor o opisu zadatka.

Ova se logika primjenjuje na svako postavljeno pitanje sa slika.



Slika 2. Sučelje aplikacije nakon faze treniranja modula

8. Zaključak

Modeliranje poslovnih procesa je od velike važnosti pošto opisuje kako se izvode poslovni procesi unutar nekog poslovanja. Integracija konverzacijskih sučelja za odgovaranje na pitanja o procesima omogućuje korisnicima da lakše dođu do informacija koje se tiču određenog poslovnog procesa. Prednost konverzacijskih sučelja u odnosu na metode agenta uživo i kontakta putem elektroničke pošte je ta što su konverzacijska sučelja konstantno aktivna i mogu vršiti interakciju bez potrebe ljudskog nadzora. Nedostatak ovog pristupa je što je teško razviti sučelje s visokom kvalitetom prepoznavanja pitanja i donošenja pravih odgovora.

Ovom se dokumentacijom opisala arhitektura te način rada izrađenog programskog modula za odgovaranje na pitanja o procesima. Koristeći postojeći modul za prepoznavanje domene pitanja korisnika, model za zadatke predviđanja maskiranog izraza te rečenični transformator za kvantifikaciju sličnosti rečenica uspješno se izradio modul za odgovaranje na pitanja o procesima. Modul je trenutno namijenjen da odgovara samo na pitanja o linearnim procesima, to jest najjednostavnijim tipovima procesa u kojem zadatci procesa teku slijedno jedan za drugim.

Ograničenje trenutnog modula je što korisnik mora pitati pitanje na sličnu shemu treniranih podataka, što znači da ako korisnik postavi kompliciraniji upit velike su šanse da će modul krivo odgovoriti ili ne razumjeti postavljen upit. Drugo je ograničenje što korisnik mora znati o kojem procesu želi dobiti odgovor. Naizgled ne liči na ograničenje, no ako se radi o većem procesu koji sadrži manje potprocese (koji su svi linearne prirode), onda dolazi do problema pošto se može dogoditi da korisnik ne zna o kojem je potprocesu pitanje vezano. Sljedeće je ograničenje to što modul ne može generalizirati upite na bilo koji proces, već je potrebno trenirati modul za svaki očekivani proces. Posljednje je ograničenje same domene procesa, to jest linearnosti procesa. Ako se radi o procesu koji ima grananje, petlje i slične elemente, ovaj modul ne može interpretirati takav proces.

Ovaj je modul moguće unaprijediti tako da se koristeći isti model kao i onaj za raspoznavanje domene pitanja donese zaključak o kojem se procesu radi. Temeljeno na samom upitu korisnika pozadinska logika dohvaćanja domene pitanja i domene procesa je vrlo slična. Taj bi napredak uklonio drugo ograničenje navedeno u prethodnom odlomku. Nadalje, koristeći sofisticiranije metode razumijevanja korisničkog pitanja moguće je unaprijediti modul kako bi mogao odgovarati na kompleksnije upite s većom preciznošću. Posljednje, kako modul trenutno daje

unaprijed formulirane odgovore, koristeći neki od generatora teksta (primjerice GPT-2), moguće je poboljšati kvalitetu dobivenih odgovora to jest učiniti odgovore prirodnijim i bližim svakodnevnoj ljudskoj komunikaciji.

9. Literatura

[1] Aguilar-Saven, R. S. (2004). Business process modelling: Review and framework. *International Journal of production economics*, *90*(2), 129-149. https://www.sciencedirect.com/science/article/abs/pii/S0925527303001026

[2] Burton-Payne, L. (n.d.). *Why Business Process is important?* •. Checkify. Retrieved September 4, 2022, from https://checkify.com/blog/why-business-process-is-important/

[3] White, S. A. (2004). Introduction to BPMN. *Ibm Cooperation*, *2*(0), 0. http://yoann.nogues.free.fr/IMG/pdf/07-04_WP_Intro_to_BPMN_-_White-2.pdf

[4] Adamopoulou, E., & Moussiades, L. (2020). An Overview of Chatbot Technology. *IFIP Advances in Information and Communication Technology*, 373–383. https://doi.org/10.1007/978-3-030-49186-4_31

[5] IEEExplore Digital Library. (2010). *Choice Reviews Online*, *47*(11), 47–6268. https://doi.org/10.5860/choice.47-6268

[6] Haristiani, N. (2019, November). Artificial Intelligence (AI) Chatbot as Language Learning Medium: An inquiry. *Journal of Physics: Conference Series*, 012020. https://doi.org/10.1088/1742-6596/1387/1/012020

[7] Wikipedia contributors. (n.d.). *Natural language processing*. Wikipedia. Retrieved September 4, 2022, from https://en.wikipedia.org/wiki/Natural_language_processing

[8] Brandtzaeg, P. B., & Følstad, A. (2017). Why People Use Chatbots. *Internet Science*, 377–392. https://doi.org/10.1007/978-3-319-70284-1_30

[9] Pérez, J. Q., Daradoumis, T., & Puig, J. M. M. (2020). Rediscovering the use of chatbots in education: A systematic literature review. *Computer Applications in Engineering Education*, *28*(6), 1549–1565. https://doi.org/10.1002/cae.22326

[10] x*lm-roberta-base · Hugging Face*. (n.d.). Xlm-Roberta-Base. Retrieved September 4, 2022, from https://huggingface.co/xlm-roberta-base

[11] *sentence-transformers/all-MiniLM-L6-v2 · Hugging Face*. (n.d.). All-MiniLM-L6-V2. Retrieved September 4, 2022, from https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[12] *wikiann · Datasets at Hugging Face*. (n.d.). WikiAnn. Retrieved September 4, 2022, from https://huggingface.co/datasets/wikiann

[13] Camunda. (2022, May 3). *Real-world BPMN 2.0 examples and answers to common questions*. Retrieved September 15, 2022, from https://camunda.com/bpmn/examples/

[14] Wikipedia contributors. (2022, September 12). *Transformer (machine learning model)*. Wikipedia. Retrieved September 18, 2022, from https://en.wikipedia.org/wiki/Transformer_(machine_learning_model)