ALMA MATER EUROPAEA
EVROPSKI CENTER, MARIBOR
Arhivske znanosti

DOKTORSKA DISERTACIJA

Miroslav Milovanović

ALMA MATER EUROPAEA
Evropski center, Maribor

Doktorska disertacija
študijskega programa tretje bolonjske stopnje

ARHIVSKE ZNANOSTI

MODEL UREJANJA IN POPISOVANJA NESTRUKTURIRANIH BESEDIL Z UPORABO STROJNEGA UČENJA

Mentor: izr. prof. dr. Miroslav Novak

Kandidat: Miroslav Milovanović

Maribor, oktober 2024

ZAHVALA

Zahvaljujem se svojemu mentorju izr. prof. dr. Miroslavu Novaku za pripravljenost za sodelovanje in diskusije ter pomoč pri raziskavi in pisanju doktorske disertacije. Za strokovna predavanja in nasvete se zahvaljujem vsem predavateljem v okviru študijskega programa arhivske znanosti.

Zahvaljujem se družini, ki me je in me podpira na moji življenjski poti.

Posebna zahvala gre moji partnerki in sinu, ki sta mi ves čas študija in ob času pisanja doktorske disertacije stala ob strani, me motivirala in spodbujala, da sem dosegel zastavljene cilje.

POVZETEK

Namen: Namen doktorske disertacije je raziskati, ali je možna izdelava modela za urejanje in popisovanje nestrukturiranih besedil z uporabo strojnega učenja. Pri izdelavi modela je bila raziskava razdeljena na tri ključne segmente in povezana raziskovalna vprašanja, in sicer, ali je izdelava modela za samostojno klasifikacijo nestrukturiranih vsebin, samostojno prepoznavo imenskih entitet in samostojno izdelavo naslova popisne enote izvedljiva in uporabna.

Metodologija: V raziskavi sta uporabljeni metoda analize vsebine in metoda eksperimenta. Raziskani so bili različni pristopi za izdelavo izvedbenega modela za urejanje in popisovanje nestrukturiranih besedil, ravno tako je bilo raziskana uporabnost izdelanega modela in učinkovitost izdelave popisne enote z uporabo izdelanega modela.

Rezultati: Ugotovljeni rezultati raziskav kažejo, da je izdelava modela za urejanje in popisovanje nestrukturiranih besedil z uporabo strojnega učenja za vse tri segmente izvedljiva in uporabna, izdelani model pa predstavlja celoten formalni in aplikativni okvir za obdelavo nestrukturiranih besedil, ki se ga lahko neposredno uporabi za obdelavo nestrukturiranih podatkov.

Izvirnost/uporabnost: Raziskava omogoča natančen vpogled v izdelavo modela za urejanje in popisovanje nestrukturiranih besedil ter izpostavlja prednosti in obliko uporabe izdelanega modela. Hkrati izdelani model in spremna dokumentiranost izdelave modela predstavljata podlago za uporabo modela v praksi in potencialno podlago za nadaljnje raziskave.

Ključne besede: urejanje in popisovanje, strojno učenje, imenske entitete, nestrukturirano besedilo, klasifikacija.

ABSTRACT

Purpose: The purpose of the doctoral dissertation is to investigate whether it is possible to create a model for arrangement and archival description of unstructured texts using machine learning. When creating the model, the research was divided into three key segments and related research questions, namely whether the creation of a model for the standalone classification of unstructured texts, the standalone recognition of name entities and the standalone creation of the title of the unit of description is feasible and useful.

Methodology: Methods content analysis and experiment are used in the research. Different approaches for creating an implementation model for arrangement and archival description of unstructured texts were investigated, as well as the usability of the developed model and the efficiency of creating a unit of description using the developed model.

Results: The results of the research indicate that the creation of a model for arrangement and archival description of unstructured texts using machine learning is feasible and useful for all three segments, and the created model represents the entire formal and application framework for the processing of unstructured texts, which can be directly used for processing unstructured data.

Originality/Usability: The research provides a detailed insight into the creation of a model for arrangement and archival description of unstructured texts and highlights the advantages and use cases of the created model. At the same time, the created model and the accompanying documentation of the model creation represent the basis for the use of the model in practice and a potential basis for further research.

Keywords: Arrangement and description, machine learning, named entities, unstructured text, classification.

KAZALO

SEZNAM TABEL

SEZNAM SLIK

1 UVOD

Temeljni problem z velikimi količinami gradiva v digitalni obliki, ki ima neznano vsebino in nestrukturirano obliko, je v tem, da je lahko pri izjemno velikih količinah takšnega gradiva težko zagotoviti hitro, učinkovito in ustrezno vsebinsko obravnavo za namen klasificiranja gradiva oziroma za namen urejanja in popisovanja gradiva v digitalni obliki.

Glede na to, da se v primeru nestrukturiranega in vsebinsko neznanega gradiva lahko nahaja veliko zapisov, ki jih potencialno ne želimo hraniti oziroma ne predstavljajo takšne vrednosti, da bi bila potrebna ohranitev zapisov, je treba razviti pristop identifikacije gradiva. Ta bo gradivo ločil na tisto, ki ga želimo ohraniti, in tisto, ki ga ne želimo hraniti (vsebinska klasifikacija in vrednotenje), ter omogočil ustrezno dostopnost in uporabnost gradiva, ki ga želimo ohraniti.

V raziskavi za samostojno klasifikacijo elektronske pošte z uporabo strojnega učenja (Milovanović 2020) so bile pridobljene dodatne informacije, ki ravno tako osnovno izkazujejo, da lahko pri vsebinski interpretaciji tovrstnega dokumentarnega gradiva prihaja do različne interpretacije (npr. v primeru spletnih strani oziroma spletnega gradiva ali npr. pri obravnavi elektronske pošte) glede na vsebino ali njegovo pojavno obliko. Slednje lahko povzroči tudi različno obravnavo enakih vrst gradiva.

Iz problema, opredeljenega v prvem odstavku tega poglavja, izhaja, da je v praksi treba najti način, kako preseči omejitve, ki izhajajo iz predstavljenega problema. Okvirno bi pri urejanju in popisovanju gradiva z uporabo strojnega učenja lahko opredelili, da gre za tri temeljne izzive oziroma podprobleme:

- izziv št. 1 – vsebinska klasifikacija individualnega zapisa,
- izziv št. 2 – zajem ključnih informacij (imenskih entitet) iz individualnega zapisa,
- izziv št. 3 – zajem informacij za izdelavo naslova popisne enote individualnega zapisa.

Namen izziva št. 1 je predvsem ugotoviti, kako pri velikih količinah nestrukturiranih zapisov v digitalni obliki pridobiti informacijo o osnovni vsebinski opredelitvi oziroma temi vsebine, ki bi lahko predstavljala podlago za klasificiranje takšnih zapisov, posledično tudi podlago za morebitno vrednotenje gradiva.

Pri izzivu št. 2 je namen uporabiti postopek in vzpostaviti način, kako pri vsakem individualnem zapisu v digitalni obliki pridobiti osnovne informacije, tj. identifikacija imenskih entitet, na podlagi katerih bi naknadno lažje uporabljali ohranjeno gradivo (točke dostopa, angl. Access points) (ISAD(G) 2000), kjer bi takšne informacije lahko predstavljale tudi podlago za kontekstualizacijo vsebine.

Namen izziva št. 3 pa je v vsebini nestrukturiranega zapisa pridobiti informacije in najti način, na osnovi katerega bi bilo možno izdelati naslov popisne enote (individualnega zapisa).

Za čim boljše razumevanje izvajanja postopkov masovnega urejanja in popisovanja gradiva, ki ima nestrukturirano obliko oziroma neznano vsebino, s pomočjo strojnega učenja je bilo treba izvesti vzorčno analizo takšnih zapisov in potrditi možnosti za uporabo strojnega učenja pri:

- zagotavljanju samostojne klasifikacije nestrukturiranega gradiva na podlagi vsebinske analize,
- identifikaciji in ekstrakciji imenskih entitet za potrebe popisovanja gradiva ter
- identifikacijo in povzemanje ključnih informacij za izdelavo naslova popisne enote (individualnega zapisa).

Za izvedbo raziskave sta bili uporabljeni kvalitativni metodi vsebinska analiza in eksperiment.

Strukturirani podatki so informacije, organizirane na način, ki je vnaprej določen. Primer strukturiranih podatkov so npr. podatki, ki so urejeni v tabelah z vrsticami in stolpci. Takšni podatki podatkov se običajno nahajajo v relacijski bazi podatkov, ki ima vnaprej predvideno strukturo in potencialne relacije med različnimi entitetami, zato je do strukturiranih podatkov pogosto lažje dostopati, jih upravljati ali analizirati. Nestrukturirani podatki nimajo vnaprej določene oblike, podatkovnega modela ali strukture. Tako v praksi predstavljajo katerokoli obliko zapisa, ki lahko vključuje kakršnokoli informacijo. Ker ta vrsta podatkov ni organizirana na vnaprej določen način, jih je težje obdelati in analizirati s tradicionalnimi metodami (Nestrukturirani podatki 2024).

Glede na to, da je nestrukturiranih podatkov vse več – po določenih projekcijah naj bi jih bilo do leta 2025 okoli 80 % (Burgener 2021), in glede na to, da količina teh podatkov izjemno hitro narašča (Burgener 2021), je urejanje in popisovanje nestrukturiranih besedil velik problem, s katerim se arhivisti soočajo vsakodnevno. Da bi lahko na področju arhivistike v navezavi z identificiranim temeljnim problemom učinkovito izvajali arhivska opravila, je treba pristopiti k integraciji tehnoloških konceptov in metod, ki bi takšno učinkovitost lahko zagotovile. Razvoj modela za urejanje in popisovanje nestrukturiranih besedil z uporabo strojnega učenja je eden od takšnih pristopov, ki bi omogočil učinkovito urejanje in popisovanje gradiva, s tem pa tudi postavil temelje za razvoj arhivske znanosti z uporabo umetne inteligence in strojnega učenja.

Nestrukturirane podatke lahko kategoriziramo kot besedila, slike, videoposnetke, zvok, spletne strani, spletne bloge in številne druge oblike. Med raznolikostjo podatkov se praktično povsod najde tudi nestrukturirano besedilo, ki lahko z vidika arhiviranja gradiva vsebuje pomembne informacije. Pri osnovnem konceptu razumevanja in obdelave nestrukturiranih besedila tako ne gre le za analizo uporabljenega jezika, temveč tudi za vključitev različnih kriterijev, kot sta kontekst in pomen vsebin, zato takšna obdelava

nestrukturiranih besedil predstavlja ključni izziv na področju dolgoročne hrambe in uporabnosti gradiva (Adnan idr. 2021).

Obdelava nestrukturiranih podatkov z neznano vsebino predstavlja izziv tudi za različna strokovna, etična ali pravna področja, predvsem v smeri ustrezne identifikacije vsebine, na podlagi katere je možno izvajati nadaljnje obdelave, kot so npr. identifikacija osebnih podatkov, avtorskih vsebin, intelektualne lastnine, sovražnega govora, nezakonitih vsebin itd. Z uporabo strojnega učenja tako lahko pristopimo k izdelavi različnih izvedbenih modelov za obdelavo podatkov, ki nam omogočajo lažje doseganje omenjenih obdelav podatkov.

Nekateri najpogostejši izzivi pri obdelavi nestrukturiranih podatkov so (Baig 2023; Nestrukturirani podatki 2024):

- pravočasna obdelava velikih količin podatkov. količine nestrukturiranih podatkov izjemno hitro rastejo, kar izpostavlja izziv, kako pravočasno zajeti in obdelati podatke, preden pride do kakršnihkoli potencialnih izgub;
- varnostna tveganja. V primeru obdelave zaupnih podatkov je zaščita nestrukturiranih podatkov lahko zapletena, saj se lahko te informacije hitro razširijo v različnih oblikah zapisov in na lokacijah za shranjevanje, vključno s potencialnimi težavami pri identifikaciji zaupnih vsebin;
- razpoložljivost in uporabnost nestrukturiranih podatkov. Ob hitrem razvoju informacijskih tehnologij in potencialno zaklenjenih oblik formatov ter oblik zapisa se lahko pokaže izziv berljivosti podatkov v meri, da so še uporabni za nadaljnjo obdelavo;
- zahteven postopek indeksiranja, tj. klasificiranja. Zaradi raznolikih oblik zapisov in neznanih vsebin sta indeksiranje in klasifikacija običajno zahtevna postopka, pri katerih se lahko pojavljajo napake, ki ključno vplivajo na kakovost pridobljenih rezultatov;
- potreba po specializiranih resursih in specializiranem znanju. Nestrukturirani podatki danes predstavljajo večino ustvarjenega gradiva, zato je predmet obdelave velika količina podatkov, za katere je treba z vidika računske zmogljivosti zagotoviti ustrezno zmogljivo strojno opremo, ki bo tolikšno količino podatkov lahko obdelala. Vzporedno z zagotavljanjem ustrezne strojne opreme pa je treba zagotoviti tudi ustrezno znanje oziroma kadrovske resurse, ki bodo lahko izdelali ustrezne rešitve za obdelavo nestrukturiranih podatkov, t. i. strokovnjake za obdelavo podatkov (angl. Data Scientists);
- visoki stroški za vzpostavitev sistema za obdelavo nestrukturiranih podatkov. Poleg zmogljive strojne opreme in zagotavljanja ustreznih človeških virov je treba upoštevati še strošek ostalih gradnikov sistema za obdelavo nestrukturiranih podatkov, kot so specializirana programska oprema, oprema za shranjevanje podatkov, ukrepi za zagotavljanje informacijske varnosti itd.;

- podpora obdelavi različnih oblik zapisov. Nestrukturirani podatki nimajo vnaprej določenih standardnih oblik zapisov oziroma posebnih formatov, kar lahko otežuje obdelavo podatkov za zagotavljanje možnosti obdelave različnih vrst formatov.

Iz identificiranih izzivov za obdelavo nestrukturiranih podatkov je tako razvidno, da je treba pristopiti k razvoju celovitega modela (teoretičnega kot tudi aplikativnega), ki bo omogočal lažjo obdelavo nestrukturiranih podatkov. Z vidika arhivistike je pomembna obdelava nestrukturiranih podatkov v vseh njihovih pojavnih oblikah, posebno pozornost pa glede na količino podatkov zahteva obdelava nestrukturiranih besedil.

V okviru predmetne raziskave so predstavljeni osnovni koncepti in tehnologije za izdelavo modela urejanja in popisovanja nestrukturiranih besedil z uporabo strojnega učenja ter gradniki, ki so bili uporabljeni pri izdelavi modela. Poleg konceptov so predstavljene tudi predhodne raziskave, še zlasti o delu rešitev, ki so bile uporabljene pri izdelavi modela urejanja in popisovanja nestrukturiranih besedil z uporabo strojnega učenja.

Izdelana raziskovalna naloga je vsebinsko in izvedbeno razdeljena na naslednja področja:

- uvodna opredelitev problema in opredelitev nestrukturiranih podatkov, ki vključuje tudi terminologijo izrazov, uporabljenih v raziskavi;
- predstavitev področij vrednotenja, urejanja in popisovanja gradiva kot elementarnih postopkov, ki so bili zasledovani pri izdelavi modela urejanja in popisovanja nestrukturiranih besedil z uporabo strojnega učenja;
- predstavitev standardov in dobrih praks pri popisovanju gradiva z namenom prikaza uporabljene standardizirane izhodne oblike popisne enote, ki je rezultat obdelave nestrukturiranih podatkov z uporabo modela urejanja in popisovanja nestrukturiranih besedil z uporabo strojnega učenja;
- predstavitev področja umetne inteligence in razlogov za uporabo ter ključnih vplivov uporabe umetne inteligence, na katere je treba biti pozoren pri obdelavi podatkov;
- predstavitev področja strojnega učenja kot del umetne inteligence in pristop, ki je bil uporabljen za izdelavo modela urejanja in popisovanja nestrukturiranih besedil z uporabo strojnega učenja;
- predstavitev predhodnih raziskav specifično v delu o tehnologijah in postopkih, ki so bili uporabljeni pri izdelavi modela urejanja in popisovanja nestrukturiranih besedil z uporabo strojnega učenja;
- predstavitev namena in ciljev raziskave ter raziskovalnih hipotez in raziskovalnih vprašanj;
- predstavitev metod in tehnik zbiranja podatkov, vključno z opisom vzorcev podatkov, ki so bili uporabljeni za izdelavo modela urejanja in popisovanja nestrukturiranih besedil z uporabo strojnega učenja;
- natančen opis obdelave podatkov ter vseh korakov, aktivnosti, delotokov in procesov, ki so bili uporabljeni za izdelavo modela urejanja in popisovanja nestrukturiranih besedil z uporabo strojnega učenja;

- predstavitev pridobljenih rezultatov uporabe izdelanega modela urejanja in popisovanja nestrukturiranih besedil z uporabo strojnega učenja;
- pregled in interpretacija pridobljenih rezultatov v kontekstu izpostavljene problematike, ki je bila podlaga za izvedbo raziskave, in predstavitev odgovorov na postavljene znanstvene hipoteze;
- predstavitev zaključkov in odgovorov na postavljena raziskovalna vprašanja.

Terminologija

»**Arhiv** je ustanova, ki opravlja javno službo pridobivanja, hrambe in posredovanja arhivskega gradiva« (Arhivski slovarček 2024).

»**Arhivski fond je** celota arhivskega gradiva enega ustvarjalca, ki je bilo zbrano oziroma je nastalo v času njegovega delovanja« (Arhivski slovarček 2024).

»**Arhivsko gradivo** je dokumentarno gradivo, ki ima trajen pomen za zgodovino, druge znanosti in kulturo ali trajen pomen za pravni interes pravnih in fizičnih oseb; arhivsko gradivo je kulturni spomenik« (Zakon o varstvu dokumentarnega in arhivskega gradiva ter arhivih 2006, 2. čl.; v nadaljevanju ZVDAGA, 2. čl.).

»**Dokumentarno gradivo** so vse vrste in oblike zapisov, ki so nastali ali bili prejeti pri poslovanju pravnih in fizičnih oseb« (ZVDAGA, 2. čl.).

»**Dolgoročna hramba gradiva** je hramba gradiva v digitalni obliki za daljše časovno obdobje in se nanaša na gradivo, katerega rok hrambe je več kot pet let« (ZVDAGA, 2. čl.).

»**Hramba gradiva** je tista hramba izvirnega ali zajetega dokumentarnega gradiva, ki izpolnjuje pogoje po tem zakonu in zagotavlja uporabnost ter avtentičnost vsebine hranjenega gradiva« (ZVDAGA, 2. čl.).

»**Izvedbeni aplikativni model**, izvedbeni model oziroma ekspertni sistem je namenska programska oprema – računalniški program, ki uporablja umetno inteligenco« (Balič 2004, 184).

»**Klasifikacija** je razvrstitev, razporeditev česa glede na enake ali podobne lastnosti« (Klasifikacija 2020).

»**Naslov** je beseda ali skupina znakov, ki določajo individualno popisno enoto« (ISAD(G) 2000).

»**Oblika zapisa** je vrsta zapisa, ki ga povezujejo določene fizične ali intelektualne skupne lastnosti« (ISAD(G) 2000).

»**Popisna enota** je vsebinsko zaključena enota ureditve arhivskega gradiva, ki je bila oblikovana pri ustvarjalcu arhivskega gradiva ali pozneje v arhivu v kontekstu popisovanja« (Uredba o varstvu dokumentarnega in arhivskega gradiva 2017).

»**Pripomoček za uporabo (arhivski pripomoček)** v najširšem pomenu obsega vse vrste popisov in pripomočkov, nastalih pri urejanju in popisovanju v arhivu (arhivski popis, inventar, vodnik), ali prvotnih pripomočkov, nastalih zaradi preverjanja gradiva in pregleda nad njim« (Arhivski slovarček 2024).

»**Zajem** je vsak uvoz metapodatkov o gradivu ali gradiva samega v strojno berljivi obliki v informacijski sistem za upravljanje z dokumenti ali v informacijski sistem za hrambo« (ZVDAGA, 2. čl.).

»**Zapis**. Skupek informacij v katerikoli obliki ali mediju, ustvarjen ali prejet in vzdrževan s strani določene organizacije« (ISAD(G) 2000).

2 UREJANJE IN POPISOVANJE GRADIVA Z UPORABO STROJNEGA UČENJA

2.1 Urejanje in vrednotenje gradiva

Urejanje kot tudi vrednotenje gradiva sodita med temeljne procese, ki se izvajajo v okviru arhivske službe oziroma na področju zagotavljanja varovanja dokumentarnega in arhivskega gradiva. Pri obdelavi večjih količin gradiva v elektronski obliki, kot je to v primeru obdelave besedil v nestrukturirani obliki, proces urejanja in vrednotenja predstavlja poseben izziv, ki terja posebne rešitve. Danes vse več gradiva nastaja izvirno v digitalni obliki ali pa je del procesa digitalizacije, skupna lastnost v obeh primerih pa je, da je treba zagotoviti ustrezno okolje za dolgoročno hrambo in uporabnost gradiva (Stančić 2018).

2.1.1 Urejanje gradiva

Kot navaja Uredba o varstvu dokumentarnega in arhivskega gradiva (2017) je »dokumentarno gradivo urejeno, če so posamezni ali združeni dokumenti (spis, zadeva, dosje) razvrščeni v skladu z načinom upravljanja dokumentov oziroma pisarniškega poslovanja ustvarjalca«. Uredba o varstvu dokumentarnega in arhivskega gradiva (2017) nadalje opredeljuje, da je »pri hrambi, urejanju in uporabi treba z dokumentarnim in arhivskim gradivom ravnati previdno, da se ne poškoduje ali uniči in se ohrani njegova izvirna pričevalnost«. Urejanje gradiva tako predstavlja pomemben del celotnega procesa ohranjanja gradiva tako v fizični kot elektronski obliki. Zgolj pojavna oblika gradiva ne vpliva na obveznosti, ki izhajajo iz zakonodaje, standardov in etičnih načel glede hrambe samega gradiva in dolgoročne uporabe hranjenega gradiva.

Pri urejanju arhivskega gradiva je treba slediti načelom urejanja arhivskega gradiva, ki v osnovi postavljajo nekaj jasnih pravil kako pristopiti k urejanju gradiva v primeru, ko ima lastnosti arhivskega gradiva. Uredba o varstvu dokumentarnega in arhivskega gradiva (2017) opredeljuje, da je treba:

- arhivsko gradivo urejati po strokovnih načelih provenience, tj. skladno z načelom izvora in prvotne ureditve;
- arhivsko gradivo obravnavati kot celoto glede na načelo provenience, če je nastalo med delovanjem enega ustvarjalca;
- arhivsko gradivo ohraniti po enakem sistemu ureditve, kot ga je imel ustvarjalec, skladno z načelom prvotne ureditve;
- arhivsko gradivo urediti »vsebinsko, geografsko, številčno, kronološko, abecedno, po vrstah gradiva oziroma na podlagi kombinacije navedenih načinov«, če prvotna ureditev ni ohranjena oziroma je ni mogoče obnoviti.

V praksi se velikokrat srečujemo z odločitvami o izvedbi postopka urejanja gradiva šele v točki, ko se skuša rešiti probleme, ki nastajajo zaradi kopičenja velikih količin gradiva oziroma zaradi regulatornih zahtev za zagotavljanje ustrezne hrambe. Omenjeno je redni

pokazatelj, da bi moral biti proces urejanja gradiv integralni proces, s katerim je treba upravljati že pred nastankom gradiva, med njim pa vse do njegove končne destinacije in hrambe. S sistematičnim pristopom k urejanju gradiva bi lahko prihranili veliko časa za izvedbo postopka urejanja gradiva, predvsem pa bi bila zgotovljena transparentnost pri obdelavi, s katero bi lahko dosegli lažjo izvedbo postopkov, ki sledijo urejanju gradiva.

2.1.2 Vrednotenje gradiva

Vrednotenje kot eden od temeljnih procesov za zagotavljanje dolgoročne uporabnosti in varnosti arhivskega gradiva predstavlja vhodno točko za ugotavljanje, katero gradivo je treba hraniti kot »dokumentarno gradivo, ki ima trajen pomen za zgodovino, druge znanosti in kulturo ali trajen pomen za pravni interes pravnih in fizičnih oseb« (Zakon o varstvu dokumentarnega in arhivskega gradiva ter arhivih 2006). Hajtnik in Babić (2018) izpostavljata, da vrednotenje gradiva že leta predstavlja eno od najpomembnejših nalog javnih arhivov po svetu, ki temelji na uporabi različnih meril glede na njihovo pravno, kulturno, upravno, znanstveno in zgodovinsko vrednost.

Pri vrednotenju besedil, ki so v elektronski nestrukturirani obliki, ostaja enak problem kot v primeru vrednotenja kateregakoli gradiva, katerega vsebina ni znana. Da bi lahko učinkovito izvedli postopek vrednotenja, je treba v prvi vrsti razumeti vsebino in oceniti njeno vrednost skozi postavljena merila za vrednotenje. Pojavna oblika na vrednotenje nima vpliva, zato tudi rešitve za vrednotenje v elektronski obliki, kjer vsebina ni znana, težko sledijo standardnim postopkom, kot bi to veljalo za vrednotenje gradiva v npr. fizični/papirni obliki.

Vrednotenje je v osnovi posledica predhodno določene selektivne politike »kaj hraniti«, da bi se izognili potencialni prekomerni hrambi oziroma t. i. kopičenju gradiva, ki ne izpolnjuje zadanih meril za trajno hrambo. Pri tem je vedno treba upoštevati, kaj bo skozi čas predstavljalo vrednost oziroma katero gradivo ima trajen pomen za zgodovino, druge znanosti in kulturo ali trajen pomen za pravni interes pravnih in fizičnih oseb (Zakon o varstvu dokumentarnega in arhivskega gradiva ter arhivih 2006). S tega vidika vrednotenje predstavlja precejšen izziv za osebe, ki izvajajo omenjeni proces, v primeru vrednotenja izjemno velikih količin nestrukturiranih besedil pa je izziv še toliko večji, saj količina nastalega gradiva in njegov prirast predstavljata izjemno veliko časovno obremenitev za katerokoli arhivsko službo v primeru, da niso na voljo sodobna informacijska orodja, ki bi takšno vrednotenje optimizirala.

V okviru slovenske nacionalne zakonodaje Zakon o varstvu dokumentarnega in arhivskega gradiva ter arhivih (2006) navaja naslednja merila za izvedbo vrednotenja:

- potrebe zgodovinopisja, drugih znanosti in kulture,
- potrebe za trajno pravno veljavo za doseganje pravic oseb,
- pomembnost vsebine gradiva,
- specifičnost dogodkov in pojavov v določenem času,
- specifičnost kraja ali območja,

- pomen javnopravne osebe,
- pomen avtorja,
- pomen gradiva z vidika kulturne raznolikosti,
- izvirnost dokumentov,
- izvirnost podatkov in informacij,
- reprezentativni izbor,
- notranje in zunanje značilnost gradiva in
- druga merila, ki jih določi pristojni arhiv.

Couture (2005) je podobna merila opisovala tudi v svoji raziskave, kjer je prišla do zaključka, da bi morali arhivisti zasledovati pet temeljnih principov pri izvajanju vrednotenja gradiva, in sicer:

- da zapisi predstavljajo pomembne informacije o delovanju družbe kot celote,
- da je vrednotenje izvedeno na način, da odraža ocenjeno vrednost zapisa v času izvajanja vrednotenja,
- da je zagotovljena in upoštevana povezava med vrednotenjem in ostalimi arhivskimi procesi,
- da je vzpostavljeno ravnovesje med administrativnimi cilji in cilji dediščine,
- da je vzpostavljeno ravnovesje med kontekstom ustvarjenih zapisov in uporabo zapisov.

Vrednotenje kot proces se lahko izvaja na različne načine. Z upoštevanjem načel varstva dokumentarnega in arhivskega gradiva in uporabo meril za izvedbo vrednotenja nestrukturiranih besedil v velikih količinah pa se pojavlja vprašanje, kje in na kakšen način naj bi se takšno vrednotenje izvedlo. Ali je ustvarjalec tisti, ki je odgovoren za vrednotenje takšnega gradiva, ali sta odgovornost in izvedba na strani javnih arhivov? S tem povezano se odpirajo dodatna vprašanja, in sicer, ali naj tisti, ki izvaja vrednotenje, poleg meril uporablja še prej določene okvire, standarde, smernice, modele idr., kako naj se izvede nadzor nad ustreznostjo izvedenega vrednotenja, katera oprema naj se uporabi in kateri modeli itd. Operativno je vsekakor smotrno, da bi bil pristop k vrednotenju nestrukturiranih besedil v elektronski obliki poenoten. Iz tega bi sledilo, da bi bila tudi nadzor in izvedba postopkov lažja ter bolj pregledna.

Problem pri obdelavi besedil v nestrukturiranih obliki je tudi njihov izvor oziroma izvirna oblika, saj v praksi lahko tovrstni nestrukturirani zapisi spreminjajo obliko zapisa, v določenih primerih tudi obseg, kar predstavlja svojevrsten izziv v primeru zagotavljanja verodostojnosti gradiva, ko izvirna oblika ni več na voljo (Novak 2015).

2.2 Popisovanje gradiva

»Popisovanje arhivskega gradiva sodi med temeljne naloge arhivskega strokovnega dela, saj je popisovalec posrednik med ustvarjalcem arhivskega gradiva in njegovim uporabnikom« (Čargo 2018). Da bi končni uporabniki arhivskega gradiva lažje uporabljali gradivo

predvsem v kontekstu iskanja, je proces popisovanja izjemno pomemben, saj se lahko zaradi neustrezno izvedenega postopka popisovanja potencialno izgubi tudi uporabna vrednost hranjenega gradiva, če uporabnik ne bo na učinkovit način mogel najti tistega, kar išče.

Metod in oblik popisovanja gradiv je več, vse od ročnega popisovanja, uporabe sistemov za upravljanje z zapisi (Popovici 2021) do uporabe namenskih arhivskih sistemov za popisovanje gradiva. Pri obdelavi velikih količin podatkov, zlasti pri obdelavi besedil v nestrukturirani obliki, je izhodiščno stališče, da je gradivo lahko razpršeno skozi uporabo množice različnih informacijskih sistemov in oblik hrambe velikokrat brez kakršnihkoli informacij, ki bi nam bile v pomoč pri izvajanju popisovanja gradiva. Določeno problematiko za popisovanje v elektronski obliki lahko predstavlja tudi gnezdenje vsebin ali veriženje vsebin v obliki niti ali zaključenih sklopov, ki lahko združujejo več individualno ločenih zapisov, ki bi morali biti ločeni in popisani glede na njihovo izvirno obliko, pri čemer je zaradi velikih količin gradiva zelo težko uporabiti standardne metode ročnega pregledovanja hranjenih zapisov in potencialno ločevanje na izvirne zapise, ki bi morali biti individualno popisani (Zajšek in Novak 2021).

2.2.1 Nivoji popisa

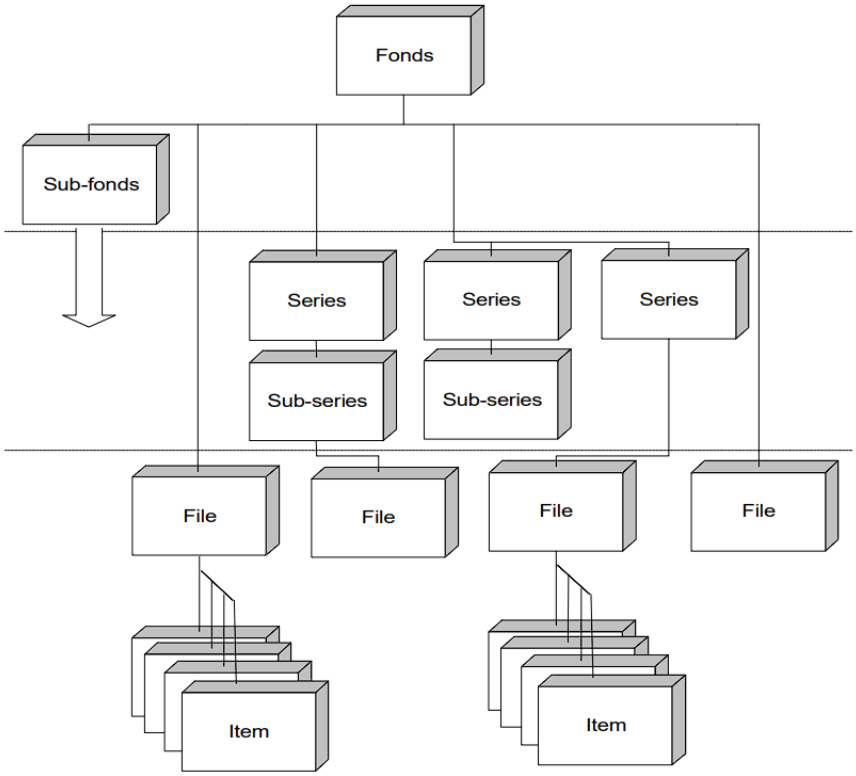
Uredba o varstvu dokumentarnega in arhivskega gradiva (2017) navaja, da se arhivsko gradivo popisuje »po nivojih, ki odražajo položaj popisne enote v strukturi fonda ali zbirke«. Nivoji popisa so (Uredba o varstvu dokumentarnega in arhivskega gradiva 2017):

- fond ali zbirka,
- podfond,
- serija,
- podserija,
- združeni dokument (na primer spis, zadeva, dosje) in
- dokument.

Uredba o varstvu dokumentarnega in arhivskega gradiva (2017) nadalje navaja, da mora vsaka popisna enota ne glede na nivo popisovanja »vsebovati naslednje obvezne elemente popisovanja: signaturo, naslov, čas nastanka popisne enote, nivo popisa, popisne enote na nivoju fonda ali zbirke, podfonda, pa tudi količino in obseg gradiva popisne enote«.

Iz slike 1 je razvidna hierarhična struktura nivojev popisovanja skladno s smernicami v standardu ISAD(G) (2000).

Slika 1: Nivoji popisa v hierarhični razdelitvi



Vir: ISAD(G) 2000.

2.2.2 Elementi popisa

»Elementi popisovanja arhivskega gradiva so podatki, ki identificirajo popisno enoto, navajajo njen izvor, vsebino in ureditev popisne enote, pogoje dostopnosti in uporabe ter povezave popisne enote« (Uredba o varstvu dokumentarnega in arhivskega gradiva 2017).

Težava pri obdelavi velikih količin nestrukturiranih besedil, kjer vsebina ni znana, je, da večina elementov za izvedbo celovitega popisovanja ni na voljo, zato je izjemno težko izvesti natančno ali celovito popisovanje posameznega izvirnega zapisa brez predhodne natančne analize celotne vsebine gradiva. Za določitev nivojev popisa gradiva je pomembno poznavanje vira besedil oziroma izvirnih zapisov, ki so v nestrukturirani obliki, ter hierarhije in povezljivosti zadnje samostojne popisne enote glede na predhodne nivoje v isti hierarhični strukturi.

Uredba o varstvu dokumentarnega in arhivskega gradiva (2017) navaja naslednji okvir elementov za izvedbo popisovanja:

- elementi identifikacije so:
  - signatura popisne enote,
  - prejšnje signature popisne enote,
  - klasifikacijska oznaka popisne enote,
  - številka tehnične enote,
  - naslov popisne enote,

    - čas nastanka arhivskega gradiva popisne enote,
    - zvrst arhivskega gradiva,
    - obseg popisne enote v tekočih metrih oziroma za digitalno gradivo v bajtih,
    - količina arhivskega gradiva popisne enote,
    - zunanje značilnosti arhivskega gradiva,
    - nivo popisa,
- elementi izvora so:
    - ime ustvarjalca,
    - obdobje obstoja ustvarjalca,
    - historiat ustvarjalca,
    - historiat fonda,
    - ime izročitelja arhivskega gradiva,
    - datum in številka prevzema arhivskega gradiva,
- elementi vsebine in ureditve so:
    - vsebina popisne enote arhivskega gradiva,
    - sistem ureditve arhivskega gradiva,
    - merila vrednotenja, odbiranja in izločanja,
- elementi dostopnosti in uporabe so:
    - pravni status arhivskega gradiva,
    - omejitve dostopnosti,
    - pravice intelektualne lastnine,
    - jezik in pisava arhivskega gradiva,
    - stanje ohranjenosti,
    - pripomočki za uporabo,
- elementi povezav so:
    - drugi imetniki delov arhivskega gradiva popisne enote,
    - kopije arhivskega gradiva ne glede na tehnologijo kopiranja,
    - vsebinsko sorodno arhivsko gradivo,
    - objave arhivskega gradiva.

2.2.3 Standardi na področju popisovanja gradiva

Standardov in smernic na področju popisovanja gradiv je več, odvisno od zastavljenih pristopov in namenov uporabe. Za dva poglavitna pristopa za popisovanje bi tako lahko šteli standard ISAD(G) – General International Standard Archival Description (ISAD(G) 2000) ter konceptualni model RIC – Records in Context – Conceptual model (RIC 2023).

ISAD(G) (2000) predstavlja temeljne smernice za izvedbo popisa arhivskega gradiva. Pri tem je treba opozoriti, da standard ne nadomešča potencialne nacionalne zakonodaje, ki ureja področje popisovanja arhivskega gradiva, ampak ga je treba uporabiti skupaj z nacionalnimi okviri. Standard poleg okvira za izvedbo popisa izpostavlja tudi določena pravila kako pristopiti k izvedbi popisovanja arhivskega gradiva. Vzpostavlja okvir za izvedbo popisa 26 elementov, ki predstavljajo celoto popisa za individualno popisno enoto. Standard tudi sledi temeljnim načelom s področja arhivistike. Poleg temeljnega standarda ISAD(G) so bili

pripravljeni tudi komplementarni standardi, ki pokrivajo specifična področja, in sicer ISDF – International standard for Describing Functions« (ISDF 2007), ISAAR (CPF) – International Standard Archival Authority Record for Corporate Bodies, Persons and Families (ISAR CPF 2004) in ISDIAH – International Standard for Describing Institutions with Archival Holdings (ISDIAH 2008).

Pravila ISAD(G) (2000) so organizirana v sedem področij za izvedbo popisa, in sicer:

- področje identifikacije, ki vključuje ključne informacije za identifikacijo enote opisa,
- področje konteksta, ki vključuje informacije o izvoru in ohranjenosti enote opisa,
- področje vsebine in strukture, ki vključuje informacije o predmetu in urejenosti enote opisa,
- področje pogojev dostopnosti in uporabe, ki vključuje informacije o dostopnosti enote popisa,
- področje povezanih virov, ki vključuje informacije o gradivu, ki je pomembno povezani z enoto opisa,
- področje opomb, ki vključuje posebne ali ostale informacije o enoti popisa, ki jih ni mogoče umestiti v nobeno drugo področje,
- področje kontrole opisa, ki vključuje informacije na kakšen način, kdaj in kdo je izvedel popis.

Podobno kot Uredba o varstvu dokumentarnega in arhivskega gradiva (2017) tudi standard ISAD(G) (2000) predvideva minimalen obvezen okvir elementov, ki bi moral biti zagotovljen pri izvedbi popisa individualne popisne enote. Ti elementi so:

- identifikacijska oznaka,
- naslov,
- ustvarjalec gradiva,
- čas nastanka gradiva,
- količina enot opisa,
- nivo opisa.

RIC – Records in Context – Conceptual model (v nadaljevanju RIC 2023) je konceptualni model visoke ravni, ki se osredotoča na intelektualno prepoznavanje in opisovanje zapisov in ljudi, ki so jih ustvarili in jih uporabljajo, ter dejavnosti, ki jih izvajajo ljudje, ki jih zapisi opredeljujejo in dokumentirajo. Ker gre za model visoke ravni, predstavlja širok izvedbeni okvir, ki pa ne vključuje smernic za vse aktivnosti, ki so potrebne za upravljanje z zapisi.

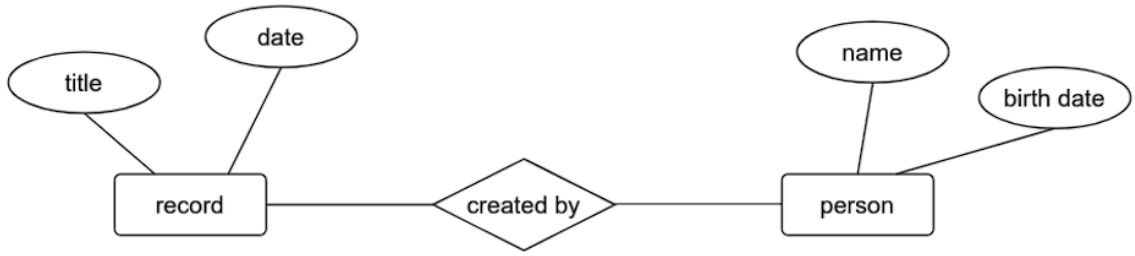
RIC (2023) pokriva celoten obseg ključnih vsebin standarda ISAD(G) (2000) in komplementarnih standardov ter s tem predstavlja enovito obliko oziroma enovit standard za popisovanje gradiva. Za razliko od ISAD(G) (2000), ki temelji na izdelavi pripomočkov, RIC (RIC 2023) gradi pristop na modeliranju entitet in relacij med njimi kot podlagi za dokumentiranje in popisovanje popisnih enot in je primarno namenjen arhivistom. Posledično se RIC (2023) opira na intelektualni opis zapisa, vključno z določenimi

standardnimi pristopi k popisovanju, tj. z zajemanjem potrebnih informacij v obliki, kot jih predvideva ISAD(G) (2000).

RIC (2023) kot konceptualni model služi kot osnova za popisovanje zapisov z namenom zagotavljanja kratkoročne in dolgoročne hrambe ter uporabnosti gradiva. Zagotavlja konceptualni okvir, ki temelji na arhivskih načelih za načrtovanje in implementacijo standardiziranih sistemov za intelektualni nadzor in opis analognih in digitalnih zapisov (RIC 2023).

Na sliki 2 je prikazan primer konceptualnega modela relacije entitet, ki pojasnjuje identificirane entitete in povezave med njimi.

Slika 2: Primer konceptualnega modela relacije entitet

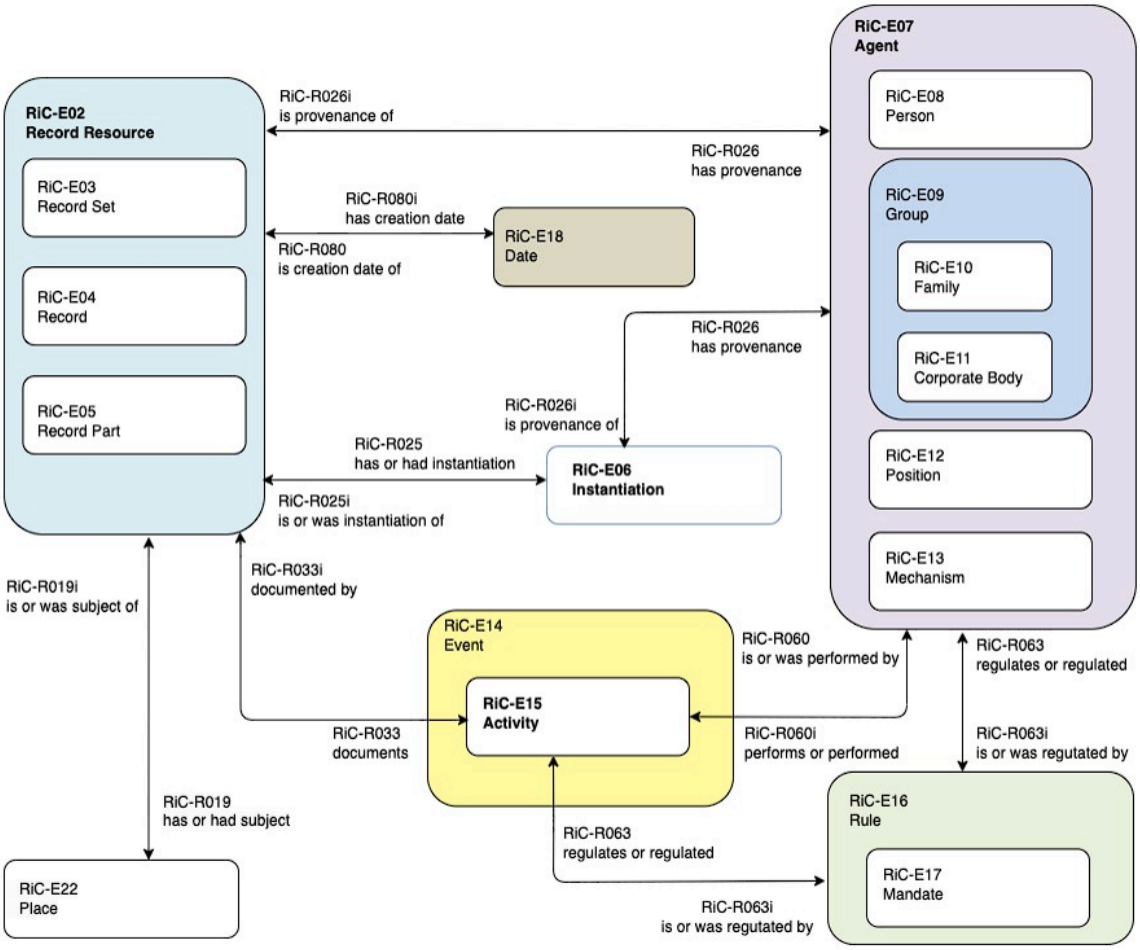


Vir: RIC 2023.

Ključna razlika med ISAD(G) (2000) in RIC (2023) se kaže v razumevanju, kako izvesti popis popisne enote predvsem z vidika linearnega oziroma nelinearnega pristopa. Z uporabo RIC (2023) je bil namen preseči linearni večnivojski pristop k popisovanju zapisov, ki so bili pridobljeni s strani enega ustvarjalca gradiva, ter se funkcionalno osrediniti bolj v smeri večdimenzijskega popisovanja s predominantnim pristopom, v okviru katerega izstopajo dokumentirane relacije in odvisnosti med posameznimi zapisi.

Slika 3 prikazuje pregled konceptualnega modela RIC, vključno z relacijami in odvisnostmi.

Slika 3: Konceptualni model RIC



Vir: RIC 2023.

Poleg omenjenih standardov in konceptualnih modelov se v praksi uporabljajo tudi drugi standardi, odvisno od področja uporabe in namenov, zaradi katerih so bili razviti. V Združenih državah Amerike tako poleg že omenjenih mednarodnih standardov uporabljajo tudi svoj standard Encoded archival description (EAD 2019), vendar je namenjen mednarodni uporabi.

Kot je opredeljeno v standardu je Encoded archival description (EAD 2019) mednarodni standard za izdelavo hierarhičnih opisov arhivskega gradiva. Razvoj standarda je omogočil ustvarjanje elektronskih pripomočkov za iskanje v okviru arhivskih podatkovnih struktur, ki so skladne z zahtevami splošnega mednarodnega standarda ISAD(G) (2000), predvsem zaradi hitrega prehoda hrambe in dostopnosti opisov arhivskega gradiva na internet.

2.4 Umetna inteligenca in strojno učenje

2.4.1 Kaj je umetna inteligenca

> »Umetna inteligenca je obsežno in raznoliko področje računalništva. Lahko jo opišemo kot študij inteligentnih agentov, ki zaznavajo in sprejemajo podatke iz okolice in s pomočjo

> implementirane funkcije izvršujejo akcije. Tako so inteligentni agenti postali glavna unifikacijska tema, ki povezuje zelo različna področja umetne inteligence. Z agenti lahko predstavimo algoritme z različnih področij, kot so reševanje problemov z iskanjem, igranjem iger, avtomatsko sklepanje, planiranje, verjetnostno sklepanje, klasifikacija, razpoznavanje itd.« (Guid in Strnad 2007, 1).

Razumevanje delovanja umetne inteligence je predpogoj za njeno uporabo. Danes bolj kot kadarkoli je s približevanjem konceptov umetne inteligence in strojnega učenja povprečnemu uporabniku, ki je osnovno informacijsko ozaveščen, močno dvignilo zavedanje o vplivu na družbo in delovanje posameznikov ter zmožnosti, ki jih prinašajo umetna inteligenca in rešitve, ki temeljijo na uporabi konceptov umetne inteligence. Umetna inteligenca za večino posameznikov tako ne predstavlja več abstraktnega pojma, ki je povezan z uporabo kompliciranih informacijskih sistemov, saj so jim bile sodobne rešitve, ki uporabljajo umetno inteligenco oziroma strojno učenje, ponujene na prijazen in uporaben način.

Russell in Norvig (2010, 3) sta raziskovala področje umetne inteligence z vidika pristopa k razumevanju človeškega razmišljanja, ki je podlaga za delovanje oziroma primerjavo delovanja umetne inteligence. Določila sta tri kategorije, na osnovi katerih je možno spremljati oziroma razumeti človeško razmišljanje, in sicer psihološke eksperimente, introspekcijo in opazovanje delovanja možganov. Uporaba umetne inteligence na področju arhivistike praviloma ne odstopa od uporabe umetne inteligence na ostalih znanstvenih področjih. V osnovi je cilj uporabe umetne inteligence enak, tj. olajšanje, optimiziranje in učinkovitejše izvajanje arhivskih del.

Obstaja več področij uporabe umetne inteligence. Hlede (2009, 13) izpostavlja, da »obstajajo vsaj štiri glavna področja umetne inteligence, to so vizualna inteligenca, ki se ukvarja s prepoznavanjem oblik, obrazov, prstnih odtisov ipd., govorna inteligenca, ki preučuje prepoznavanje govora ter njegovo sintezo, manipulativna inteligenca, ki se ukvarja z nadzorom gibanja robotske roke ter nadzorom nožnih mehanizmov, ter racionalna inteligenca, ki se ukvarja z ekspertnimi sistemi, podatkovnimi bazami ipd.« Nadalje navaja, da je »poleg glavnih področij še kup podpodročij, kot so strojno učenje (stroji s sposobnostjo učenja), nevronske mreže (najbolj znani stroji, ki se lahko učijo in so izdelani po vzoru naših možganov), ekspertni sistemi (pomoč pri odločanju), računalniški vid (sistemi, ki računalnikom omogočajo videti objekte) itd.« (13).

V strokovni literaturi je izpostavljenih več različnih opredelitev koncepta umetne Na splošno ne obstaja prednostna ali skupna opredelitev, ki bi se jo masovno uporabljalo za temeljno razlago. Balič (2004, 183) jo je opredelil kot »znanost, katere cilj je narediti stroj, ki bo delal takšne stvari, pri katerih je potrebna inteligenca človeka«, Barredo Arrieta idr. (2020) so definirali umetno inteligenco kot sistem, ki lahko snuje ali se izvaja samostojno brez posegov človeka, Cath (2018), Vollmer idr. (2020) in Padilla (2020) so umetno inteligenco opredelili kot sisteme, katerih delovanje je tako kompleksno, da jih je zelo težko obrazložiti, vendar je treba zagotoviti jasne omejitve pri njihovi uporabi skozi uvajanje ustreznih regulativnih

ukrepov in kontrol, ki bi omogočali večjo informiranost o delovanju takšnega sistema. Cordell (2020) primerja umetno inteligenco s skupkom metod in raziskovalnih ciljev, ki omogočajo samostojno delovanje. Duan idr. (2019) so umetno inteligenco opredelili kot sposobnost stroja, da se uči iz izkušenj, prilagaja glede na vhodne informacije in izvaja človeška opravila. Podobne zaključke je sprejelo tudi več raziskovalcev, ki so ugotavljali potencialne meje umetne inteligence (Lake idr. 2017; Miller 2019; Hagendorff 2021).

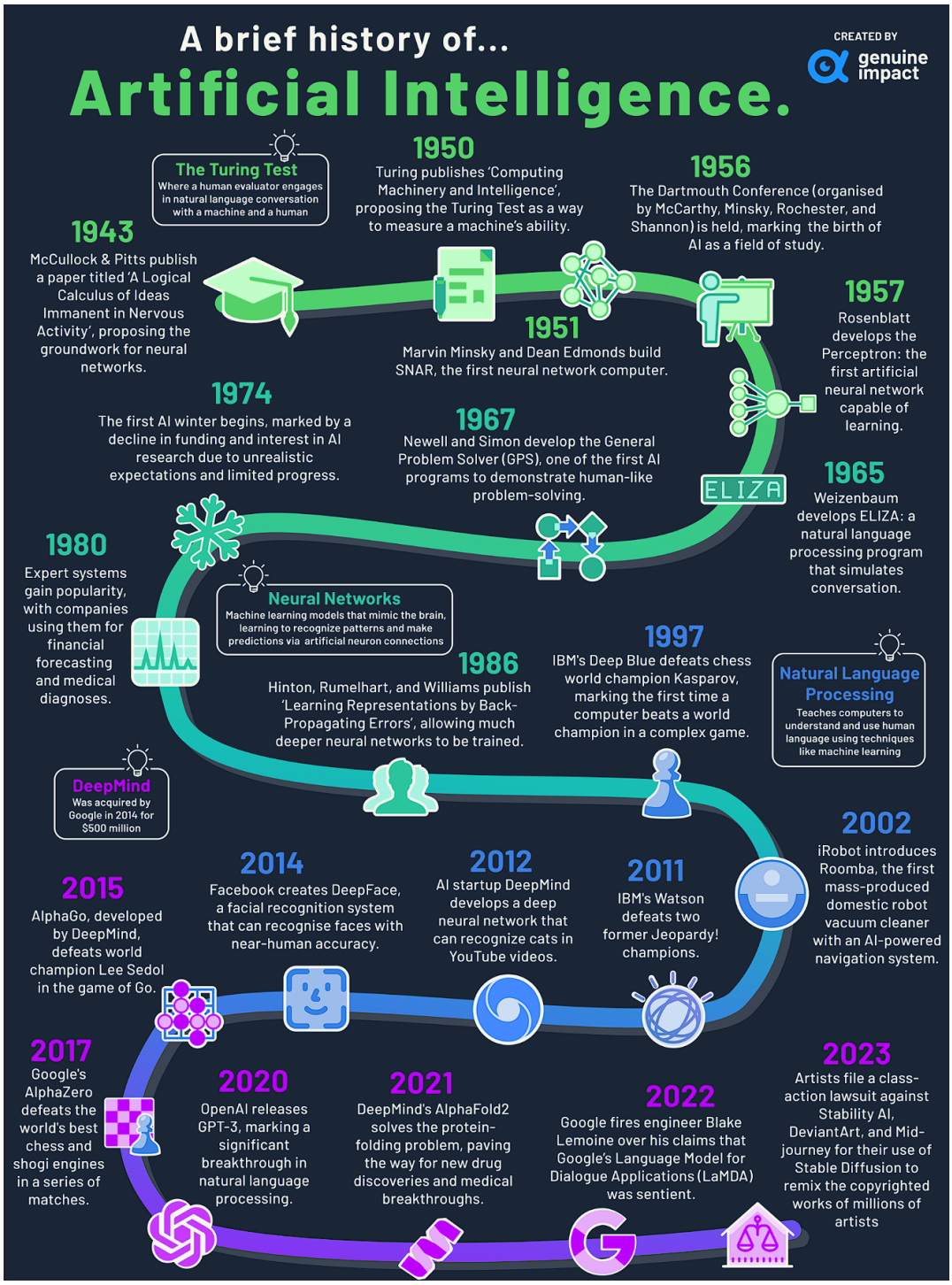
Kot je izpostavljeno, za termin in koncept umetne inteligence obstaja več različnih definicij, ne glede na to pa imajo vse skupni imenovalec, in sicer, da naj bi pod pojmom umetna inteligenca na splošno iskali predvsem ravnanje in delovanje, ki je enako človeškemu razmišljanju, ravnanju in delovanju.

2.4.2 Zgodovina umetne inteligence

Umetna inteligenca ni nov ali sodoben koncept. Njeni začetki segajo nekaj desetletij nazaj, splošna uporaba predvsem v praksi pa se kaže šele v zadnjem času. Skozi čas se je umetni inteligenci velikokrat pripisovalo tudi negativne, kar apokaliptične posledice, če bi dosegla nivo samostojnega razmišljanja in ustvarjanja, predvsem pa samostojnega delovanja.

Na sliki 4 je prikazan pregled razvoja umetne inteligence skozi različna časovna obdobja od prvih omemb do danes.

Slika 4: Razvoj umetne inteligence skozi čas



Vir: Genuine Impact 2023.

Prevladuje splošno mnenje, da se je današnja pot umetne inteligence začela v sredini 20. stoletja. Razvoj umetne inteligence bi zato lahko razdelili na več smiselno zaokroženih obdobij (Dey 2021):

- obdobje v štiridesetih in petdesetih letih 20. stoletja, ko je bila izoblikovana osnovna definicija umetne inteligence ter opredelitev možnih testiranj za merjenje strojnih (inteligentnih) zmogljivosti;

- obdobje v šestdesetih in sedemdesetih letih 20. stoletja, ko so bili razviti prvi sistemi za obdelavo naravnega jezika;
- obdobje v sedemdesetih in osemdesetih letih 20. stoletja, ko so bili razviti prvi ekspertni sistemi, tj. učeči pristopi za izvajanje ekspertnih opravil;
- obdobje v osemdesetih in devetdesetih letih 20. stoletja, ko je prišlo do ekspanzije uporabe nevronskih mrež in vedenjsko usmerjene robotike;
- obdobje v devetdesetih letih 20. stoletja in prvem desetletju 21. stoletja, ko sta se močno uveljavila strojno učenje in razvoj odločitvenih modelov;
- obdobje v prvem in drugem desetletju 21. stoletja, ko je razvoj umetne inteligence prinesel praktične rešitve enostavnim uporabnikom, kot so npr. virtualni asistenti, avtonomno delovanje strojev, optična prepoznava slik itd.;
- obdobje od leta 2020 naprej, ko je razvoj umetne inteligence dosegel točko vključenosti v različne družbene in poslovne vidike, npr. uporaba umetne inteligence v zdravstvu, kvantno preračunavanje, robotika, razvoj praktičnih rešitev z uporabo strojnega učenja, učinkoviti konverzacijski modeli idr.

Obdobje od petdesetih do sedemdesetih let se pri dokumentiranju zgodovine umetne inteligence zaradi izjemnega optimizma, hitrega razvoja in doseženih napredkov velikokrat omenja kot t. i. zlato obdobje umetne inteligence (Wooldridge 2021). V sredini sedemdesetih se je napredek na področju umetne inteligence počasi začel ustavljati, vzroke pa se je večinoma pripisovalo pomanjkanju zanimanja za težko dosegljive praktične rešitve, ki so bile še vedno zelo okorne glede na strojne zmogljivosti tistega časa, zaradi česar se je tudi financiranje razvoja v precejšnji meri ustavilo. Razvoj umetne inteligence se je nato konec sedemdesetih nadaljeval z uvedbo ekspertnih sistemov, ki so začeli uporabljati izkušnje ljudi za reševanje specifičnih problemov, kar je predstavljalo gonilno silo razvoja umetne inteligence za naslednje desetletje (Wooldridge 2021).

Razvoj umetne inteligence od konca sedemdesetih let naprej ni več doživel večjih zastojev in je do danes približal koncept in uporabo rešitev umetne inteligence tako večjim uporabnikom (organizacijam) kot posameznikom, kar je spodbudilo večje zanimanje in pritisk za zagotavljanje večje podpore nadaljnjemu razvoju umetne inteligence.

2.4.3 Področja uporabe umetne inteligence

Danes se koncepti in rešitve, ki uporabljajo umetno inteligenco, pojavljajo v vseh družbenih segmentih in različnih izvedbenih oblikah. Rešitve, ki uporabljajo umetno inteligenco, lahko zasledimo na poslovnih področjih, v zdravstvu, izobraževanju, kulturi, arhivistiki in drugod. Praktičnih omejitev za uporabo umetne inteligence ni, zato se rešitve lahko uporabljajo v različnih primerih uporabe ne glede na področje, kjer je namen takšne rešitve uporabiti.

Davenport in Ronanki (2018) sta z vidika poslovne uporabnosti opredelila tri pristope, s pomočjo katerih lahko umetna inteligenca izboljša poslovanje posamezne organizacije:

- avtomatizacija poslovnih procesov,

- razumevanje poslovnega okolja skozi analizo podatkov ter
- delo s strankami in zaposlenimi.

Danes so prednosti uporabe umetne inteligence opazne na različnih področjih. V primeru avtomatizacije poslovnih procesov so lahko doseženi cilji, ki so bili predhodno težko ali celo nemogoče dosegljivi. V primeru arhivistike je eden od takšnih problemov tudi temeljni problem, ki je zajet v tej raziskavi, in sicer, kako izvesti urejanje in popisovanje nestrukturiranih besedil v velikih količinah. Brez uporabe rešitev, ki jih ponuja umetna inteligenca, bi bile takšne aktivnosti praktično neizvedljive oziroma v najboljšem primeru neracionalne in neoptimalne. Reševanje takšnega problema z uporabo umetne inteligence sodi med t. i. optimizacijo oziroma avtomatizacijo poslovnega procesa na področju arhivistike.

Do podobnih ugotovitev je prišel tudi Dilmegani (2024), ki je uporabo umetne inteligence v poslovnem segmentu razdelil na dve osnovni področji: (1) splošne rešitve s področja umetne inteligence, kot so analiza podatkov za zaposlene, analiza trga in avtomatizacija procesov z uporabo strojnega učenja, in (2) specializirane rešitve, kot so npr. konverzacijska analiza, analiza e-poslovanja, geoanaliza, prepoznava slik in analiza v resničnem času. Dilmegani (2024) in Kreutzer in Sirrenberg (2020) so nadalje še podrobneje razdelili uporabo specializiranih rešitev v poslovnem segmentu glede na različne poslovne potrebe, kot so rešitve, ki so usmerjene v podporo:

- strankam (analiza klicev, klasifikacija dogodkov, konverzacijski roboti itd.),
- pri upravljanju s podatki (zagotavljanje kakovosti podatkov, validacija, transformacija, vizualizacija, NLP, itd.),
- finančnim in računovodskim službam (avtomatizacija, analitika),
- pri upravljanju s kadri (zaposlovanje, merjenje učinkovitosti),
- marketingu (analiza trga, personalizacija, kontekstualni marketing),
- produkciji (kognitivna in inteligentna avtomatizacija, robotizacija, rudarjenje podatkov in analize, izdelava prediktivnih modelov, izdelava odločitvenih modelov, optimizacija itd.),
- prodaji (izdelava prediktivnih modelov prodaje, avtomatizacija opravil, merjenje učinkovitosti, treniranje in izobraževanje, personalizacija prodaje, chatboti itd.),
- na področju tehnologije (avtomatiziran in samostojen razvoj programske opreme, varnostne analize in prediktivni modeli na področju informacijske varnosti, upravljanje z znanjem, obdelava naravnega jezika, varnost komunikacij, pametni varnostni nadzori, varovanje osebnih podatkov itd.).

V industriji bi umetno inteligenco lahko izpostavili za (Chan 2022; Dilmegani 2024):

- uporabo na področju prevoza in avtonomije (asistenti za vožnjo, kibernetska varnost, vizualni sistemi, samovozeča vozila),
- uporabo na področju izobraževanja (samodejna izdelava izobraževalnih programov in vsebin za izobraževanje, samodejno ocenjevanje, tutoring, personalizacija, varovanje podatkov, restavriranje starih izobraževalnih vsebin),

- uporabo na področju oblikovanja in mode (kreativno ustvarjanje, virtualizacija, analiza trendov, samodejno ustvarjanje slik),
- uporabo na področju finančnih sistemov (samodejna detekcija prevar, avtomatizacija, finančna analiza, upravljanje s stroški, kreditiranje in izračun tveganj, izterjave, zagotavljanje skladnosti, pridobivanje podatkov, konverzacijski roboti),
- uporabo na področju medicine (analiza zdravstvenih podatkov, personalizacija pri negi in določanju zdravljenja, razvoj novih zdravil, analiza in optimizacija triaže in prioretizacijskih pristopov, hitrejše in učinkovitejše določanje diagnoz, avtomatizacija pri izdaji receptov, upravljanje z zdravljenjem, obdelava slik, genska analiza itd.).

Na področju razvoja informacijskih rešitev (programov) razvoj umetne inteligence danes praviloma predstavlja matematično vsebino, ki se prevaja v izvorno kodo. Kar v precejšnji meri povzroča težave tradicionalnim razvijalcem (programske opreme), ki sledijo drugačni paradigmi razvoja programske opreme. Večji tehnološki velikani se zato usmerjajo v uporabo oblačnih rešitev, ki bodo v prihodnosti potrebovale čim manj tovrstnih inženirjev (Rothman 2018, 8).

2.4.4 Vplivi umetne inteligence

Umetna inteligenca ima svoje prednosti in slabosti. Okvirno bi lahko delovanje umetne inteligence razdelili na močno in šibko umetno inteligenco (Collins idr. 2021). Za močno umetno inteligenco je značilno, da v svoje delovanje ne vključuje potencialnih omejitev, ki bi jih postavili ljudje, in razpolaga s sposobnostjo, da samostojno interpretira in obdeluje informacije. Treba je ločiti med odločanjem umetne inteligence na osnovi pravil, ki jih umetna inteligenca strogo spoštuje in so jih postavili razvijalci sistema, in odločanjem na osnovi pravil, ki ji niso bila strogo določena. Odločanje na podlagi pravil, ki so jih postavili razvijalci sistema, se pripisuje šibki umetni inteligenci, medtem ko se odločanje po pravilih poskusa nagiba k močni umetni inteligenci. Primer odločanja po prej določenih pravilih so nevronske mreže, ki algoritmom omogočajo, da se učijo sami. Močno umetno inteligenco bi tako predstavljali sistemi, ki bi ustvarjali lastna pravila, ki bi jim nato sledili, kar v tem trenutku še ni mogoče in izvedljivo (Collins idr. 2021).

Khanzode in Sarode (2020) ter Bhbosale idr. (2020) so predstavili nekaj prednosti in slabosti umetne inteligence:

- prednosti umetne inteligence:
  - hitrejše izvajanje nalog od ljudi pri enaki količini, načinu obdelave in enakem pristopu dela (avtomatizacija opravil),
  - ni stresa in preobremenitev,
  - ni utrujenosti, ni potrebnih premorov,
  - lokacijska neodvisnost,
  - težka opravila se lahko opravijo učinkoviteje in v krajšem času,
  - izvajanje več nalog hkrati,

    - doseganje dobrih rezultatov glede na masovno obdelavo in zagotavljanje informacij za optimizacijo procesov,
    - standardizirano neodstopajoče ponavljanje nalog,
    - manj napak pri obdelavi podatkov,
    - večja učinkovitost pri večopravilnosti,
    - majhni stroški glede na opravljeno delo in dosežene rezultate,
    - uporaba z namenom zmanjševanja tveganj za ljudi (npr. uporaba robotov v nevarnih okoljih),
    - odkrivanje relacij, povezav, vzorcev pri predhodno neznani vsebini;
- slabosti umetne inteligence:
    - v primeru napak potenciranje morebitne škode,
    - vpliv na potrebe po človeških virih (zaposlitvah),
    - kreativnost opravil in pravil je odvisna od snovalcev sistema,
    - spodbujanje lenobe,
    - potencialni visoki stroški razvoja in postavitve,
    - odvisnost od tehnologije,
    - potencialna zloraba in neetična uporaba.

V zadnjem času se pojavlja skrb zaradi uporabe umetne inteligence, in sicer predvsem glede njenega vpliva na trg dela, saj se vse več opravil, ki so jih izvajali ljudje, izvaja v sistemih, s katerimi upravlja umetna inteligenca. Spet drug vidik, ki vzbuja veliko skrbi, je etična uporaba umetne inteligence. Za izvajanje opravil umetna inteligenca potrebuje vhodne informacije, pri čemer lahko pride do zlorab, zato se sproža vse več iniciativ za uvedbo regulacij, ki bi opredeljevale meje izdelave in uporabe sistemov, ki uporabljajo umetno inteligenco.

Z razvojem strojnih zmogljivosti in zmožnosti obdelave podatkov oziroma izvedbe računsko intenzivnih opravil se je umetna inteligenca izkazala za uporabno v različnih panogah, kar je ustvarilo dodatno povpraševanje po rešitvah, ki so bile v preteklosti večini nedosegljive.

2.4.5 Regulacija uporabe umetne inteligence

Ker je večina sistemov, ki uporabljajo umetno inteligenco, omejenih s pravili, ki jih določajo snovalci tovrstnih sistemov, ki za svoje delovanje niso imeli jasnih regulatornih okvirov, je bilo zgolj vprašanje časa, kdaj in na kakšen način se bo sprožil mehanizem za regulacijo izdelave in uporabe sistemov umetne inteligence.

Ključni premiki na omenjenem področju se dogajajo ravno v Evropski uniji, ki je v fazi sprejema regulatornega okvira, ki bo urejal področje uporabe umetne inteligence. Evropska komisija je že leta 2021 pripravila prvi regulatorni okvir za uporabo umetne inteligence, predvsem v pogledu bolj jasnega pristopa k njenemu razvoju kot tudi uporabi. Regulatorni okvir, ki bo veljal v Evropski uniji, naj bi tako zajemal upravljanje s tveganji na štirih nivojih

(Uredba evropskega parlamenta in sveta o določitvi harmoniziranih pravil o umetni inteligenci – predlog 2022):

- nesprejemljiva tveganja,
- visoka tveganja,
- splošna in generativna uporaba (omejeno tveganje),
- nizko tveganje.

Nesprejemljiva tveganja in z njimi prepoved uporabe umetne inteligence tako vključujejo delovanje in uporabo umetne inteligence tam, kjer bi lahko prišlo do kognitivne manipulacije z ljudmi ali ranljivimi skupinami, uporabe sistemov vrednotenja, kjer bi se klasificiralo ljudi glede na njihovo obnašanje, ekonomski status ali osebne lastnosti, uporabe biometričnih podatkov za namen identifikacije ali izdelave kategorizacij ljudi, ter identifikacije oseb v resničnem času, kot je npr. prepoznava obrazov, prepoznava čustev idr. Pri tem so predvidene določene izjeme za omenjeno uporabo s strani organov pregona, vendar le na osnovi sodne odločitve in le v ozko opredeljenih primerih uporabe (Uredba Evropskega parlamenta in Sveta o določitvi harmoniziranih pravil o umetni inteligenci – predlog 2022).

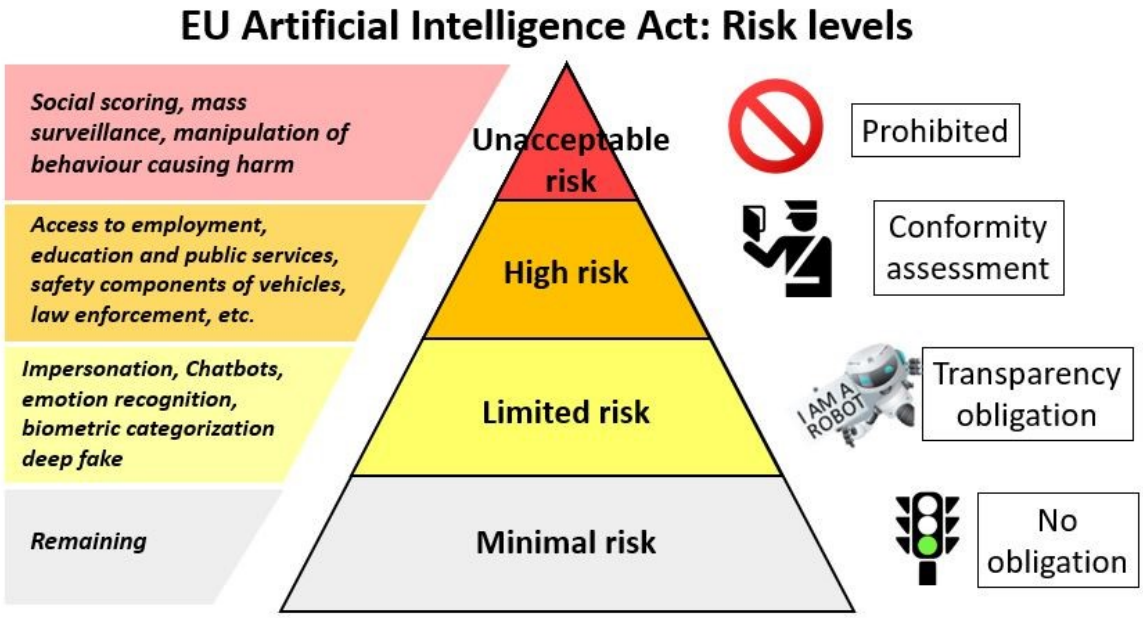
Visoka tveganja vključujejo izdelavo in uporabo sistemov umetne inteligence, pri katerih bo za zmanjševanje tveganj potrebno izvajanje določenih obveznosti. Med omenjena tveganja se uvršča uporaba umetne inteligence, kjer bi se lahko pojavile potencialno večje negativne posledice na področju zdravja, splošne varnosti, temeljnih pravic, varovanja okolja in varovanja demokratičnih načel ter pravne države. Predvidene so unificirane kontrole, kot je npr. izdelava ocene učinka za temeljne pravice, ki bodo veljale tudi za zavarovalniški in bančni sektor. Uporaba umetne inteligence z namenom vplivanja na izid volitev se ravno tako uvršča med visoko tvegano uporabo. S tem je povezan in predviden tudi inštrument pritožb s strani posameznikov, ki bo omogočal, da lahko posamezniki prejmejo jasna pojasnila o odločitvah, temelječih na uporabi umetne inteligence, ki sodijo med visoko tvegano uporabo (Uredba Evropskega parlamenta in Sveta o določitvi harmoniziranih pravil o umetni inteligenci – predlog 2022).

Pri splošni in generativni uporabi sistemov umetne inteligence je bilo predvideno, da se pri upoštevanju širokega spektra nalog, ki jih lahko sistemi umetne inteligence opravijo, in hitri širitvi njihovih zmogljivosti zagotovi, da bodo sistemi umetne inteligence za splošne namene in modeli, na katerih temeljijo, morali upoštevati zahteve glede preglednosti. Ti vključujejo predvsem pripravo tehnične dokumentacije, zagotavljanje skladnosti z evropsko zakonodajo o avtorskih pravicah in razširjanje podrobnih povzetkov o vsebini, ki se uporablja za usposabljanje. Pri splošni in generativni uporabi, ki ima velik vpliv in zato pomeni sistemsko tveganje, so predvidene strožje obveznosti. Če bodo ti modeli izpolnjevali določena merila, bodo upravljavci morali izvesti ocene modelov, oceniti in ublažiti sistemska tveganja, izvesti ustrezno testiranje, poročati pristojnim organom o resnih incidentih, zagotoviti ustrezno kibernetsko varnost in poročati o njihovi energetski učinkovitosti (Uredba Evropskega parlamenta in Sveta o določitvi harmoniziranih pravil o umetni inteligenci – predlog 2022).

Pri nizkih tveganjih je pripravljeno osnovno izhodišče, da bi uporabljeni sistemi umetne inteligence morali izpolnjevati minimalne zahteve glede preglednosti, kar bi uporabnikom omogočilo sprejemanje premišljenih odločitev. Kot primer se tako lahko izpostavi nadzorovano interakcijo z aplikacijami, ki uporabljajo umetno inteligenco, kjer se lahko uporabnik v kateremkoli delu procesa uporabe sam odloči, ali jo želi še naprej uporabljati. Uporabniki morajo biti jasno obveščeni, da komunicirajo s sistemom umetne inteligence. To vključuje npr. sisteme umetne inteligence, ki ustvarjajo ali manipulirajo s slikovno, zvočno ali video vsebino (Uredba Evropskega parlamenta in Sveta o določitvi harmoniziranih pravil o umetni inteligenci – predlog 2022).

Na sliki 5 je prikazana struktura predvidenih tveganj in obveznosti, ki sledijo identificiranim tveganjem v primeru uporabe umetne inteligence.

Slika 5: Struktura tveganj in povezanih obveznosti



Vir: Telefonica 2024.

Druge države (izven Evropske unije) so trenutno še v fazi izdelave lastnih regulatornih okvirov, skupni konsenz večine pa je, da se področje umetne inteligence v zadnjem času hitro razvija in da je treba zagotoviti ustrezne ukrepe za nadzor uporabe umetne inteligence. V Združenih državah Amerike (v nadaljevanju ZDA) podobno kot v Evropski uniji (v nadaljevanju EU) poteka postopek za sprejem regulatornega okvira, ki bi zagotavljal varno uporabo umetne inteligence. Določeni regulatorni okviri so bili v delih ZDA začasno sprejeti za specifično uporabo umetne inteligence. Tak je npr. kalifornijski zakonodajni akt, ki ureja področje globokih ponaredkov. Pričakovani regulatorni okvir bo osredotočen predvsem na zahtevanje odgovornosti podjetij vseh velikosti, zlasti na področju generativnih rešitev umetne inteligence z visokim tveganjem, kot so zdravje in varnost potrošnikov, zasebnost občutljivih informacij (zdravje, finance, preverjanje pristnosti identitete, biometrija idr.) ter preverjanje prijav za zaposlitev (Foley 2024).

2.4.6 Umetna inteligenca in etika

Umetna inteligenca lahko močno vpliva na način, kako ljudje razmišljamo (npr. vpliv na volitve), zato se vzporedno z razvojem umetne inteligence vedno znova pojavljajo vprašanja, kako zagotoviti njeno etično izdelavo in uporabo. Z vidika etike se bo vedno postavljalo vprašanje, kako umetna inteligenca deluje in pod kakšnimi pogoji se razvija. Zaradi vse lažje dostopnosti do sistemov umetne inteligence se pojavljajo skrbi, kako bo umetna inteligenca globalno vplivala na družbo in posameznike in kakšen nadzor na izdelavo in uporabo takšnih sistemov je možno zagotoviti.

Etika na področju umetne inteligence tako pomeni predvsem pogled na vpliv in potencialne posledice, ki ga predstavljata razvoj in uporaba umetne inteligence (Boddington 2023). Ena od ključnih potencialnih posledic je tudi, kako bo razvoj umetne inteligence vplival na potrebo po uporabi ljudi za opravljanje dela (Kumar idr. 2021, 10).

Zaradi vplivov zabavne industrije je umetna inteligenca prepoznana kot ena največjih groženj, s katerimi se sooča človeštvo. Iz občutka nezaupanja nastajajo scenariji, v katerih se pri uvajanju kakršnekoli nove tehnologije stvari potencirajo do nečesa, kar se na koncu zdi neobvladljivo, in sicer vse od strahu pred izgubo delovnih mest zaradi avtomatizacije človeških dejavnosti do zatekanja k uporabi umetne inteligence za opravljanje temeljnih človeških nalog, kot je odločanje. Pomisleki glede varovanja zasebnosti, socialne neenakosti in potencialne diskriminacije lahko prav tako povečajo omenjene skrbi (Benedetti del Rio 2023, 179).

Schmidpeter in Funk (2023, 11) izpostavljata, da nam temeljni optimizem pri uporabi umetne inteligence lahko pomaga pri ustvarjanju družbene in podjetniške dodane vrednosti. Računalniški algoritmi zagotovo lahko pomagajo pri boljšem in hitrejšem reševanju kompleksnih izzivov, vendar pa ne morejo v celoti kopirati človeške inteligence, predvsem pa ne morejo nadomestiti »razuma« in sposobnosti ravnati etično (ena najbolj kritičnih človeških edinstvenih točk ravnanja). Umetna inteligenca bi zato morala služiti ljudem in prepoznati človekovo dostojanstvo in človekove pravice v vseh okoliščinah. Pomemben izziv pri tej preobrazbi je vprašanje ustvarjanja. Razvoj umetne inteligence je že prišel do točke, ko je sposobna samostojno pisati ali optimizirati programsko kodo. Največji izziv pri tem je uporabljati takšne možnosti ter jih povezati z etičnim delovanjem in razumom.

Boddington (2023) pri uporabi umetne inteligence v povezavi z etiko izpostavlja naslednje teme, ki so aktualne na trenutni stopnji razvoja in uporabe umetne inteligence:

- svoboda in avtonomija pri sprejemanju odločitev,
- transparentnost in razložljivost,
- pravičnost in nediskriminatoren odnos,
- dobronamernost in neškodljivost,
- odgovornost,
- zasebnost,

- zaupanje,
- trajnost in vzdržnost,
- dostojanstvo in
- solidarnost.

Svoboda in avtonomija pri sprejemanju odločitev sistemov umetne inteligence je ena od ključnih točk, zaradi katere se pojavljajo pozivi k regulaciji in etičnemu ravnanju. Ko je namen prepustiti sprejemanje odločitev določenemu sistemu umetne inteligence, se vedno pojavi tudi vprašanje, v kolikšnem obsegu naj se tovrstno delovanje omogoči. Sprejemanje odločitev pri enakih situacijah se lahko pri ljudeh spreminja, odvisno od subjektivnih kriterijev, ki so del odločitve, zato bo vedno tudi vprašanje, ali so splošne rešitve/odločitve, ki bi jih praviloma sprejemala umetna inteligenca, tudi vedno najboljša rešitev za vse posameznike, na katere bi takšne rešitve vplivale. Določanje omenjenega obsega oziroma zmogljivosti in sposobnosti avtonomije samostojnega sprejemanja odločitev bi tako moralo sprožiti širšo diskusijo in raziskavo, še posebej glede potencialnih posledic.

Transparentnost in razložljivost sta pomembni zato, da so posamezniki in končni uporabniki jasno seznanjeni z uporabo in delovanjem sistemov umetne inteligence. Predhodno je bilo omenjeno, da je danes veliko sistemov izdelanih na način, da je razložljivost delovanja takšnega sistema izjemno zahtevna ali pa v določenih delih tudi neizvedljiva, kar ne vliva zaupanja v podlago za sprejem odločitev, saj ni jasno razvidno, zakaj so bile določene odločitve ali delovanje izvedeni na način, kot so bili izvedeni. Da bi lahko zagotovili transparentno in razložljivo obliko, bi bilo treba vzpostaviti jasne referenčne in po potrebi regulatorne okvire, ki bi standardizirali način zagotavljanja transparentnosti za aktualne oblike kot tudi pričakovane oblike izdelave in uporabe umetne inteligence. Zaradi potencialnega masovnega vpliva na posameznike in potencialno ključnega vpliva na delovanje družbe je to eno od vprašanj in področij, ki bi se moralo obravnavati prioritetno, saj je veliko takšnih sistemov umetne inteligence danes že v uporabi in že vplivajo na posameznike in družbo. Z ustrezno transparentnostjo bi lažje dosegli cilj višje ozaveščenosti posameznikov, ki bi imeli možnost izbire, kako in če sploh uporabljati določen sistem umetne inteligence. Posledice izdelave in uporabe umetne inteligence vplivajo na posameznike v celotnem spektru družbe. Način doseganja ustrezne transparentnosti in razložljivosti bi moral zaradi tega zasledovati pristop, kjer bi bila obrazložitev podana na čim bolj preprost način, ki bi bil razumljiv širši množici in ne zgolj določeni skupini ljudi, npr. razvijalcem ali raziskovalcem sistemov umetne inteligence (Boddington 2023).

Pravičen in nediskriminatoren odnos v osnovi pomeni prenos etičnih in moralnih načel družbe v izdelavo, delovanje in uporabo sistemov umetne inteligence. Med takšna načela bi lahko všteli politična, socialna in zakonodajna pravila (Mazzi idr. 2023), kar lahko predstavlja tudi določeno omejitev v primeru, da se pravila razlikujejo glede na različno družbeno sredino ali družbene nazore. Kar v določeni družbi predstavlja pravično in nediskriminatorno delovanje, lahko pomeni ravno nasprotno v drugi ločeni družbi, zato je treba pristopiti k urejanju omenjenega področja na način, da se upošteva tako splošno sprejeta pravila kot potencialna specifična pravila, ki veljajo v družbi, kjer naj bi se izvedla

regulacija sistemov umetne inteligence, pri tem pa je treba upoštevati zgodovinski in socialni razvoj takšnega okolja.

Dobronamernost in neškodljivost naj bi zasledovala pristop za izdelavo in uporabo sistemov umetne inteligence z namenom, da ne povzroča škode oziroma da ne deluje zlonamerno, kar je ponovno precej odvisno od družbene sredine ter etičnih in moralnih načel in povezanih pravil, ki veljajo v določenem okolju. Ker sta izdelava in uporaba sistemov umetne inteligence danes postali široko dostopni, so se pojavili tudi apetiti po uporabi takšnih sistemov za doseganje lastnih individualnih ciljev, ki nujno ne predstavljajo primarne ali nameravane uporabe izdelanih sistemov. Kot primer se lahko izpostavi uporabo konverzacijskih sistemov umetne inteligence, ki so zmožni ustvarjati koncipirano besedilo z veliko količino podatkov, ki so jih predhodno obdelali. Na področju izobraževanja se je zato hitro začelo izkoriščati omenjene sisteme za pripravo raznoraznih besedil in nalog ter s tem zasledovati individualne cilje posameznikov. Na področju arhivistike je eno od osnovnih etičnih načel, da je treba zagotoviti ustrezno varnost gradiva in se izogniti morebitnim poškodbam ali malomarnostim, ki bi povzročile poškodovanje gradiva (Giménez-Chornet 2017).

Odgovornost je tesno povezana z vzpostavitvijo regulatornih okvirov za izdelavo in uporabo umetne inteligence. Izhodiščno je treba določiti okvir za pripisovanje odgovornosti, v katerega bi morali biti vključeni jasne ločnice, sodila in kriteriji, ki so potrebni za ocenjevanje ustreznosti sistemov umetne inteligence. Predvideni okvir bi moral vključevati odgovore na večino vprašanj, ki so predstavljena v tem poglavju. Zaradi kompleksnosti področja in marsikaterih neodgovorjenih vprašanj in nedodelanih stališč glede posameznih področij je tudi v praksi razvidno, da je sprejem takšnih regulatornih okvirov dolgotrajen proces, ki terja veliko usklajevanj, vključno s potencialnimi omejitvami pri uporabi ali dostopu do vsebin, ki so predmet obdelav z uporabo sistemov umetne inteligence (Holterhoff 2017).

Zasebnost je eden od ključnih elementov z vidika etike pri obravnavi ustreznosti izdelave in uporabe sistemov umetne inteligence. Danes so informacije in v določenem obsegu osebni podatki ključna dobrina delovanja organizacij. Da bi lahko zaščitili posameznike pred nezakonito in nedovoljeno obdelavo osebnih podatkov, je treba predvideti ustrezne kontrole, ki bi preprečevale zlorabo osebnih podatkov in potencialni vdor v zasebnost. Pri tem je zasledovanje varovanja zasebnosti in osebnih podatkov tesno povezano s transparentnostjo in morebitno odgovornostjo, saj se zaradi pomanjkanja transparentnosti lahko posamezniki prostovoljno odrečejo določenim pravicam glede varovanja zasebnosti ali osebnih podatkov zaradi želje po uporabi sistemov umetene inteligence. Ob zagotavljanju ustrezne transparentnosti bi bila takšna odločitev vsaj v delu uporabe na strani posameznika (Siau in Wang 2020). V določenem delu bodo vedno obstajale tudi izjeme glede varovanja zasebnosti, posebej v primerih, ko bi takšna obdelava z družbenega ali individualnega stališča predstavljala upravičen poseg v pravice posameznikov (npr. uporaba s strani organov pregona ali v zdravstvene namene). Uporaba umetne inteligence na področju

varovanja zasebnosti tako predstavlja oboje, potencialne grožnje kot tudi priložnosti za zagotavljanje ustrezne zasebnosti (Liu idr. 2022).

Ko je namen prenesti sprejemanje odločitev od človeka na sistem umetne inteligence, je zaupanje prva stvar poleg svobode pri odločanju in avtonomiji, kateri bo posvečena posebna pozornost. Glede na posledice, ki jih lahko imajo neustrezen razvoj, delovanje in uporaba sistemov umetne inteligence, je zaupanje v ustreznost uporabe takšnega sistema na prvem mestu z vidika zagotavljanja dolgoročne uporabnosti. Na področju razvoja sistemov umetne inteligence tako kot na drugih področjih vlada konkurenčen odnos razvijalcev, ki vsak na svoj način skušajo prepričati končne uporabnike o tem, katera rešitev je za njih najboljša. Pri tem se zaupanje gradi predvsem na konceptih transparentnosti in stabilnosti delovanja, odprtosti tehnologij, neodvisnosti uporabe idr. Pri tem je treba izpostaviti, da se velik del razvoja sistemov umetne inteligence izvaja v okviru komercialnih rešitev, kjer so določeni koncepti zaupanja okrnjeni. Zaupanje je tako lahko zaradi nepopolne odprtosti in nezadostne transparentnosti omajano, pričakovanja pa v veliki meri usmerjena v razvijalce sistemov, saj so oni tisti, ki v osnovi določajo delovanje sistema (Morley idr. 2020).

Trajnost in vzdržnost se z razvojem strojne opreme in zmogljivosti sistemov umetne inteligence kažeta predvsem v načinu izrabe potrebnih resursov za delovanje takšnih sistemov. V osnovi gre pri večini opravil, ki jih izvajajo sistemi umetne inteligence, za računsko intenzivne operacije, kar povzroča vpliv na porabo resursov za izvedbo takšnih operacij. Vplivi tako vključujejo porabo energije, ustrezno distribucijo porabe in pridobljene koristi glede na družbene sredine in geografska področja, potrebno infrastrukturo in vzpostavljanje odvisnosti od uporabe sistemov umetne inteligence. Marsikatera odločitev o razvoju in uporabi sistemov umetne inteligence bo tako odvisna od zmožnosti vzdrževanja takšnih sistemov in trajnosti njihove uporabe. Področje umetne inteligence se razvija hitro, zato je tudi dinamika izrabe resursov visoka in lahko za marsikoga ali marsikatero družbo predstavlja pomembno oviro za trajno uporabo takšnih rešitev. Ob tem se pojavlja vprašanje, kako doseči pravično distribucijo bremen, ki sledijo iz razvoja in uporabe sistemov umetne inteligence. Na področju arhivistike je tako treba zagotoviti ustrezna sredstva, da bo uporaba takšnih sistemov izvedljiva kot podpora dolgoročni hrambi gradiva v digitalni obliki, da se lahko zagotovi pravočasno obdelavo in prepreči izgubo gradiva (Runardotter idr. 2011).

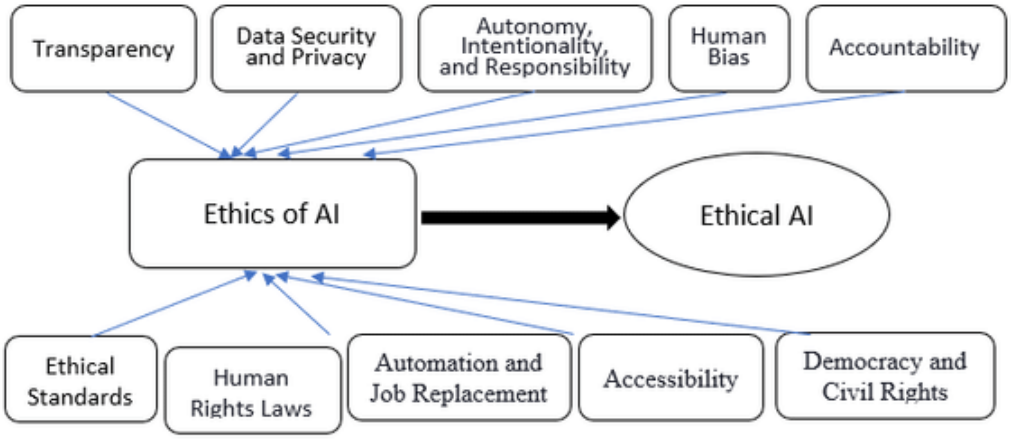
Dostojanstvo je koncept, povezan z etično obravnavo predvsem zaradi vrednotenja dela ali opravil, ki jih izvajajo sistemi umetne inteligence, ki so jih predhodno izvajali ljudje. Ena od temeljnih skrbi in potencialnih negativnih učinkov uporabe sistemov umetne inteligence je zamenjava ljudi na določenih delovnih področjih in s tem vzpostavljanje stališča, ali so njihove izkušnje, znanje in sposobnosti sploh še zaželeni in uporabni. Za posameznike, ki so lahko porabili veliko časa za pridobivanje izkušenj za opravljanje dela, ki bi ga po novem opravljali sistemi umetne inteligence, lahko to predstavlja poseben pritisk na dostojanstvo in razvrednotenje njihovega truda. Ker se funkcionalni vidiki delovanja sistemov umetne inteligence razvijajo na način, da določena rešitev predstavlja rešitev za celoten istovrstni sklop opravil, se lahko cele generacije ljudi, ki so opravljali istovrstno delo in ga lahko opravi sistem umetne inteligence, znajdejo v enaki stiski. Enako je možno sklepati o delovanju v

prihodnosti in povezanih tveganjih pri izbiri profesionalne ali študijske usmeritve, ne vedoč, kaj bo razvoj umetne inteligence prinesel in katera delovna področja in delovna mesta bodo lahko v prihodnosti ogrožena. Spet druga skrb je vpliv na posameznike v primeru uporabe umetene inteligence, ki bi lahko vplivala na njihovo dostojanstvo, npr. distribucija lažnih novic, uporaba in distribucija globokih ponaredkov itd. (Zannettou idr. 2019).

Solidarnost je tesno povezana s pravičnostjo in regulacijo v smislu, da naj razvoj umetne inteligence koristi vsem in ne zgolj ožjemu krogu interesentov. Na tej podlagi in v navezavi s koncepti, povezanimi s trajnostjo in vzdržnostjo, je treba zagotoviti solidaren razvoj rešitev in solidarno uporabo rešitev na področju umetne inteligence. Uporaba nečesa, kar bi lahko koristilo družbi kot celoti in posameznikom, naj praviloma ne bi bilo omejena, ampak bi se moralo takšen razvoj spodbujati v dobro vseh. Ker so v določenih družbenih sredinah ali na geografskih področjih resursi lahko omejeni do te mere, da razvoj rešitev umetne inteligence ni možen ali izvedljiv, bi morala biti odgovornost na strani tistih, kjer takšnih omejitev ni, da zagotovijo solidarno udejstvovanje pri razvoju in uporabi sistemov umetne inteligence ter pridobljenih rezultatov (Jaillant 2022).

Slika 6 prikazuje etične koncepte za etično izdelavo, delovanje in uporabo sistemov umetne inteligence.

Slika 6: Etika in umetna inteligenca



Vir: Siau in Wang 2020.

2.4.7 Strojno učenje

Strojno učenje je v osnovi ena od oblik sistema oziroma uporabe umetne inteligence. Murty in Avinash (2023) strojno učenje opredeljujeta kot zrelo in aktivno ter pomembno področje, razvoj katerega traja že več kot šest desetletij. Med drugim je mogoče pripisati hitro rast strojnega učenja tudi temu, da je šele v zadnjem času prišlo do večje razpoložljivosti podatkov, ki jih je možno strojno obdelati, ter izboljšanja in dostopnosti strojne opreme, ki omogoča obdelavo tako velikih količin podatkov.

Podobno kot v primeru več različnih opredelitev umetne inteligence velja v primeru opredelitve strojnega učenja, za katerega obstaja več razlag. Gupta idr. (2024) opredeljujejo strojno učenje hierarhično kot vmesno točko med umetno inteligenco, ki predstavlja krovni okvir oziroma nadokvir za izvajanje, in globokim učenjem, ki predstavlja podokvir strojnega učenja. Nichols idr. (2019) so opredelili strojno učenje kot krovni izraz, ki se nanaša na široko paleto algoritmov za izvajanje inteligentne napovedi na podlagi obdelanih podatkov, ki so pogosto obsežni, morda sestavljeni iz milijonov edinstvenih zapisov. Jordan in Mitchell (2015) in Sarker (2021) navajajo, da je strojno učenje pristop, ki se ukvarja z vprašanjem, kako zgraditi sistem oziroma računalnik, ki se bo samodejno izboljševal skozi izkušnje. Lahko bi rekli, da je strojno učenje pristop, s katerim želimo doseči sistem, ki bi imel sposobnost samostojnega učenja in izboljševanja na osnovi izkušenj.

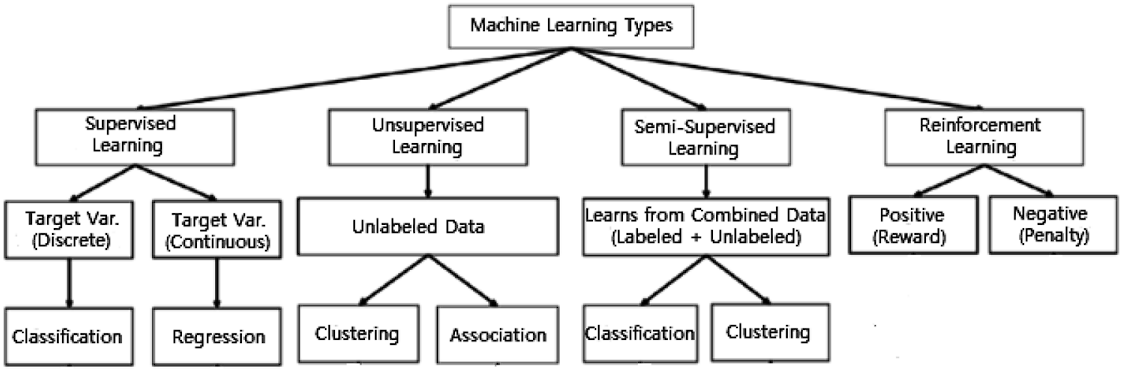
Strojno učenje primarno temelji na obdelavi podatkov, ki omogočajo učenje, zato je pomembno, v kakšni obliki se podatki nahajajo. V osnovi bi lahko podatke razdelili na strukturirane, nestrukturirane, delno strukturirane podatke in opisne ali metapodatke. Poleg tega je pomembna tudi kakovost oziroma ustreznost vhodnih podatkov. Od te je odvisen tudi pristop in pričakovani rezultati, ki se lahko močno razlikujejo, če podatki niso ustrezni oziroma zajemajo veliko nepotrebnih informacij (Caliskan idr. 2017).

Algoritmi strojnega učenja so v glavnem razdeljeni na štiri kategorije oziroma pristope (Sarker 2021):

- nadzorovano učenje, ki predstavlja obdelavo podatkov na podlagi povezljivosti vhodnih in izhodnih podatkov, kjer so na voljo predobdelani podatki, ki so pripravljeni za učenje/trening. Tipičen primer nadzorovanega učenja predstavlja klasifikacija vsebin;
- nenadzorovano učenje predstavlja obdelavo podatkov brez predhodne obdelave podatkov oziroma brez posegov človeka. Pogosto se uporablja za ekstrakcijo generativnih lastnosti, prepoznavanje trendov in struktur, zmanjševanje dimenzionalnosti, iskanje asociacijskih pravil, odkrivanje nepravilnosti itd.;
- delno nadzorovano učenje je kombinacija nadzorovanega in nenadzorovanega pristopa, kar v praksi pomeni, da se pri obdelavi uporablja tako predobdelane podatke kot tudi podatke, ki niso bili predhodno obdelani. Namen je izvajati dodatno kontrolo in učenje ter pridobiti boljše rezultate od tistih, ki so pridobljeni izključno z nadzorovanim pristopom. Delno nadzorovani pristop se uporablja na področju strojnega prevajanja, detekcije goljufij, samodejnega označevanja in klasifikacije vsebin;
- učenje s krepitvijo je pristop strojnega učenja, ki omogoča samodejno vrednotenje optimalnega ravnanja v določenem kontekstu ali okolju z namenom izboljšati učinkovitost. Tovrstno učenje temelji na nagrajevanju oziroma kaznovanju, končni cilj pa je uporaba pridobljenih povratnih informacij, na podlagi katerih se sprejmejo ukrepi za povečanje nagrade ali zmanjšanje tveganj. Je močno orodje za usposabljanje modelov umetne inteligence, ki lahko pomagajo povečati avtomatizacijo ali optimizirati operativno učinkovitost sofisticiranih sistemov, kot so robotika, izvajanje avtonomne vožnje, optimizacija proizvodnje in logistike itd.

Na sliki 7 so prikazani glavni pristopi strojnega učenja z algoritmi.

Slika 7: Pristopi strojnega učenja

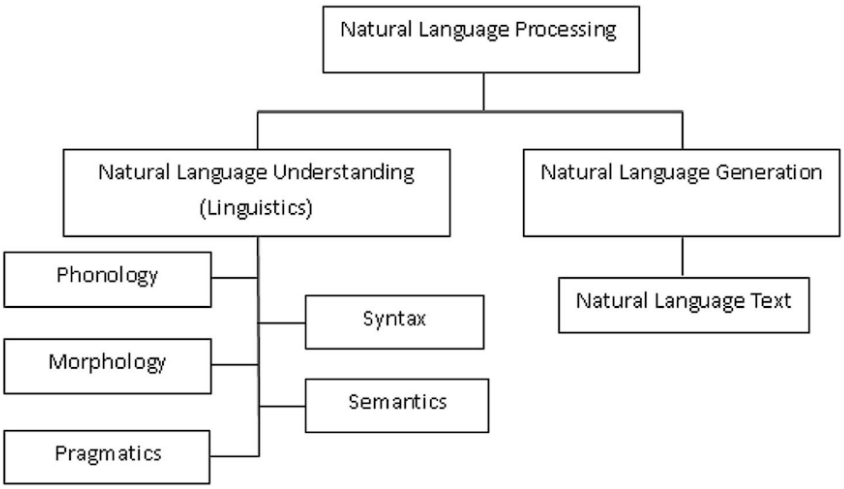


Vir: Sarker 2021.

Eno od področij, kjer se veliko uporablja strojno učenje in je tudi predmet te raziskave, je t. i. obdelava naravnega jezika (angl. Natural language processing – NLP). Eisenstein (2018) obdelavo naravnega jezika neposredno primerja z izrazom računalniško jezikoslovje, vendar izpostavlja, da ne glede na to, da se v precejšnji meri prekrivata, je med njima vseeno določena razlika. V primeru računalniškega jezikoslovja je v središču zanimanja jezikoslovje, kakršnakoli računalniška obdelava pa predstavlja podporno vlogo. V primeru obdelave naravnega jezika je poudarek na načrtovanju in analizi računalniških algoritmov in pristopov za obdelavo naravnega človeškega jezika. Cilj obdelave naravnega jezika je zagotoviti nove računalniške zmožnosti v zvezi s človeškim jezikom: na primer pridobivanje informacij iz besedil, prevajanje med jeziki, odgovarjanje na vprašanja, vodenje pogovora. Khurana idr. (2022) so obdelavo naravnega jezika opredelili kot vejo umetne inteligence in jezikoslovja, namenjeno temu, da bi računalniki razumeli izjave ali besede, zapisane v človeških jezikih. Obdelava naravnega jezika je nastala, da bi uporabniku olajšala delo in zadovoljila željo z računalnikom komunicirati v naravnem jeziku. Nadalje avtorji delijo področje obdelave naravnega jezika na razumevanje naravnega jezika in ustvarjanje naravnega jezika.

Slika 8 prikazuje osnovno razdelitev obdelave naravnega jezika.

Slika 8: Klasifikacija obdelave naravnega jezika



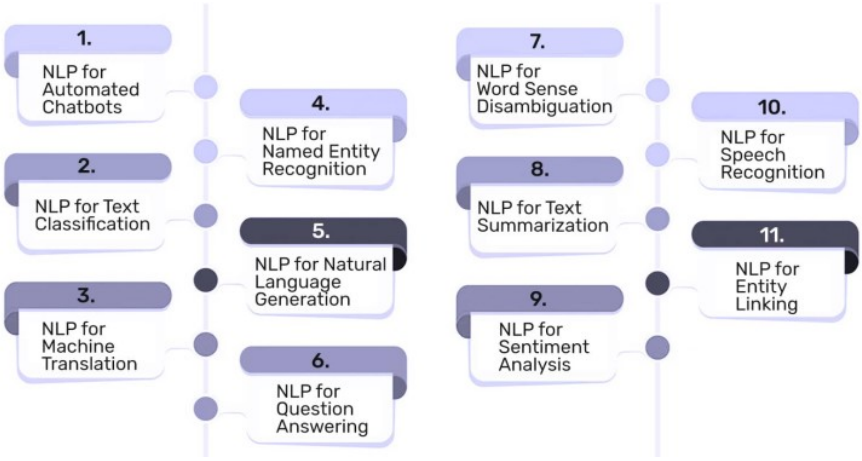
Vir: Khurana idr. 2022.

Primeri uporabe obdelave naravnega jezika z uporabo strojnega učenja so različni. McMullen (2023) je navedel nekaj najbolj tipičnih primerov obdelave naravnega jezika z uporabo strojnega učenja:

- avtomatizirane klepetalnice,
- klasifikacijo besedil,
- strojno prevajanje,
- prepoznavanje imenskih entitet,
- ustvarjanje naravnega jezika,
- odgovarjanje na vprašanja,
- razločevanje pomena besed,
- povzemanje besedila,
- analizo razpoloženja,
- prepoznavanje govora in
- povezovanje entitet.

Na sliki 9 so predstavljeni tipični primeri uporabe strojnega učenja v primeru obdelave naravnega jezika.

Slika 9: Klasifikacija obdelave naravnega jezika



Vir: McMullen 2023.

Predmet raziskave je ugotoviti, ali je možna izdelava izvedbenih modelov za obdelavo nestrukturiranih besedil za nekatere od predvidenih primerov obdelave naravnega jezika. Primeri, ki bodo neposredno zajeti v raziskavo, so klasifikacija besedila, prepoznava imenskih entitet in povzemanje besedila.

Klasifikacija besedil

Kowsari idr. (2019) izpostavljajo, da so bile težave s klasifikacijo besedil v zadnjem času obsežno raziskane in tudi obravnavane v številnih resničnih scenarijih in primerih uporabe. Zaradi napredka v obdelavi naravnega jezika in analizi besedil je pozornost zdaj primarno usmerjena v razvoj aplikativnih sistemov, ki izkoriščajo metode klasifikacije besedil. Večino sistemov za razvrščanje in kategorizacijo besedil je mogoče razdeliti na naslednje štiri faze: ekstrakcija funkcij, zmanjšanje dimenzij, izbira klasifikatorja in ocenjevanje učinkovitosti. Obseg ustvarjanja nestrukturiranih besedil izjemno narašča zlasti v spletnem okolju (Liu idr. 2017) in predstavlja večino vseh besedil, ki se danes ustvarijo. Zaradi velikih količin gradiva se pojavljajo velike težave, kako omenjeno gradivo vsebinsko obdelati oziroma klasificirati za potrebe arhivistike, vključno z morebitno detekcijo, obdelavo in klasifikacijo besedil in objektov iz slik (Tian idr. 2016). Klasifikacija besedil z uporabo strojnega učenja (Luo 2021) tako predstavlja eno od učinkovitih možnih rešitev za reševanje omenjenega problema.

Metod za klasifikacijo besedila je več, od standardnih pristopov, kot je npr. uporaba metode odločitvenih dreves do pristopov z uporabo globokega učenja (Wu idr. 2020) ali z uporabo predizdelanih modelov, ki lahko učinkovito izvajajo klasifikacijo besedil brez dodatnega učenja (Trajanov idr. 2023). Zadnje čase se v praksi za izvedbo klasifikacije vsebine vse več uporabljajo modeli, ki so bili izdelani na podlagi predizdelanih modelov. Ti imajo v osnovi specifično prednost pred izdelavo lastnega samostojnega modela, in sicer, da so bili izdelani iz velikih količin podatkov in z računskimi resursi, ki so v večini primerov za povprečnega uporabnika še vedno nedosegljivi. Omenjeni pristop tako omogoča izdelavo specializiranih ali dodelanih modelov (angl. Fine tuned), ki lahko močno izboljšajo pridobljene rezultate izhodiščnih modelov. Pri tem še vedno gre izdelavo odločitvenega modela na podlagi učenja.

Prepoznava imenskih entitet

Aparna in Srinivasa (2023) prepoznavo imenskih entitet umeščata med pomembnejša področja obdelave naravnega jezika, ki je v zadnjih letih postalo tudi zelo aktivno raziskovalno področje. Naloga prepoznave imenskih entitet je, da samodejno prepozna in razvrsti imenske entitete v besedilu, ki je predmet obdelave. Imenske entitete tako lahko predstavljajo lastna imena, lokacije, organizacije oziroma karkoli, čemur lahko dodelimo neko ključno skupno lastnost. Pri tem se uporablja določanje tipov entitet, velikokrat pa se uporablja tudi izraz razred entitete. Izraze ali količine, kot so valuta, telefonske številke, datum, čas, naslov se ravno tako lahko obravnava kot tipe oziroma razrede imenskih entitet, čeprav ne sodijo neposredno pod opredelitev lastnih imen. Tipe ali razrede je mogoče določiti tudi glede na določeno vsebinsko področje, kot je npr. medicina, kemija itd. Pri identifikaciji imenskih entitet tako obstajata dve ključni nalogi: prepoznava entitete v podanem besedilu v okviru obdelave naravnega jezika in razvrstitev entitete v vnaprej določene tipe oziroma razrede entitet (Anthony idr. 2013).

Ena od pomembnih lastnosti, ki jo je treba upoštevati pri prepoznavi imenskih entitet, je, v katerem jeziku je besedilo, ki je predmet obdelave. Določeni pristopi za izdelavo izvedbenih modelov za prepoznavo imenskih entitet morajo tako upoštevati tudi vhodne podatke in jezik besedila kot ključni vidik učinkovite prepoznave imenskih entitet. Ob tem se pri določenih jezikih lahko pojavijo težave zaradi pomanjkljivih podatkov, ki bi lahko bili uporabljeni za izdelavo modela. V osnovi se za reševanje omenjenega problema, kjer ni na voljo ustreznega obsežnega korpusa za izvedbo učenja oziroma za izdelavo korpusa za učenje, uporabljata dva pristopa, ki lahko olajšata izdelavo modela. Prvi pristop je usmerjen v ustvarjanje indeksiranih podatkov za izvedbo učenja v predvidenem jeziku. Eden od načinov za ustvarjanje indeksiranih podatkov je lahko izvedba prevoda korpusov med izvornim in predvidenim jezikom, spet drug način se lahko izvede s podatkovnim rudarjenjem znotraj javno dostopnih indeksiranih vsebin. Drugi pristop temelji na neposrednem prenosu že izdelanega modela. Pristop sledi ideji, da se izdela le en model, ki je narejen na način, da je neodvisen od lastnosti specifičnega jezika. Tako je možno takšen izdelani model uporabiti za obdelavo besedila, ki je v različnih jezikih (Ni idr. 2017).

Povzemanje besedila

Povzemanje besedil z uporabo strojnega učenja je tako kot klasifikacija besedil in prepoznava imenskih entitet del obdelav naravnega jezika. Karjule idr. (2023) navajajo, da je v današnji digitalni dobi vidna eksponentna rast besedil na internetu in v različnih repozitorijih, kot so novičarski članki, knjige, pravni dokumenti in znanstveni članki. Obseg teh besedil raste vsak dan. Osebno (ročno) branje in povzemanje vsega tega je preprosto prepočasno in nepraktično. Povzemanje besedila kot del obdelave naravnega jezika se je pokazalo kot zmogljiva rešitev za reševanje tega izziva. Z uporabo strojnega učenja je tako možna izdelava modela za povzemanje besedila, ki lahko učinkovito ustvarja kratke povzetke, ki zajamejo bistvo vsebine in hkrati zmanjšajo obseg podatkov za morebitno kasnejšo obdelavo.

Danes se večinoma uporablja abstraktno in ekstrakcijsko povzemanje (Rajendran idr. 2023). Ekstrakcijski in abstraktni pristop tako predstavljata dve osnovni obliki povzemanja informacij. Ekstrakcijsko povzemanje izdvaja podmnožico stavkov iz izvirnega besedila z namenom izdelave povzetka, medtem ko gre pri abstraktnem povzemanju za ustvarjanje kratkega in jedrnatega povzetka, ki zajame bistvene ideje izvornega besedila. Ustvarjeni povzetki abstraktnega povzemanja lahko potencialno vsebujejo nove besedne zveze in stavke, ki morda niso vsebovani v izvirnem besedilu.

Podobno kot v primeru klasifikacije vsebin se v praksi za izvedbo povzemanja lahko neposredno uporabijo predizdelani modeli, ki so bili izdelani na velikih količinah podatkov. Ker je lahko izdelava lastnih modelov na osnovi lastnih podatkov za učenje izjemno problematična predvsem zaradi zagotavljanja ustreznega obsega podatkov in potrebnih resursov za izdelavo takšnega modela, predstavlja uporaba predizdelanih modelov odlično izhodišče za izvedbo povzemanja (Abdel-Salam in Rafea 2022). Pri tem lahko večjezičnost vsebine tako kot v primeru prepoznave imenskih entitet vpliva na uspešnost izvedenega povzemanja, zato je priporočljivo, da se uporabi predizdelane modele, ki podpirajo večjezično vsebino oziroma so bili izdelani na podlagi vsebine, ki je v jeziku, v katerem se nahaja besedilo, ki je predvideno za izvedbo povzemanja.

Izdelava odločitvenih modelov z uporabo strojnega učenja

Izdelava tipičnega celotnega izvedbenega in/ali odločitvenega modela na podlagi učenja lahko vključuje naslednje ključne korake (Murty in Avinash 2023; Luo idr. 2016; Medvedeva idr. 2020; Vinyals idr. 2015):

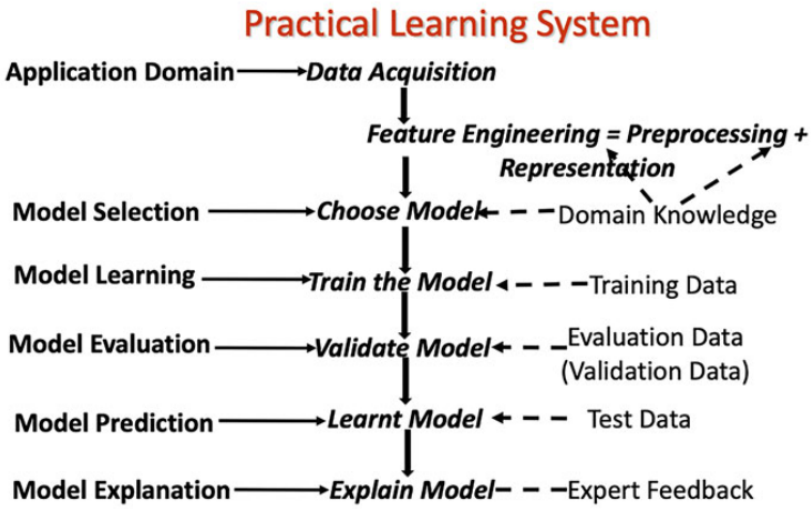
- *pridobivanje podatkov*. V večji meri je korak pomemben predvsem zaradi poznavanja vhodnih vrednosti in povezane problematike, ki jo želimo rešiti z izdelavo modela, kar je ključnega pomena za nadaljnje korake. Pri tem je pomembno, da se upošteva potencialna kakovost podatkov;
- *razvoj in določanje značilnosti modela*. Ta korak vključuje kombinacijo predhodnih obdelav podatkov, kot so določanje reprezentativnega obsega ali npr. zmanjšanje dimenzij ali predhodno urejanje podatkov z namenom zagotavljanja višje kakovosti podatkov. Pri tem se lahko pojavijo tri različne težave: manjkajoči podatki, velika razdrobljenost in različnost podatkov ter potencialno izstopajoči podatki, npr. nepotrebni podatki (šum ali hrup med podatki, ki jih dejansko želimo);
- *izbira metode*. Izbira metode je v prvi vrsti odvisna od metodologije, ki jo želimo uporabiti za izdelavo modela. Na primer če želimo obdelati izključno numerične podatke, so določene metode primernejše kot tiste, ki so specializirane le za obdelavo teksta;
- *izdelava modela*. Ta korak se v praksi velikokrat imenuje tudi učenje modela. Ker gre večinoma za računsko intenzivne operacije, ta korak z vidika resursov predstavlja enega od najbolj obremenjujočih korakov v celotnem procesu izdelave modela. Čas in izvedljivost izdelave modela sta tako precej odvisna od razpoložljivih resursov (strojna oprema) in obsega ter vrste podatkov, ki so uporabljeni za izdelavo modela. V tem koraku se praviloma izvaja preverjanje učinkovitosti izdelanega modela z

razdvojitvijo podatkov na del, ki se nanaša na izdelavo modela, in del, ki se nanaša na preverjanje učinkovitosti delovanja izdelanega modela;

- *ocenjevanje učinkovitosti modela*. Ta korak se imenuje tudi validacija modela. Zanj je za razliko od prejšnjega koraka potrebno ločeno preverjanje delovanja izdelanega modela na ločenih podatkih, ki so bili predhodno že obdelani z namenom potrjevanja uspešnosti uporabe izdelanega modela;
- *razlaga modela*. Korak je pomemben predvsem za pridobivanje povratnih informacij od oseb, katerim je uporaba takšnega modela namenjena. Oni so tisti, ki bodo najlaže in najstrokovneje ocenili pridobljene rezultate. V primeru pozitivnih povratnih informacij se izvede potrditev ustreznosti izdelanega modela.

Na sliki 10 so prikazani ključni koraki za izdelavo modela za obdelavo podatkov na podlagi strojnega učenja.

Slika 10: Ključni koraki za izdelavo modela za obdelavo podatkov na podlagi strojnega učenja



Vir: Murty in Avinash 2023.

2.4.8 Kakovost informacij

Kakovost informacij je na področju strojnega učenja eno od pomembnih področij, ki se velikokrat spregleda, lahko pa, odvisno od izbranega pristopa in metode, ključno vpliva na izdelavo odločitvenih modelov in rezultate, ki so pridobljeni z uporabo takšnega modela. Pri uporabi definicije informacij se teh ne sme mešati z definicijo podatkov.

Rožanec (2017, 12) je podatek opredelila z naslednjo definicijo:

- »Podatek je poljubna predstavitev s pomočjo simbolov ali analognih veličin, ki ji je pripisan ali se ji lahko pripiše pomen.

- Podatek je predstavitev dejstva, koncepta ali instrukcije na formaliziran način, ki je primeren za komunikacijo, interpretacijo ali obdelavo s strani človeka ali stroja.
- Podatki so dejstva, predstavljena z vrednostmi (številke, znaki, simboli), ki imajo pomen v določenem kontekstu.«

»Vsem trem definicijam pa je skupno, da se podatku lahko pripiše nek pomen na osnovi predpisa oziroma znotraj nekega konteksta. Podatek je tako le nosilec informacije oziroma njegova fizična predstavitev« (12).

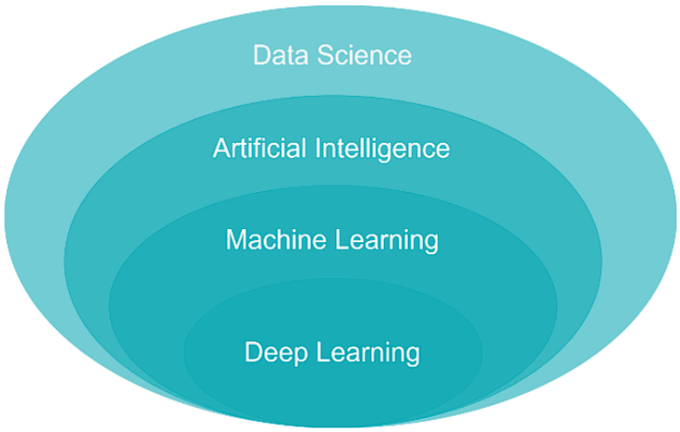
Težave s kakovostjo podatkov in informacij se praviloma pojavijo že v fazi pridobivanja podatkov, kjer zgolj razpoložljivost podatkov še ne pomeni tudi ustrezne kakovosti podatkov in informacij (v nadaljevanju poenotena opredelitev »podatki«). Omenjene težave z vidika kakovosti podatkov se pojavijo zlasti pri delu z odprtimi podatki, s podatki, ki jih ustvarijo uporabniki, ali podatki, ki prihajajo iz več različnih virov. Posledično se lahko pojavi potreba po predhodni obdelavi oziroma. urejanju podatkov, pri čemer se izvede proces zagotavljanja kakovosti podatkov, kot so urejanje manjkajočih podatkov, urejanje podvojenih podatkov, upravljanje z mejnimi vrednostmi, normalizacija, standardizacija, odstranjevanje nepotrebnih podatkov itd. (Diericx idr. 2023).

Na področju umetne inteligence se uporablja velike količine t. i. velikih podatkov (angl. Big data), ki jih je praktično nemogoče urejati ročno. Taleb idr. (2021) velike podatke opredeljujejo kot univerzalne podatke, sestavljene iz velikih količin podatkov z nekonvencionalnimi vrstami. Te vrste podatki so lahko strukturirani, nestrukturirani ali v dinamični obliki. Ne glede na to, ali so ustvarjeni oziroma uporabljeni na specifičnem področju, je za podporo odločitvam, ki temeljijo na podatkih, nujen nov način ravnanja z velikimi podatki tako s tehnološkega vidika za zagotavljanje raziskovalnih pristopov kot pri njihovem upravljanju (Elouataoui idr. 2022).

Ker upravljanje s podatki in uporaba podatkov predstavljata predpogoj oziroma nadokvir za uporabo umetne inteligence, ki v osnovi temelji na obdelavi podatkov, je treba posebno pozornost posvetiti tudi tej temi. Pomanjkanje visokokakovostnih podatkov za izvedbo učenja pri izdelavi odločitvenih modelov lahko predstavlja precejšnjo omejitev za uporabo strojnega učenja, zlasti globokega učenja (Chen idr. 2022).

Na sliki 11 sta prikazani odvisnost in povezanost med področjem upravljanja s podatki in področjem umetne inteligence in bolj podrobno strojnega učenja.

Slika 11: Povezanost med področjem upravljanja s podatki in področjem umetne inteligence



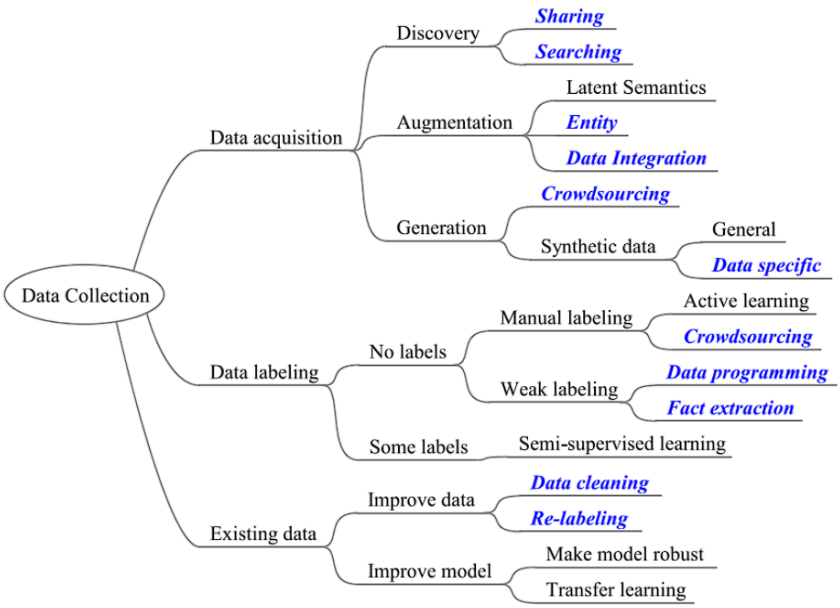
Vir: Chen idr. 2022.

Chen idr. (2022) nadalje opredeljujejo, da je za zagotavljanje kakovosti podatkov treba zagotoviti strukturiran in sistematičen pristop, ki so ga v fazi odločitve o uporabi podatkov razdelili na tri različne veje:

- uporaba obstoječega niza podatkov brez dodatne izvedbe procesa urejanja podatkov,
- uporaba obstoječega niza podatkov z dodatno izvedbo procesa urejanja podatkov in
- izdelava novega niza podatkov.

Slika 12 prikazuje visokonivojski pristop za zbiranje in obdelavo podatkov, ki se lahko uporabijo za izdelavo odločitvenih modelov z uporabo strojnega učenja.

Slika 12: Visokonivojski pristop za zbiranje in obdelavo podatkov



Vir: Chen idr. 2022.

Določen izziv se pojavlja na strani potrjevanja ustreznosti in kakovosti podatkov, ko so izvedeni vsi procesi za zagotavljanje kakovosti podatkov. Možnosti za preverjanje (validacijo) kakovosti podatkov po izvedenih procesih (npr. samodejno indeksiranje) je več, od vzorčenja do križnih primerjav itd. Glede na to, da se lahko kakovost podatkov interpretira na več načinov (npr. potencialna identifikacija nepravilnih podatkov ali identifikacija manjkajočih podatkov), se tudi pristopi za izvedbo validacije ustreznosti lahko razlikujejo, zato je treba k izvedbi validacije pristopiti na način, da bodo upoštevani vsi kriteriji, ki so bili identificirani kot problematični z vidika kakovosti podatkov.

Lastnosti kakovosti podatkov so ključne za identifikacijo kriterijev, na podlagi katerih lahko ocenjujemo kakovost podatkov. Lastnosti kakovosti podatkov lahko v osnovi razdelimo na naslednje kategorije (European Commission 2014):

- Točnost: ali podatek pravilno predstavlja realni svet ali dogodek?
- Konsistentnost: ali podatek morda ne vsebuje nasprotij?
- Razpoložljivost: ali je do podatka mogoče dostopati v tem trenutku in tudi čez nekaj časa?
- Popolnost: ali podatek vsebuje vse dele podatka, ki predstavljajo en subjekt ali dogodek?
- Skladnost: ali je podatek v skladu s sprejetimi standardi?
- Verodostojnost: ali podatek temelji na verodostojnih virih?
- Zmožnost obdelave: ali je podatek strojno berljiv?
- Relevantnost: ali podatki vsebujejo ustrezno količino podatkov?
- Pravočasnost: ali podatek predstavlja dejansko situacijo in ali je objavljen dovolj hitro?

Pri zagotavljanju kakovosti podatkov je treba zagotoviti ustrezen pregled nad podatki, kar se lahko izvede tudi skozi njihovo profiliranje, tj. z analiziranjem nabora podatkov in zbiranjem podatkov o podatkih (metapodatkov) z uporabo številnih različnih tehnik. Zato je bistvena naloga pred kakršnokoli meritvijo kakovosti ali obdelavo podatkov pridobiti vpogled v dani nabor podatkov. Primeri informacij, ki se zbirajo med profiliranjem podatkov, so tako lahko število različnih ali manjkajočih (tj. ničelnih) vrednosti, tipi atributov ali pojavljajoči se vzorci in njihova pogostost (Ehrlinger in Wöß 2022).

Da bi dosegli ustrezno kakovost podatkov, ki so namenjeni učenju pri uporabi strojnega učenja, je tako treba izbrati pristop, ki bo vključeval tudi morebitne vplive nekakovostne vsebine, če kakovosti podatkov v predvideni obliki ni možno zagotoviti. V izogib takšnim situacijam je bil kot ena od rešitev razvit pristop z izdelavo in predučenjem t. i. maskiranih jezikovnih modelov (angl. Masked language model), ki je za izdelavo modelov predvidel postopek prekrivanja določenega dela podatkov z namenom, da model samostojno predvidi, za kakšne podatke gre. Praviloma se takšni modeli učijo/trenirajo na velikih količinah podatkov z namenom, da se pri učenju bolje razume kontekst uporabljenih besed in ugotovi ustreznost podatkov na podlagi omenjenega konteksta (Salazar idr. 2020; Zhou idr. 2022).

Maskirani jezikovni modeli zagotavljajo odlične rezultate pri številnih nalogah v okviru obdelave naravnega jezika. Kot izpostavljata Kaneko in Bollegala (2022), pa lahko

maskirani jezikovni modeli kažejo tudi zelo zaskrbljujočo stopnjo družbene pristranskosti oziroma se lahko pojavi precejšen odklon v smeri diskriminacije. Avtorja poudarjata, da so predhodno predlagane metrike vrednotenja za kvantifikacijo družbenih pristranskosti in diskriminacije pri uporabi maskiranih jezikovnih modelov problematične, ker: (1) je natančnost napovedi maskiranih podatkov v nekaterih maskiranih jezikovnih modelih običajno nizka, kar vodi do nezanesljivih metrik vrednotenja; (2) se pri večini nadaljnjih nalog z vidika obdelave naravnega jezika maskiranje več ne uporablja, zato takšna predikcija ni več neposredno povezana z izvirnim maskiranjem; (3) so pogoste besede v izvirnih podatkih za učenje pogosteje maskirane, kar povzroča šum zaradi izbirne pristranskosti.

Kakovost podatkov je tako pomembna predvsem zaradi izbire pristopa za izdelavo modela. V primeru, da gre za izdelavo celotnega modela brez uporabe predizdelanih modelov, bo imela kakovost podatkov ključen vpliv na izdelavo odločitvenega izvedbenega modela in pridobljene rezultate. Če je izbran pristop za izdelavo končnega modela z uporabo predizdelanega maskirnega jezikovnega modela, potem se lahko izkaže tudi, da kakovost podatkov ne bo ključno vplivala na izdelavo odločitvenega izvedbenega modela in pridobljene rezultate, vse pa je odvisno tudi od izbranega predizdelanega maskirnega jezikovnega modela.

2.5 Ocena dosedanjih raziskovanj na obravnavanem področju

Klasifikacija gradiva

Na področju uporabe strojnega učenja pri klasifikaciji nestrukturiranega gradiva je bilo v preteklosti izvedenih več raziskav predvsem v smeri ugotavljanja in določanja specifične kategorizacije vsebine, vse od namenske prepoznave in ločevanja nezaželene elektronske pošte (Dada idr. 2019), identifikacije sumljivih, sovražnih in nevarnih vsebin (Sharif 2020) pa do izdelav tematskih modelov v podporo končni klasifikaciji gradiva (Jelodar idr. 2013; Agrawal idr. 2018; Gao idr. 2015; Onan 2019; Hashimoto 2016). Izdelava tematskih modelov se je skozi čas izkazala za izjemno učinkovit pristop za celovito vsebinsko klasifikacijo gradiva, saj omogoča uporabo tako nadzorovanega ali nenadzorovanega pristopa k določanju obsega in števila (Zhao idr. 2015) kot tudi hierarhičnega povezovanja tematskih modelov (Yang idr. 2016). Suominem in Toivanen (2016) sta v svoji raziskavi preverjala obliko nenadzorovanega učenja za izvajanje samostojne klasifikacije na podlagi izdelave tematskih modelov in na podlagi izdelanih tematskih modelov izdelavo zemljevida, ki prikazuje strukturo, obseg in povezanost finske znanosti prek analize bibliometričnih podatkov (člankov), pri čemer je vsaj eden od avtorjev povezan z državo Finsko. Wang idr. (2019) in Hartmann idr. (2019) so raziskovali različne metode, vključno z metodo za izdelavo tematskih modelov za doseganje čim boljših rezultatov na področju klasifikacije besedil na podlagi šibko indeksirane oziroma pomanjkljivo indeksirane vsebine, ki so jo uporabili kot podlago za strojno učenje. Na področju vsebinske klasifikacije nestrukturiranih zdravstvenih besedil so Kreimeyer idr. (2017) izvedli raziskavo z namenom identificirati in izdelati čim bolj standardizirano obliko oziroma strukturirano obliko relevantnih informacij, ki bi lahko bile v operativno pomoč zdravstvenemu osebju. V zadnjem času se v praksi

veliko uporablja predizdelani model BERT NLP (González-Carvajal in Garrido 2023), ki prednjači predvsem v tem, da lahko izjemno učinkovito podpre obdelavo različnih podatkov.

Patel in Gadhavi (2017) sta v svoji raziskavi proučevala možnosti vsebinske klasifikacije z uporabo različnih metod strojnega učenja, vključno z izdelavo semantičnih modelov, pri kateri izpostavljata predvsem učinkovitost različnih metod strojnega učenja za doseganje najboljših rezultatov pri samodejni vsebinski klasifikaciji različnih vsebin in digitalnih zapisov. Arras idr. (2017) so raziskovali, na kakšen način se izvaja samodejna klasifikacija v primeru uporabe strojnega učenja, pri čemer so se opirali na tehniko za razčlenitev odločitvenih modelov do njihovih temeljnih gradnikov LRP (angl. Layer-wise relevance propagation). Rāts in Pede (2020) sta v svoji raziskavi pristopila k izdelavi aplikativnega izvedbenega modela za klasifikacijo besedil z namenom vzpostavitve integralne funkcije obdelave besedil v okviru sistemov za upravljanje vsebin.

V povezavi z vsebinsko klasifikacijo so bile določene raziskave usmerjene v optimizacijo in doseganje čim bolj natančnih oziroma uspešnih rezultatov z uporabo in kombiniranjem obstoječih metod. Bennasar idr. (2015) so preverjali učinkovitost dveh nelinearnih metod JMIM (angl. Joint Mutual Information Maximisation) in NJMIM (angl. Normalised Joint Mutual Information Maximisation), ki uporabljata t. i. skupne informacije in maksimalni izkoristek glede na minimalni obseg informacij z namenom doseganja čim boljših in točnejših rezultatov pri vsebinski klasifikaciji. Podobno raziskavo so izvedli Bottou idr. (2018) z uporabo Convex optimizacije. Da bi dosegli določene zadovoljive rezultate, je pomembna tudi optimizacija pri hitrosti obdelave podatkov predvsem v razmerju kakovosti in točnosti obdelave v navezavi s potrebnimi viri za obdelavo. Približati je torej treba uporabo strojnega učenja in možnost vsebinske klasifikacije na obsežnih podatkovnih virih z uporabo standardnih delovnih postaj (Joulin idr. 2017) ali z uporabo ustreznih metod (Canedo in Mendes 2020). Določeni izzivi se kažejo tudi v učinkovitosti obstoječih metod pri vsebini, ki se nahaja v različnih jezikih, in s tem povezano učinkovitostjo prepoznave in ustrezne klasifikacije vsebine (Howard in Ruder 2018).

Prepoznava imenskih entitet

Na področju identifikacije in ekstrakcije vsebinskih gesel (imenskih entitet) z uporabo strojnega učenja za potrebe popisovanja gradiva so bile izvedene predvsem parcialne raziskave, ki so proučevale, ali je izvedba takšne identifikacije izvedljiva in v kolikšni meri je učinkovita.

Štajner idr. (2013) so proučevali možnost identifikacije imenskih entitet v slovenskem jeziku z uporabo nadzorovanega učenja s pogojnimi naključnimi polji CFR (angl. Conditional Random Fields). V raziskavi so poskušali identificirati osebna, zemljepisna, organizacijska in stvarna imen z uporabo leksikonov in brez nje. Pri prepoznavi imenskih entitet je bilo ugotovljeno, da sistem učinkovito identificira vsa imena razen stvarnih imen. Ljubešić idr. (2012) so v raziskavi predstavili izgradnjo modelov za prepoznavo imenskih entitet in samodejno klasifikacijo vsebine za hrvaški in slovenski jezik ter kakšen vpliv ima količina

podatkov oziroma uporaba morfosintaktičnih ali oblikoslovnih oznak na učinkovitost učenja in prepoznave. Weston idr. (2019) so proučevali ekstrakcije in normalizacijo imenskih entitet na podlagi obdelave večjih podatkov znanstvene literature z namenom vzpostavitve strukturiranega podatkovnega skladišča, ki bi lahko bilo namenjeno programskim poizvedbam. V raziskavi so identificirali več kot 80 milijonov imenskih entitet, ki so bile shranjene v strukturiranem podatkovnem skladišču, znotraj katerega je možno izvajati kompleksne metapoizvedbe. Dernoncourt idr. (2017) so skozi prepoznavo imenskih entitet raziskovali možnosti za anonimizacijo osebnih podatkov v zdravstvenih zapisih pacientov z namenom izkoriščanja ostalih podatkov za izvedbo raziskav na področju medicine.

Samodejna identifikacija in upravljanje z imenskimi entitetami se podobno kot samodejna klasifikacija vsebine sooča s specifičnimi izzivi, kot so npr. jezikovne spremembe in različno izrazoslovje (Augenstein idr. 2017), povezovanje oziroma vzpostavitev relacij med identificiranimi imenskimi entitetami (Derczynski idr. 2015), NEL (angl. Named Entity Linking), identifikacija imenskih entitet z namenom izgradnje baze znanj (Eftimov idr. 2017), identifikacija imenskih entitet v različnih jezikih (Mayhew idr. 2017), povezava metode MD (angl. Mention Detection) z identifikacijo imenskih enititet (Xu idr. 2017), zagotavljanje ustreznih predoznačenih podatkov za izvedbo strojnega učenja (Peters idr. 2017; Ficek idr. 2022) ter identifikacija in ločevanje ugnezdenih imenskih entitet (Sun idr. 2019).

Samodejno povzemanje besedil

Identifikacija in povzemanje ključnih informacij za izdelavo individualnega naslova popisne enote (individualnega zapisa) je področje, ki predstavlja največji izziv za pridobivanje kakovostnih rezultatov, saj na tem področju ni bilo izvedenih veliko raziskav. Conneau idr. (2017) so v svoji raziskavi izdelali model za nadzorovano reprezentacijo stavkov, ki omogoča podlago za naknadno pomensko vrednotenje celotnih stavkov. Ena od možnih metod za doseganje omenjenega cilja je tudi uporaba samodejnega povzemanja (angl. summarization) do oblike, ki bi lahko uspešno ponazorila naslov individualne popisne enote na podlagi vsebinske analize individualnega zapisa (Afantenos idr. 2005; Gui idr. 2018; Kurian in Sheena 2020; Tabak in Vesile 2020).

Andhale in Bewoor (2016) sta raziskovala različne oblike samodejnega povzemanja, vključno z učinkovitostjo in različnimi pristopi specifično v delu ekstrakcijskega in abstraktnega povzemanja.

3 EMPIRIČNI DEL

3.1 Namen in cilji raziskovanja

Namen raziskave je raziskati možnost samodejne klasifikacije in identifikacije različnih vsebin v digitalni obliki z uporabo strojnega učenja. Namen in cilj raziskave je preveriti, ali je z uporabo strojnega učenja možno vzpostaviti model za masovno urejanje in popisovanje nestrukturiranih zapisov, oziroma natančneje, ali je možno z uporabo strojnega učenja samodejno izdelati uporabno in razumljivo individualno popisno enoto.

Ugotovitve raziskave naj bi zagotovile podlago za uporabo strojnega učenja pri urejanju in popisovanju nestrukturiranih besedil neposredno pri tistih, ki takšne rešitve glede na identificirani problem potrebujejo.

Cilj raziskave je vzpostaviti delujoč teoretični in aplikativni izvedbeni model za samodejno masovno urejanje in popisovanje nestrukturiranih zapisov ter skušati odgovoriti na naslednja vprašanja:

- Kako različna pojavna oblika posameznih digitalnih zapisov vpliva na možnost vsebinske analize z uporabo strojnega učenja?
- Kako nestrukturiranost podatkov vpliva na možnost vsebinske analize z uporabo strojnega učenja?
- Kakšna je učinkovitost masovnega urejanja in popisovanja nestrukturiranih zapisov z uporabo strojnega učenja?

Rezultati disertacije omogočajo podrobnejši vpogled v možnosti urejanje in popisovanja nestrukturiranih besedil z uporabo strojnega učenja na področju arhivistike.

Rezultati raziskave smiselno zaokrožujejo:

- temeljit vsebinski pregled oblik in možnosti klasifikacije vsebin nestrukturiranih besedil z uporabo strojnega učenja,
- temeljit vsebinski pregled oblik in možnosti pristopov k identifikaciji imenskih entitet z uporabo strojnega učenja,
- vsebinski pregled oblik in možnih pristopov k ekstrakciji informacij, potrebnih za izdelavo naslova individualne popisne enote,
- izdelavo teoretičnega modela urejanja in popisovanja nestrukturiranih besedil z uporabo strojnega učenja, ki bo vključeval popis procesov, delotokov in posameznih aktivnosti, potrebnih za samostojno urejanje in popisovanje nestrukturiranih besedil,
- temeljit pregled, kateri dejavniki ključno vplivajo na kakovost pridobljenih rezultatov,
- razvoj aplikativnega izvedbenega modela za urejanje in popisovanje nestrukturiranih besedil z uporabo strojnega učenja,
- popis procesov, delotokov in posameznih aktivnosti ter izbiro in opredelitev delovanja specializiranih orodij, ki bodo predstavljeni kot navodila oziroma

priporočila za samostojno izdelavo aplikativnega modela za urejanje in popisovanje nestrukturiranih besedil z uporabo strojnega učenja.

3.2 Raziskovalne hipoteze, raziskovalna vprašanja

V okviru raziskave pri razvoju modela urejanja in popisovanja nestrukturiranih besedil z uporabo strojnega učenja se postavljajo naslednja raziskovalna vprašanja:

- Ali je v primeru nestrukturiranih besedil za namen popisovanja možno izdelati model samodejne klasifikacije vsebin z uporabo strojnega učenja?
- Ali je v primeru nestrukturiranih besedil za namen popisovanja možno izdelati model samodejne identifikacije imenskih entitet z uporabo strojnega učenja?
- Ali je v primeru nestrukturiranih besedil za namen popisovanja možno izdelati model izdelave naslova individualne popisne enote z uporabo strojnega učenja?

Hipoteze

Skladno z opredeljenim raziskovalnim problemom, izpostavljenimi raziskovalnimi vprašanji ter predvidenim pristopom k razvoju modela urejanja in popisovanja nestrukturiranih besedil z uporabo strojnega učenja so v nadaljevanju oblikovane naslednje hipoteze:

H1: Izdelava modela samodejne klasifikacije vsebin z uporabo strojnega učenja v primeru prepoznavanja nestrukturiranih besedil je izvedljiva in uporabna za potrebe popisovanja individualnih zapisov.

H2: Izdelava modela samodejne identifikacije imenskih entitet z uporabo strojnega učenja v primeru prepoznavanja nestrukturiranih besedil je izvedljiva in uporabna za potrebe popisovanja individualnega zapisa.

H3: Izdelava modela samodejne izdelave naslova individualne popisne enote z uporabo strojnega učenja v primeru prepoznavanja nestrukturiranih besedil je izvedljiva in uporabna za potrebe popisovanja individualnega zapisa.

3.3 Raziskovalna metodologija

V okviru vsebinske analize bo analizirana in pregledana tako tuja kot domača literatura s področja urejanja in popisovanja dokumentarnega gradiva, strojnega učenja, identifikacije in upravljanja z imenskimi identitetami, obvladovanja informacij ter upravljanja z digitalnimi zapisi. S kvantitativno metodo eksperiment bo preizkušena vzpostavitev testnega aplikativnega modela za vsebinsko analizo in klasifikacijo posamičnih identificiranih digitalnih zapisov ter ekstrakcijo imenskih entitet za namen popisovanja identificiranega gradiva, vključno z ekstrakcijo informacij, potrebnih za izdelavo naslova popisne enote.

3.3.1 Metode in tehnike zbiranja podatkov

V raziskavi bosta uporabljena vsebinska analiza kot kvalitativna metoda in eksperiment kot kvantitativna metoda.

Informacije o področju urejanja in popisovanja dokumentarnega gradiva, strojnega učenja, identifikacije in upravljanja z imenskimi identitetami, obvladovanja informacij ter upravljanja z digitalnimi zapisi bodo primarno pridobljene iz namenskih podatkovnih zbirk COBISS, Web of Science, Scopus in ProQuest (2024), ki vsebujejo različne serijske in monografske publikacije ter ostale vire. Poleg uporabe omenjenih podatkovnih zbirk bo izvedeno dodatno iskanje literature prek spleta oziroma v znanstvenih raziskavah, strokovnih člankih in ostalih prosto dostopnih informacijah, ki se nahajajo na spletu.

Vsa pridobljena literatura in vsebina bo analizirana in pregledana, vsebina predhodnih raziskav pa bo smiselno uporabljena med pripravo te raziskovalne naloge. Za podajanje dejstev in opredeljevanje dejstev bo uporabljena opisna metoda, za ugotavljanje, primerjavo različnih virov in konkretizacijo ter analizo zaključkov pa primerjalna metoda.

Eksperiment (aplikativni model) bo vključeval najmanj 50.000 vhodnih enot individualnih nestrukturiranih zapisov, od katerih bo naključnih 70 % individualnih zapisov uporabljenih za izdelavo modelov na podlagi strojnega učenja in naključnih 30 % individualnih zapisov kot kontrolna skupina, ki bo uporabljena za preverjanje uspešnosti izdelanih modelov. Eksperiment bo vključeval tudi uporabo predizdelanih modelov za strojno učenje.

Informacije in podatki za izdelavo testnega aplikativnega modela in izvedbo eksperimenta za vsebinsko analizo in klasifikacijo posamičnih identificiranih digitalnih zapisov ter vzporedno ekstrakcijo imenskih entitet za namen popisovanja identificiranega gradiva bodo pridobljeni iz namenskih korpusov za potrebe raziskovanja na področju upravljanja s podatki oziroma informacijami (KAGGLE 2024; The Guardian Open Platform 2022; CLARIN.SI 2022).

The Guardian Open Platform (2022) je odprta platforma in javna spletna storitev za dostop do vseh vsebin, ki jih objavlja medijska hiša The Guardian (2022). Podatki so razvrščeni po oznakah in vsebinskih sklopih. Platforma omogoča dostop do vsebine več kot 1,9 milijona člankov za neprofitno uporabo ter za namene raziskovanja, potrebe študija in razvoj neprofitnih informacijskih rešitev.

»CLARIN.SI (2022) je slovenski nacionalni konzorcij v mreži evropske raziskovalne infrastrukture CLARIN (2022). Raziskovalcem na področju humanistike, družboslovja in drugih z jezikom povezanih ved zagotavlja jezikovne vire in tehnologije ter strokovno podporo in prenos znanja. Na ta način želi razširiti tehnološke možnosti raziskovanja jezikov, predvsem slovenščine in drugih južnoslovanskih jezikov, ter spodbujati meddisciplinarno sodelovanje« (CLARIN.SI 2022).

V okviru raziskave in izdelave teoretičnega in aplikativnega modela bo uporabljena izključno le vsebina člankov, ne pa tudi morebitni povezani metapodatki individualnih člankov. Omenjeni viri so bili izbrani zaradi raznolikosti vsebine, ki jih korpusi in podatkovne zbirke vsebujejo, z namenom preverjanja delovanja aplikativnega modela na čim širšem vsebinskem vzorcu.

3.3.2 Opis instrumentarija

Predvideni koncept oziroma instrument raziskave za razvoj teoretičnega in aplikativnega modela bo postavljen na osnovi predhodno izvedenih raziskav ter lastnih idejnih razvojnih konceptov skladno s smernicami in mednarodnimi standardi na področju urejanja in popisovanja arhivskega gradiva.

3.3.3 Opis vzorca

Izhodišče za raziskavo je postavljeno na predpostavki, da je izdelava modela urejanja in popisovanja nestrukturiranih besedil z uporabo strojnega učenja izvedljiva in da so na voljo orodja, ki bi omogočala razvoj in uporabo takšnega modela. Poleg izvedbe procesa klasifikacije vsebine in prepoznave imenskih entitet smo predpostavili tudi, da je izvedljiva izdelava opisne enote v obliki, ki je uporabna za namen arhiviranja oziroma zagotavljanja dolgoročne hrambe.

Predpostavke pri izdelavi modela urejanja in popisovanja nestrukturiranih besedil z uporabo strojnega učenja so:

- vsebina gradiva je v nestrukturirani obliki, tj. v obliki, ki nima vnaprej določene ali predvidene strukture,
- gradivo je v digitalni obliki, ki omogoča strojno obdelavo oziroma je strojno berljivo,
- vsebina gradiva vnaprej ni znana, ravno tako pri posameznem zapisu niso na voljo metapodatki, ki bi omogočali neposredno izdelavo popisne enote,
- gradivo, ki je namenjeno obdelavi podatkov, je v angleškem in slovenskem jeziku,
- ostali popisni elementi z izjemo imenskih entitet in naslova so znani, fiksni in povezani s pričakovano klasifikacijo gradiva.

Omejitve pri izdelavi modela urejanja in popisovanja nestrukturiranih besedil z uporabo strojnega učenja se nanašajo predvsem na pridobivanje podatkov za izdelavo modela in zagotavljanje njihovega dovolj velikega obsega, da bi model lahko dosegel pričakovane rezultate.

Določena težava, ki se kaže pri raziskovanju teme uporabe strojnega učenja in splošnega koncepta uporabe umetne inteligence na področju arhivistike je precejšnje pomanjkanje ustrezne literature in predhodnih raziskav, zlasti z vidika arhivistike in dokumentologije.

3.3.4 Opis obdelave podatkov

V primeru obdelave podatkov smo se naslonili na predhodno raziskavo s področja samostojne klasifikacije elektronskih sporočil (e-pošte) z uporabo strojnega učenja (Milovanović 2020), ki je pokazala učinkovito uporabo dostopnih orodij za strojno učenje.

Aplikacijsko orodje, ki je bilo uporabljeno v raziskavi, je KNIME (2024).

V nadaljevanju so predstavljene osnovne specifikacije strojne opreme (delovne postaje), ki je bila uporabljena za obdelavo podatkov:

- procesor z osmimi jedri (16 niti),
- delovni spomin (32 GB),
- trdi disk (NVME 1 TB),
- grafična kartica z možnostjo obdelave podatkov za namen strojnega učenja NVIDIA Geforce GTX 1060 – 6 GB (NVIDIA 2023).

Namen raziskave je bil tudi preveriti, ali so vse naloge izvedljive z uporabo povprečne delovne postaje, ki ne potrebuje dodatnih investicij, in kateri del obdelave podatkov je časovno izvedljiv tudi brez uporabe grafične kartice. Uporaba jeder grafične kartice bi bila aktivirana le v primeru delotokov in aktivnosti, ki z uporabo procesorskih jeder delovne postaje ne bi mogle biti izvedene v roku 72 ur od začetka obdelave podatkov v posamezni veji delotoka.

Da bi lahko preverili izvedljivost korakov z uporabo orodja KNIME (2024), je bilo treba razviti procese in vzpostaviti aktivnosti, ki bi vključevali obravnavo vseh predvidenih postopkov za doseganje zastavljenih ciljev. V orodju KNIME (2024) so bili posledično izdelani naslednji delotoki:

- delotok za učenje in izdelavo modela za prepoznavanje imenskih entitet,
- delotok za prepoznavanje imenskih entitet na podlagi predizdelanega splošnega modela.
- krovni delotok za izdelavo tematskih modelov in z izvedbo samostojne klasifikacije nestrukturirane vsebine:
    - delotok za analizo vsebine in identifikacijo individualnih kategorij ter izdelavo tematskih modelov (Arora idr. 2020, 1),
    - delotok za učenje in izdelavo izvedbenega modela za klasifikacijo nestrukturiranih besedil na podlagi tematskih modelov, ki so bili identificirani in določeni v okviru delotoka za izdelavo tematskih modelov,
- delotok za izdelavo naslova popisne enote.

Za obdelavo podatkov v okviru vseh predvidenih delotokov je bilo uporabljenih 50.000 individualnih nestrukturiranih zapisov (člankov) v strojno berljivi tekstovni obliki iz vira The Guardian Open Platform (2022). Vsebina zapisov pred obdelavo podatkov vnaprej ni

bila znana, ravno tako niso bili zagotovljeni nobeni metapodatki ali drugi opisni podatki, s katerimi bi lahko olajšali izvedbo zastavljenih ciljev.

3.3.5 Prepoznava imenskih entitet

Nozza idr. (2021) izpostavljajo, da je naloga prepoznavanja imenskih entitet (NER) namenjena prepoznavanju imenskih entitet v danem besedilu in njihovem razvrščanju v vnaprej določene tipe domenskih entitet, kot so osebe, organizacije, lokacije itd. Ali kot navajajo Li idr. (2022), je imenska entiteta beseda ali besedna zveza, ki jasno identificira eno postavko ali vrednost iz niza drugih postavk ali vrednosti s podobnimi atributi.

Prepoznavanje imenskih entitet ima lahko različno vlogo oziroma vsebinsko obravnavo, saj je namen prepoznavati entitete v kakršnemkoli kontekstu ali tipizaciji, kot je predvidena s strani osebe, ki želi identificirati imenske entitete. Tako se lahko izvede prepoznavo imenskih entitet zgolj na enem področju, npr. na področju medicine (Leaman in Lu 2016; Abacha Asma idr. 2015; Liu idr. 2017; Sun idr. 2018).

Na splošno bi lahko opredelili tri pristope za izvedbo prepoznave imenskih entitet (Eltyeb in Salim 2014):

- pristop z uporabo slovarjev,
- pristop z uporabo pravil in
- pristop z uporabo strojnega učenja.

Pristop z uporabo slovarjev je najenostavnejši pristop za prepoznavo imenskih entitet. Za namen prepoznave imenskih entitet se uporablja predpripravljen slovar, ki vsebuje seznam besed ali besednih zvez imenskih entitet s pripadajočimi oznakami, za kakšno entiteto gre. To je zelo preprost pristop, kjer se besede ali besedne zveze iz besedila, kjer je namen prepoznati imenske entitete, primerja z vnosi v slovarju in v primeru ujemanja označi del besedila, tj. besedo ali besedno zvezo z oznako, ki je določena v slovarju. Ta pristop ima slabost predvsem v tem, da je pri večji količini podatkov (slovarju) in velikem obsegu zapisov, za katere je treba izvesti prepoznavo imenskih entitet, potrebnega veliko časa za obdelavo in prepoznavanje. Dodatna slabost je tudi, da je treba takšen slovar vedno dopolnjevati in posodabljati.

Pri pristopu z uporabo pravil se uporablja vnaprej določen nabor pravil za pridobivanje informacij. V osnovi lahko delimo pristop na uporabo dveh sklopov pravil (Eltyeb in Salim 2014):

- pravila na podlagi vzorcev in
- pravila, ki temeljijo na kontekstu.

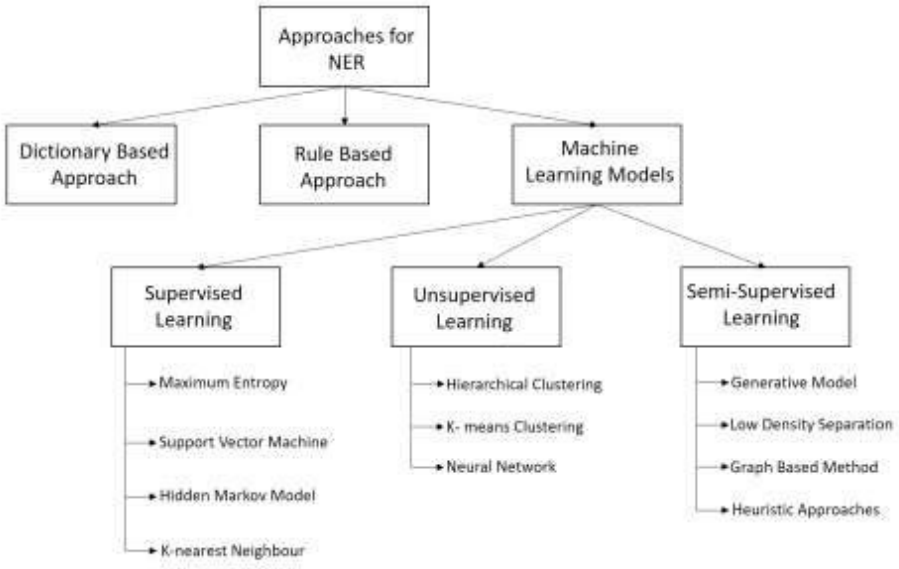
Pravila, ki temeljijo na vzorcu, uporabljajo morfološke vzorce besed, medtem ko pravila, ki temeljijo na kontekstu, uporabljajo kontekst besede, ki je podana v besedilu ali dokumentu, kjer se nahaja.

Pristopi, ki temeljijo na uporabi strojnega učenja, uporabljajo statistične modele za prepoznavanje imenskih entitet z uporabo reprezentativnih lastnosti podatkov oziroma besedila, ki je predmet učenja.

Za razvoj sistemov, ki temeljijo na strojnem učenju, sta potrebna dva osnovna koraka (Wagh idr. 2017):

- učenje: model strojnega učenja se usposablja na podlagi vnaprej označenih zapisov ali dokumentov,
- označevanje ali klasificiranje: v zapisih ali dokumentih se izvede identifikacija imenskih entitet in njihovo označevanje na podlagi prednaučenega modela (glej sliko 13).

Slika 13: Prepoznava imenskih entitet – strojno učenje



Vir: Wagh idr. 2017.

Znotraj prvega delotoka sta bila za namen preverjanja učinkovitosti prepoznave imenskih entitet vzpostavljena dva pristopa. Prvi pristop je bil usmerjen v izdelavo lastnega modela za prepoznavo imenskih entitet na osnovi lastnega testnega slovarja imenskih entitet, drugi pristop pa je vključeval postopek prepoznave imenskih entitet na osnovi predizdelanega splošnega modela, ki je bil narejen s predhodno uporabo strojnega učenja.

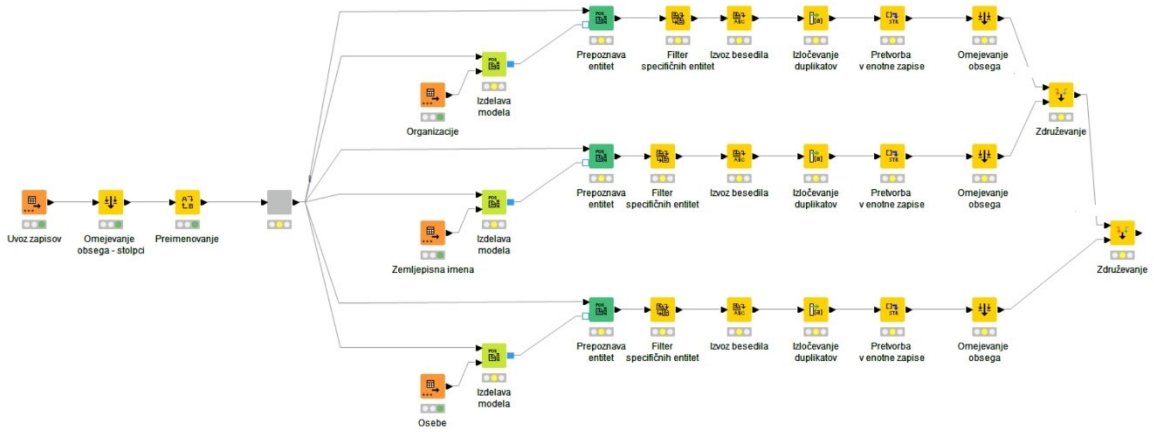
V drugi pristop, tj. prepoznavo imenskih entitet na osnovi predpripravljenega delovnega modela, so bili vključeni trije različni predizdelani modeli z namenom primerjave rezultatov učinkovitosti prepoznave imenskih entitet (SPACY 2023):

- Stanford Named Entity Recognizer (Finkel idr. 2005),
- BERT – bert-large-NER (Devlin idr. 2018; BERT-large-NER 2023) in
- SPACY EntityRecognizer – en-core-web-lg (EN-core-web-lg 2023).

Na sliki 14 je prikazan celoten proces izdelave modela za prepoznavo imenskih entitet na podlagi pristopa z uporabo slovarjev. Proces je vključeval izdelavo modelov za prepoznavo naslednjih imenskih entitet:

- lokacije,
- organizacije in
- osebe.

Slika 14: Delotok – izdelava modela za prepoznavo imenskih entitet



Vir: KNIME 2024.

V tabeli 1 so dokumentirani koraki delotoka izdelave modela za prepoznavo imenskih entitet z uporabo slovarjev, vključno z vhodno vrednostjo in obliko, izhodno vrednostjo in obliko ter parametri.

Tabela 1: Izdelava modela za prepoznavo imenskih entitet z uporabo slovarjev

Faza	Opis	Vhodni/izhodni podatki
Zajem podatkov (uvoz zapisov).	Zajem podatkov predstavlja uvoz posameznih nestrukturiranih zapisov (člankov) v digitalni obliki v orodje KNIME (2024). Format datotek, ki je bil uporabljen za uvoz podatkov člankov, je TXT (tekstovna datoteka) (TXT 2023). Datoteka, ki je bila uvožena, že vključuje agregirano obliko vseh individualnih zapisov, in sicer vsak zapis v novi vrstici.	*Vhodna oblika:* individualne datoteke v TXT-formatu (TXT 2023). *Izhodna oblika:* tabela z individualnimi zapisi, ki vključujejo besedilo posamezne tekstovne datoteke.
Omejevanje obsega – stolpci.	Izločevanje vseh nepotrebnih podatkov, ki so bili zajeti ob uvozu izvirnih zapisov (metapodatkov).	*Vhodna oblika:*

		tabela z besedili zapisov in povezanimi podatki v obliki stolpcev. *Izhodna oblika:* tabela z enim samim stolpcem, ki vključuje besedilo zapisov, kje se nahaja izvirna oblika besedila.
Preimenovanje stolpca tabele.	Preimenovanje stolpca tabele, ki tako dobi jasno oznako, da shranjuje besedilo posameznega uvoženega zapisa. Namenjeno je kasnejšemu lažjemu sledenju polne vsebine besedila, ki je bil prvotno uvoženo v delotok.	*Vhodna oblika:* tabela z besedili zapisov z generično imenovanim stolpcem. *Izhodna oblika:* tabela z besedili zapisov z jasno poimenovanim stolpcem, kje se nahaja izvirna oblika besedila.
Izločevanje praznih sekcij (podproces).	V aktivnosti se izvede izločevanje praznih sekcij v besedilu izvirnega zapisa. Iz besedila se izvede izločitev dveh ali več presledkov.	*Vhodna oblika:* tabela z besedilom zapisa, ki vključuje prazne sekcije v delu besedila. *Izhodna oblika:* tabela z besedilom zapisa, ki ima izločene prazne sekcije iz besedil izvirnih zapisov.
Izločevanje praznih sektorjev (podproces).	V aktivnosti se izvede urejanje besedil, pri čemer se izloči prazne sektorje v besedilu zapisa, t. i. »line break«. Namen je izločiti presledke, ki predstavljajo prelome vrstic v besedilu.	*Vhodna oblika:* tabela z besedilom zapisa, ki vključuje presledke, ki predstavljajo prelome vrstic v besedilu. *Izhodna oblika:* tabela z besedilom zapisa, ki ima izločene prazne sektorje v besedilu, ki predstavljajo prelome vrstic.
Pretvorba besedila v dokumente (podproces).	V aktivnosti se izvede pretvorba besedila v programski objekt, imenovan »dokument«, ki	*Vhodna oblika:* tabela z izvirnim besedilom zapisov.

	omogoča lažjo nadaljnjo obdelavo podatkov, posebej v delu izvajanja različnih operacij in obdelav besedila ter opcijskega dodajanja opisnih vsebin.	*Izhodna oblika:* tabela z izvirnim besedilom zapisov, vključno z dodanim programskim objektnim tipom »dokument«.
Pretvorba besedil izvirnih zapisov iz velikih v male črke (podproces).	Aktivnost, v kateri se izvede pretvorba besedil izvirnih zapisov (dokument) iz velikih v male črke.	*Vhodna oblika:* tabela z izvirnimi besedili zapisov, vključno s tabelo s pretvorjenim objektnim poljem tipa »dokument«. *Izhodna oblika:* tabela z izvirnimi besedili zapisov, vključno s tabelo s pretvorjenim objektnim poljem tipa »dokument« in besedili zapisov, kjer ima celotno besedilo male črke.
Zajem podatkov iz slovarja (uvoz zapisov).	Zajem podatkov predstavlja uvoz posameznih imenskih entitet v digitalni obliki v orodje KNIME (2024). Format datotek, izbran za uvoz podatkov člankov, je TXT (tekstovna datoteka) (TXT 2023). Datoteka, ki je bila uvožena, vključuje slovar imenskih entitet, ki so že omejene na en tip entitete. Slovarji so bili povzeti iz osnove StanfordNLP, ki je prosto dostopna na spletni strani Stanford Named Entity Recognizer (Finkel idr. 2005).	*Vhodna oblika:* individualne datoteke v TXT-formatu (TXT 2023). *Izhodna oblika:* tabela z individualnimi zapisi, ki vključujejo besedilo imenskih entitet, ki so že omejene na en tip entitete.
Izdelava modela za prepoznavanje imenskih entitet.	V aktivnosti se izvede treniranje/učenje na podlagi predhodno razpoložljivih vsebin zapisov, ki so namenjeni prepoznavi imenskih entitet, in predhodno določenih imenskih entitet iz slovarja.	*Vhodna oblika:* tabela s podatki, ki so namenjeni prepoznavanju vsebine, in slovar imenskih entitet. *Izhodna oblika:*

		model za prepoznavanje imenskih entitet in specifikacija za samostojno identifikacijo imenskih entitet.
Uporaba modela za določanje imenskih entitet.	V aktivnosti se izvede prepoznavanje imenskih entitet v individualnih zapisih na podlagi izdelanega odločitvenega modela za namen ugotavljanja natančnosti in uspešnosti prepoznanih imenskih entitet.	*Vhodna oblika:* model za prepoznavanje imenskih entitet s specifikacijo za samostojno prepoznavanje glede na tip entitete in tabela z izvirnimi zapisi. *Izhodna oblika:* tabela z izvirnimi zapisi in ter prepoznanimi imenskimi entitetami v obliki gruče podatkov v dokumentu.
Filtriranje prepoznanih imenskih entitet.	V tej fazi se izvede prepoznavanje in filtriranje imenskih entitet, ki pripadajo določenemu tipu entitete. Filter je določen glede na tip entitete, ki je bil prednastavljen v slovarju za učenje modela (npr. »organizacije«).	*Vhodna oblika:* tabela z izvirnimi zapisi in prepoznanimi imenskimi entitetami v obliki dokumenta. *Izhodna oblika:* tabela z izvirnimi zapisi ter prepoznanimi in filtriranimi imenskimi entitetami v obliki dokumenta.
Ekstrakcija besedila iz programskega objekta dokument.	V aktivnosti se izvede povzemanje (ekstrakcija) besedila, ki vključuje filtrirane imenske entitete v navadno besedilo (zapis) zaradi lažjega združevanja in manipulacije s podatki individualnega zapisa.	*Vhodna oblika:* tabela z izvirnimi zapisi ter prepoznanimi in filtriranimi imenskimi entitetami v obliki dokumenta. *Izhodna oblika:* tabela z izvirnimi zapisi ter prepoznanimi in filtriranimi imenskimi entitetami v obliki dokumenta in v obliki navadnega besedila (zapisa).

Izdelava zbirk prepoznanih imenskih entitet in izločevanje duplikatov.	V tej fazi se izvede izdelava zbirk prepoznanih imenskih entitet v obliki programskega objekta glede na individualen izvirni zapis. Poleg izdelave zbirk prepoznanih imenskih entitet se izvede tudi deduplikacija imenskih entitet oziroma urejanje zapisov z namenom izločevanja identičnih imenskih entitet glede na posamezno zbirko. Namen je izločiti morebitno podvojeno/identično prepoznano imensko entiteto v okviru zbirke individualnega izvirnega zapisa.	*Vhodna oblika:* tabela z izvirnimi zapisi in ter prepoznanimi in filtriranimi imenskimi entitetami v obliki dokumenta in v obliki navadnega besedila (zapisa). *Izhodna oblika:* tabela z izvirnimi zapisi ter ustvarjenimi kolekcijami imenskih entitet glede na individualen izvirni zapis, vključno z izločenimi duplikati imenskih entitet.
Pretvorba kolekcij imenskih entitet v navadno besedilo.	V fazi se izvede pretvorba programskega objekta »zbirka« v navadno besedilo (zapis) zaradi lažjega združevanja in manipulacije s podatki individualnega zapisa.	*Vhodna oblika:* tabela z izvirnimi zapisi ter ustvarjenimi kolekcijami imenskih entitet glede na individualen izvirni zapis, vključno z izločenimi duplikati imenskih entitet. *Izhodna oblika:* tabela z izvirnimi zapisi ter prepoznanimi, filtriranimi in izločenimi duplikati imenskih entitet v obliki navadnega besedila (zapisa).
Omejevanje obsega – stolpci.	Izločevanje vseh nepotrebnih podatkov, ki so bili ustvarjeni v okviru delotoka.	*Vhodna oblika:* tabela z izvirnimi zapisi ter prepoznanimi, filtriranimi in izločenimi duplikati imenskih entitet v obliki navadnega besedila (zapisa) ter ostalimi programskimi objekti. *Izhodna oblika:* tabela z zgolj dvema stolpcema, in sicer s

		stolpcem, ki vključuje besedilo zapisov, kje se nahaja izvirna oblika besedila, ter stolpcem, kjer se nahajajo prepoznane imenske entitete glede na individualen izvirni zapis.
Združevanje podatkov.	V zadnji aktivnosti delotoka se izvede združevanje podatkov vseh treh vej prepoznavanja različnih tipov imenskih entitet z namenom zagotavljanja enovitega pregleda identificiranih imenskih entitet glede na individualen izvirni zapis.	*Vhodna oblika:* tabela z izvirnimi zapisi in različnimi prepoznanimi imenskimi entitetami glede na tip entitete. *Izhodna oblika:* tabela z združenimi izvirnimi zapisi ter prepoznanimi imenskimi entitetami glede na tip entitete.

Vir: KNIME 2024.

Na sliki 15 je prikazan podproces za izdelavo modela za prepoznavo imenskih entitet, v katerem se izvede urejanje podatkov predvsem z namenom izločevanja nepotrebne vsebine ter omejevanja obsega obdelave podatkov.

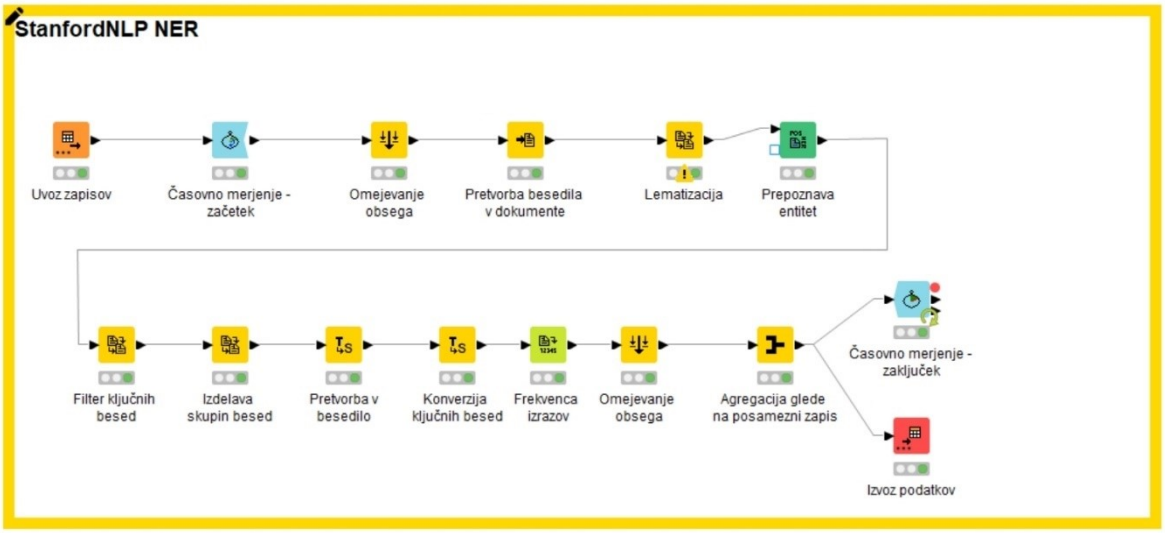
Slika 15: Delotok – podproces za izdelavo modela za prepoznavo imenskih entitet

Vir: KNIME 2024.

V raziskavi je bil namen preveriti oziroma primerjati učinkovitost vsaj dveh različnih pristopov pri prepoznavi imenskih entitet. Poleg izdelave modela z uporabo slovarjev je bil uporabljen tudi pristop prepoznavanja imenskih entitet na podlagi modelov, ki so bili narejeni z uporabo strojnega učenja.

Na sliki 16 je prikazan prvi delotok prepoznavanja imenskih entitet z uporabo predizdelanega modela, ki je bil izveden z uporabo rešitve Stanford Named Entity Recognizer (Finkel idr. 2005) in integriranega predizdelanega modela.

Slika 16: Delotok – uporaba modela za prepoznavo imenskih entitet Stanford Named Entity Recognizer (StanfordNLP)



Vir: KNIME 2024.

Tabela 2: Faze prepoznave imenskih entitet – Stanford Named Entity Recognizer

Faza	Opis	Vhodni/izhodni podatki
Zajem podatkov (uvoz zapisov).	Zajem podatkov predstavlja uvoz posameznih nestrukturiranih zapisov (člankov) v digitalni obliki v orodje KNIME (2024). Format datotek, ki je bil izbran za uvoz nestrukturiranih besedil člankov, je TXT (tekstovna datoteka) (TXT 2023). Datoteka, ki je bila uvožena, že vključuje agregirano obliko vseh individualni zapisov, in sicer vsak zapis v novi vrstici.	*Vhodna oblika:* individualne datoteke v TXT-formatu (TXT 2023). *Izhodna oblika:* tabela z individualnimi zapisi, ki vključujejo besedilo posamezne tekstovne datoteke.
Časovno merjenje – začetek.	Aktivnost je namenjena merjenju časa, ki je potreben za izvedbo delotoka prepoznavanja imenskih entitet. Aktivnost ne vpliva na vsebinsko ali strukturno obdelavo podatkov in ne spreminja obstoječih podatkov.	
Omejevanje obsega – stolpci.	Izločevanje vseh nepotrebnih podatkov, ki so bili zajeti ob uvozu izvirnih zapisov (metapodatkov).	*Vhodna oblika:*

		tabela z besedili zapisov in povezanimi podatki v obliki stolpcev. *Izhodna oblika:* tabela z zgolj enim stolpcem, ki vključuje besedilo zapisov, kje se nahaja izvirna oblika besedila.
Pretvorba besedila v dokumente.	V aktivnosti se zaradi lažjega pristopa k programski obdelavi besedil zapise pretvori v programski objekt, imenovan »dokument«, ki v omogoča lažjo obdelavo podatkov, zlasti v delu možnega dodajanja opisnih vsebin.	*Vhodna oblika:* tabela z izvirnim besedilom zapisov. *Izhodna oblika:* tabela z izvirnim besedilom zapisov in dodatnim objektnim poljem tipa »dokument«.
Izdelava lem.	V aktivnosti se programsko izvede izdelava lem (krnjenje besed) z namenom določanja osnovnih oblik posameznih besed za lažjo identifikacijo imenskih entitet.	*Vhodna oblika:* tabela z izvirnim besedilom zapisov, vključno z objektnim poljem »dokument«. *Izhodna oblika:* tabela z izvirnim besedilom zapisov in dodatnim objektnim poljem tipa »dokument« in besedilom zapisa, ki vključuje leme, tj. besede v osnovni krnjeni obliki.
Uporaba modela za določanje imenskih entitet.	V aktivnosti se izvede prepoznavanje imenskih entitet v individualnih zapisih na podlagi predizdelanega modela. Uporabljeni predizdelani model: »English 7 classes ni distsim« (Finkel idr. 2005).	*Vhodna oblika:* predizdelani model za prepoznavanje imenskih entitet s specifikacijo za samostojno prepoznavanje glede na tip entitete in tabela z izvirnim besedilom zapisov s pretvorjenim objektnim poljem

		»dokument« in besedilom zapisa, ki vključuje besede v osnovni krnjeni obliki. *Izhodna oblika:* tabela z izvirnimi zapisi in prepoznanimi imenskimi entitetami v obliki gruče podatkov v dokumentu.
Filtriranje prepoznanih imenskih entitet.	V tej fazi se izvede prepoznavanje in filtriranje imenskih entitet, ki pripadajo določenemu tipu entitete. Filter je določen na način, da filtrira in obdrži naslednje tipe entitet in sicer: »Date, Location, Organization, Person, Time« (KNIME 2024).	*Vhodna oblika:* tabela z izvirnimi zapisi in prepoznanimi imenskimi entitetami v obliki gruče podatkov v dokumentu. *Izhodna oblika:* tabela z izvirnimi zapisi ter prepoznanimi in filtriranimi imenskimi entitetami v obliki dokumenta.
Izdelava skupin (vreč) besed.	V aktivnosti se izvede izdelava skupin (vreč) ključnih besed na podlagi prepoznanih imenskih entitet.	*Vhodna oblika:* tabela z izvirnimi zapisi ter prepoznanimi in filtriranimi imenskimi entitetami v obliki dokumenta. *Izhodna oblika:* tabela s pretvorjenim objektnim poljem tipa »dokument« in besedilom zapisa s pripadajočimi skupinami (vrečami) ključnih besed.
Pretvorba tipov imenskih entitet v navadno besedilo.	V aktivnosti se izvede pretvorba programskega objekta »tag« (oznaka) individualnega tipa imenske entitete v navadno besedilo (zapis) zaradi lažjega združevanja in manipulacije s podatki individualnega zapisa.	*Vhodna oblika:* tabela s pretvorjenim objektnim poljem tipa »dokument« in besedilom zapisa s pripadajočimi skupinami (vrečami) ključnih besed. *Izhodna oblika:*

		tabela z izvirnimi zapisi in pretvorjenimi oznakami individualnih tipov imenskih entitet v obliki navadnega besedila (zapisa).
Pretvorba prepoznanih imenskih entitet v navadno besedilo.	V tej fazi se izvede pretvorba iz programskega objekta »term« (entiteta) posamezno identificirane imenske entitete v navadno besedilo (zapis) zaradi lažjega združevanja in manipulacije s podatki individualnega zapisa.	*Vhodna oblika:* tabela z izvirnimi zapisi in pretvorjenimi oznakami individualnih tipov imenskih entitet v obliki navadnega besedila (zapisa). *Izhodna oblika:* tabela z izvirnimi zapisi in pretvorjenimi oznakami individualnih tipov imenskih entitet ter prepoznanih imenskih entitet v obliki navadnega besedila (zapisa).
Omejevanje obsega – stolpci.	Izločevanje vseh nepotrebnih podatkov, ki so bili ustvarjeni v okviru delotoka.	*Vhodna oblika:* tabela z izvirnimi zapisi ter prepoznanimi, filtriranimi in imenskimi entitetami v obliki navadnega besedila (zapisa) ter ostalimi programskimi objekti. *Izhodna oblika:* tabela z zgolj dvema stolpcema, in sicer s stolpcem, ki vključuje izvirno besedilo zapisov (večkratni zapisi glede na število identificiranih imenskih entitet, ki pripadajo posameznemu zapisu), in stolpcem, kjer se nahajajo prepoznane posamezne imenske entitete

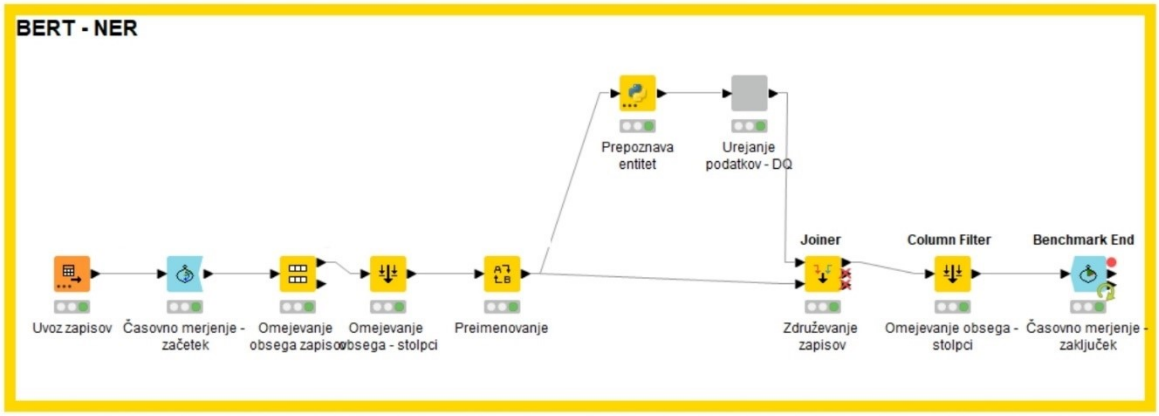
		glede na individualen izvirni zapis.
Združevanje podatkov.	V zadnji aktivnosti delotoka se izvede proces združevanja podatkov identificiranih imenskih entitet glede na individualen izvirni zapis.	*Vhodna oblika:* tabela z zgolj dvema stolpcema, in sicer s stolpcem, ki vključuje izvirno besedilo zapisov (večkratni zapisi glede na število identificiranih imenskih entitet, ki pripadajo posameznemu zapisu), in stolpcem, kjer se nahajajo prepoznane posamezne imenske entitete glede na individualen izvirni zapis. *Izhodna oblika:* tabela z izvirnimi zapisi in združenimi prepoznanimi imenskimi entitetami glede na individualen izvirni zapis.
Časovno merjenje – zaključek.	Aktivnost je namenjena merjenju časa, ki je potreben za izvedbo delotoka prepoznavanja imenskih entitet. Aktivnost ne vpliva na vsebinsko ali strukturno obdelavo podatkov in ne spreminja obstoječih podatkov. Aktivnost ustavi in v sekundah izmeri čas, ki je bil potreben za izvedbo vseh faz in aktivnosti delotoka.	
Izvoz podatkov.	V tej fazi se izvede izvoz vseh izvirnih zapisov skupaj s prepoznanimi imenskimi entitetami v eno tekstovno datoteko.	*Vhodna oblika:* tabela z izvirnimi zapisi in združenimi prepoznanimi imenskimi entitetami glede na individualen izvirni zapis. *Izhodna oblika:* tekstovna datoteka z izvirnimi zapisi in

		združenimi prepoznanimi imenskimi entitetami glede na individualen izvirni zapis.

Vir: KNIME 2024.

Na sliki 17 je prikazan drugi delotok prepoznavanja imenskih entitet z uporabo predizdelanega modela, ki je bil izveden z uporabo arhitekture in integriranega predizdelanega modela BERT (Devlin idr. 2018).

Slika 17: Delotok – uporaba modela za prepoznavo imenskih entitet BERT



Vir: KNIME 2024.

Tabela 3: Faze prepoznave imenskih entitet – BERT

Faza	Opis	Vhodni/izhodni podatki
Zajem podatkov (uvoz zapisov).	Zajem podatkov predstavlja uvoz posameznih nestrukturiranih zapisov (člankov) v digitalni obliki v orodje KNIME (2024). Format datotek, ki je bil izbran za uvoz nestrukturirane vsebine člankov, je TXT (tekstovna datoteka) (TXT 2023). Datoteka, ki je bila uvožena, že vključuje agregirano obliko vseh individualnih zapisov, in sicer vsak zapis v novi vrstici.	*Vhodna oblika:* individualna datoteka v TXT-formatu (TXT 2023). *Izhodna oblika:* tabela z individualnimi zapisi, ki vključujejo besedilo posamezne tekstovne datoteke.
Časovno merjenje.	Aktivnost je namenjena merjenju časa, ki je potreben za izvedbo delotoka prepoznavanja imenskih entitet. Aktivnost ne vpliva na vsebinsko ali strukturno obdelavo podatkov in ne spreminja obstoječih podatkov.	

Omejevanje obsega zapisov.	Uporaba treh različnih predizdelanih modelov za prepoznavo imenskih entitet je bila namenjen temu, da se preveri učinkovitost posameznega modela na podlagi istih zapisov, ki so bili podlaga za prepoznavanje. V tej fazi se je izvedla omejitev zapisov, ki so bili namenjeni primerjanju rezultatov iz prvega delotoka za prepoznavo imenskih entitet z uporabo predizdelanih modelov. Omejitev je bila postavljena na 100 individualnih in različnih zapisov.	*Vhodna oblika:* tabela z individualnimi zapisi, ki vključujejo besedilo posamezne tekstovne datoteke. *Izhodna oblika:* tabela z individualnimi zapisi, omejenimi na 100 različnih individualnih zapisov.
Omejevanje obsega – stolpci.	Izločevanje vseh nepotrebnih podatkov, ki so bili zajeti ob uvozu izvirnih zapisov (metapodatkov).	*Vhodna oblika:* tabela z besedili zapisov in povezanimi podatki v obliki stolpcev. *Izhodna oblika:* tabela z zgolj enim stolpcem, ki vključuje besedilo zapisov, kje se nahaja izvirna oblika besedila.
Preimenovanje stolpca.	Preimenovanje stolpca za lažje urejanje in izvajanje skripte za izvedbo prepoznave imenskih entitet.	*Vhodna oblika:* tabela z zgolj enim stolpcem, ki vključuje besedilo zapisov, kje se nahaja izvirna oblika besedila. *Izhodna oblika:* tabela z zgolj enim preimenovanim stolpcem, ki vključuje besedilo zapisov, kje se nahaja izvirna oblika besedila.
Uporaba modela za določanje imenskih entitet.	V fazi se izvede prepoznavanje imenskih entitet v individualnih zapisih na podlagi predizdelanega modela. Uporabljeni predizdelani	*Vhodna oblika:* predizdelani model za prepoznavanje imenskih entitet s specifikacijo za samostojno prepoznavanje

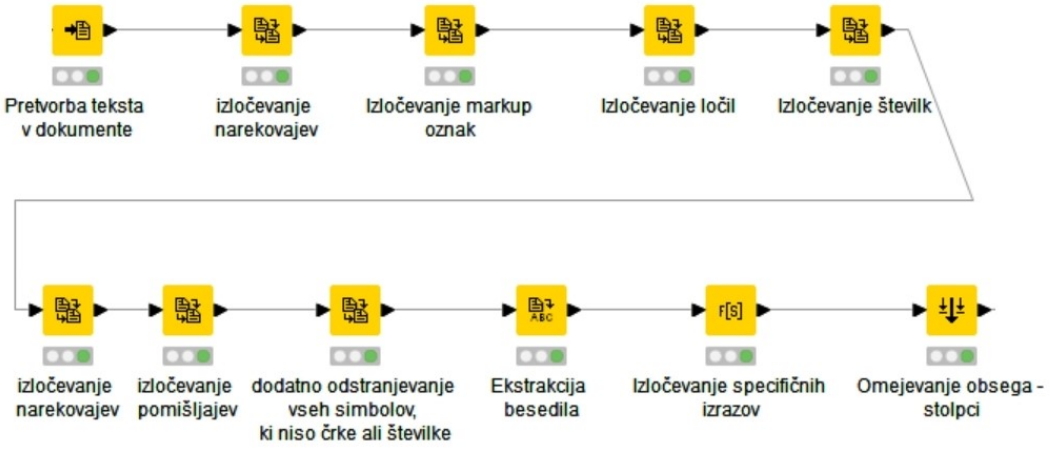
	model BERT: »bert-large-NER« (BERT-large-NER 2023).	ne glede na tip entitete in tabela z izvirnim besedilom zapisov. *Izhodna oblika:* tabela s prepoznanimi imenskimi entitetami.
Urejanje podatkov – pretvorba teksta v dokumente (podproces).	V aktivnosti so bila besedila izvornih zapisov pretvorjena v programski objekt, imenovan »dokument«, ki omogoča lažjo obdelavo podatkov, posebej za namen zagotavljanja kakovosti podatkov.	*Vhodna oblika:* tabela s prepoznanimi imenskimi entitetami. *Izhodna oblika:* tabela s prepoznanimi imenskimi entitetami, vključno z objektnim poljem »dokument«.
Urejanje podatkov – čiščenje besedila (podproces).	Faza vključuje več aktivnosti, namenjenih t. i. čiščenju besedila, in sicer z namenom izločevanja nepotrebnih podatkov zaradi lažje ekstrapolacije imenskih entitet.	*Vhodna oblika:* tabela s prepoznanimi imenskimi entitetami, vključno z objektnim poljem »dokument«. *Izhodna oblika:* tabela s prepoznanimi imenskimi entitetami, ki ne vključuje nepotrebnega besedila in nepotrebnih znakov.
Ekstrakcija besedila iz programskega objekta dokument.	V tej fazi se izvede ekstrakcija besedila, ki vključuje prepoznane imenske entitete v navadno besedilo (zapis) zaradi lažjega združevanja in manipulacije s podatki individualnega zapisa.	*Vhodna oblika:* tabela s prepoznanimi imenskimi entitetami, ki ne vključuje nepotrebnega besedila in nepotrebnih znakov. *Izhodna oblika:* tabela s prepoznanimi imenskimi entitetami, ki ne vključuje nepotrebnega besedila in nepotrebnih znakov v obliki dokumenta

		in v obliki navadnega besedila (zapisa).
Združevanje podatkov.	V aktivnosti delotoka se izvede proces združevanja podatkov identificiranih imenskih entitet glede na individualen izvirni zapis.	*Vhodna oblika:* tabela, ki vključuje izvirno besedilo zapisov, ter tabela, kjer se nahajajo prepoznane imenske entitete. *Izhodna oblika:* tabela, kjer so združeni izvirni zapisi in prepoznane imenske entitete, ki jim pripadajo.
Omejevanje obsega – stolpci.	Izločevanje vseh nepotrebnih podatkov, ki so bili ustvarjeni v okviru delotoka.	*Vhodna oblika:* tabela, kjer so združeni izvirni zapisi in prepoznane imenske entitete, ki jim pripadajo, in ostali podatki, ki so bili ustvarjeni v okviru delotoka. *Izhodna oblika:* tabela z zgolj dvema stolpcema, in sicer s stolpcem, ki vključuje izvirno besedilo zapisov, in stolpcem, kjer se nahajajo prepoznane posamezne imenske entitete glede na individualen izvirni zapis.

Vir: KNIME 2024.

Na sliki 18 je prikazan podproces za urejanje podatkov pri izdelavi modela za prepoznavo imenskih entitet, v okviru katerega se izvede urejanje podatkov predvsem z namenom izločevanje nepotrebne vsebine in omejevanja obsega obdelave podatkov.

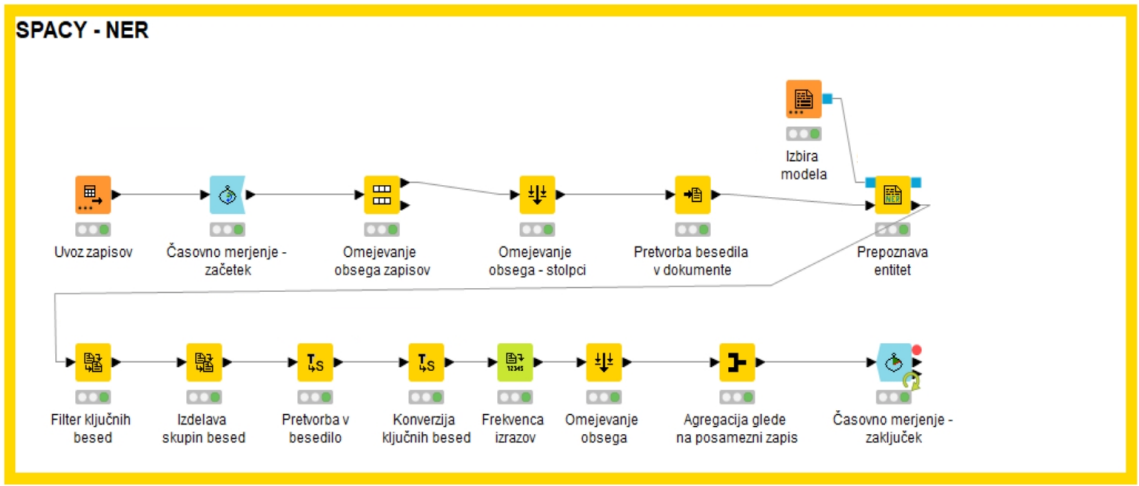
Slika 18: Delotok – podproces za urejanje podatkov pri prepoznavanju imenskih entitet – BERT



Vir: KNIME 2024.

Na sliki 19 je prikazan tretji delotok prepoznavanja imenskih entitet z uporabo predizdelanega modela, ki je bil izveden z uporabo arhitekture in integriranega predizdelanega modela SPACY (2023).

Slika 19: Delotok – uporaba modela za prepoznavo imenskih entitet SPACY



Vir: KNIME 2024.

Tabela 4: Faze prepoznave imenskih entitet – SPACY

Faza	Opis	Vhodni/izhodni podatki
Zajem podatkov (uvoz zapisov).	Zajem podatkov predstavlja uvoz posameznih nestrukturiranih zapisov (člankov) v digitalni obliki v orodje KNIME (2024). Format datotek, ki je bil izbran za uvoz podatkov člankov, je TXT (tekstovna datoteka) (TXT 2023). Datoteka, ki je bila uvožena, že vključuje agregirano obliko vseh individualni zapisov, in sicer vsak zapis v novi vrstici.	*Vhodna oblika:* individualne datoteke v TXT-formatu (TXT 2023). *Izhodna oblika:* tabela z individualnimi zapisi, ki vključujejo besedilo posamezne tekstovne datoteke.
Časovno merjenje – začetek.	Aktivnost je namenjena merjenju časa, ki je potreben za izvedbo delotoka prepoznavanja imenskih entitet. Aktivnost ne vpliva na vsebinsko ali strukturno obdelavo podatkov in ne spreminja obstoječih podatkov.	
Omejevanje obsega – stolpci.	Izločevanje vseh nepotrebnih podatkov, ki so bili zajeti ob uvozu izvirnih zapisov (metapodatkov).	*Vhodna oblika:* tabela z besedili zapisov in povezanimi podatki v obliki stolpcev. *Izhodna oblika:* tabela z zgolj enim stolpcem, ki vključuje besedilo zapisov, kjer se nahaja izvirna oblika besedila.
Pretvorba besedila v dokumente.	Zaradi nadaljnjega programskega obdelovanja besedila izvirnih zapisov so bili ti pretvorjeni v programski objekt, imenovan »dokument«, ki omogoča lažjo obdelavo podatkov, posebej v delu možnega dodajanja opisnih vsebin.	*Vhodna oblika:* tabela z izvirnim besedilom zapisov. *Izhodna oblika:* tabela z izvirnim besedilom zapisov, vključno z objektnim poljem tipa »dokument«.

Uporaba modela za določanje imenskih entitet.	V tej fazi se izvede prepoznavanje imenskih entitet v individualnih zapisih na podlagi predizdelanega modela. Uporabljeni predizdelani model: - za angleški jezik: en-core-web-lg (En-core-web-lg 2023). - za slovenski jezik: sl_core_news_lg (sl_core_news_lg 2023).	*Vhodna oblika:* predizdelani model za prepoznavanje imenskih entitet s specifikacijo za samostojno prepoznavanje glede na tip entitete in tabela z izvirnim besedilom zapisov s pretvorjenim objektnim poljem »dokument«. *Izhodna oblika:* tabela z izvirnimi zapisi in prepoznanimi imenskimi entitetami v obliki gruče podatkov v dokumentu.
Filtriranje prepoznanih imenskih entitet.	V tej fazi se izvede prepoznavanje in filtriranje imenskih entitet, ki pripadajo določenemu tipu entitete. Filter je določen na način, da filtrira in obdrži naslednje tipe entitet: Date, Location, Organization, Person, Date (KNIME 2024).	*Vhodna oblika:* tabela z izvirnimi zapisi in prepoznanimi imenskimi entitetami v obliki gruče podatkov v dokumentu. *Izhodna oblika:* tabela z izvirnimi zapisi ter prepoznanimi in filtriranimi imenskimi entitetami v obliki dokumenta.
Izdelava skupin besed.	Izdelava skupin ključnih besed na podlagi prepoznanih imenskih entitet.	*Vhodna oblika:* tabela z izvirnimi zapisi ter prepoznanimi in filtriranimi imenskimi entitetami v obliki dokumenta. *Izhodna oblika:* tabela s pretvorjenim objektnim poljem tipa »dokument« in besedilom zapisa s pripadajočimi skupinami (vrečami) ključnih besed.
Pretvorba tipov imenskih entitet v navadno besedilo.	V aktivnosti se izvede pretvorba programskega objekta »tag« (oznaka) individualnega tipa	*Vhodna oblika:* tabela s pretvorjenim objektnim poljem tipa

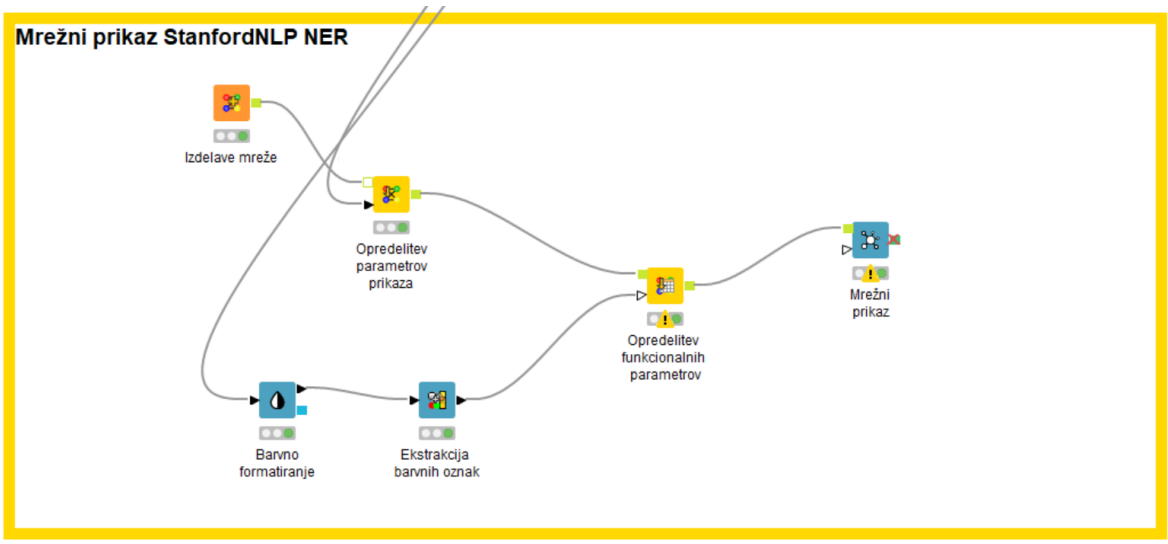
	imenske entitete v navadno besedilo (zapis) zaradi lažjega združevanja in manipulacije s podatki individualnega zapisa.	»dokument« in besedilom zapisa s pripadajočimi skupinami (vrečami) ključnih besed. *Izhodna oblika:* tabela z izvirnimi zapisi in pretvorjenimi oznakami individualnih tipov imenskih entitet v obliki navadnega besedila (zapisa).
Pretvorba prepoznanih imenskih entitet v navadno besedilo.	V tej fazi se izvede pretvorba iz programskega objekta »term« (entiteta) posamezno identificirane imenske entitete v navadno besedilo (zapis) zaradi lažjega združevanja in manipulacije s podatki individualnega zapisa.	*Vhodna oblika:* tabela z izvirnimi zapisi in pretvorjenimi oznakami individualnih tipov imenskih entitet v obliki navadnega besedila (zapisa). *Izhodna oblika:* tabela z izvirnimi zapisi in pretvorjenimi oznakami individualnih tipov imenskih entitet ter prepoznanih imenskih entitet v obliki navadnega besedila (zapisa).
Omejevanje obsega – stolpci.	Izločevanje vseh nepotrebnih podatkov, ki so bili ustvarjeni v okviru delotoka.	*Vhodna oblika:* tabela z izvirnimi zapisi in prepoznanimi, filtriranimi in imenskimi entitetami v obliki navadnega besedila (zapisa) ter ostalimi programskimi objekti. *Izhodna oblika:* tabela z zgolj dvema stolpcema, in sicer s stolpcem, ki vključuje izvirno besedilo zapisov (večkratni zapisi glede na število identificiranih

		imenskih entitet, ki pripadajo posameznemu zapisu), in stolpcem, kjer se nahajajo prepoznane posamezne imenske entitete glede na individualen izvirni zapis.
Združevanje podatkov.	V zadnji aktivnosti delotoka se izvede proces združevanja podatkov identificiranih imenskih entitet glede na individualen izvirni zapis.	*Vhodna oblika:* tabela z zgolj dvema stolpcema, in sicer s stolpcem, ki vključuje izvirno besedilo zapisov (večkratni zapisi glede na število identificiranih imenskih entitet, ki pripadajo posameznemu zapisu), in stolpcem, kjer se nahajajo prepoznane posamezne imenske entitete glede na individualen izvirni zapis. *Izhodna oblika:* tabela z izvirnimi zapisi in združenimi prepoznanimi imenskimi entitetami glede na individualen izbirni zapis.
Časovno merjenje – zaključek.	Aktivnost je namenjena merjenju časa, ki je potreben za izvedbo delotoka prepoznavanja imenskih entitet. Aktivnost ne vpliva na vsebinsko ali strukturno obdelavo podatkov in ne spreminja obstoječih podatkov. Aktivnost se ustavi in v sekundah izmeri čas, ki je bil potreben za izvedbo vseh faz in aktivnosti delotoka.	

Vir: KNIME 2024.

Na sliki 20 je prikazan delotok izdelave mrežnega prikaza za predstavitev koncentracije prepoznanih individualnih imenskih entitet glede na izvirni zapis in določen tematski model. Kot vhodna vrednost izvedbenih aktivnosti je bil uporabljen rezultat delotoka prepoznave imenskih entitet, vključno z rezultatom delotoka izdelave tematskega modela.

Slika 20: Delotok – izdelava mrežnega prikaza



Vir: KNIME 2024.

Tabela 5: Faze izdelave mrežnega prikaza

Faza	Opis	Vhodni/izhodni podatki
Izdelava mreže.	Pri izdelavi mreže se izvede inicializacija mreže z osnovnimi parametri za izgradnjo končnega prikaza.	*Vhodna oblika:* / *Izhodna oblika:* osnovna mreža z osnovnimi parametri.
Opredelitev parametrov prikaza.	V aktivnosti se izvede nastavitev parametrov prikaza za izdelavo mrežnega postavitve.	*Vhodna oblika:* podatki o prepoznanih individualnih imenskih entitetah kot individualen zapis, povezan z izvirnim zapisom in določenim tematskim modelom, vključno z inicializacijo mreže. *Izhodna oblika:* osnovna mreža s posodobljenimi parametri.
Opredelitev funkcionalnih parametrov.	V aktivnosti se izvede nastavitev funkcionalnih parametrov prikaza za izdelavo mrežne postavitve.	*Vhodna oblika:* podatki o prepoznanih individualnih imenskih entitetah kot individualen zapis, povezan z izvirnim zapisom in določenim

		tematskim modelom, vključno z osnovno mrežo s posodobljenimi parametri. *Izhodna oblika:* končna mreža, pripravljena za vizualizacijo.
Prikaz rezultatov.	V aktivnosti se izvede vizualna predstavitev rezultatov delotoka.	*Vhodna oblika:* končna mreža, pripravljena za vizualizacijo. *Izhodna oblika:* grafični prikaz rezultatov – koncentracije prepoznanih imenskih entitet glede na določen tematski model.

Vir: KNIME 2024.

3.3.6 Analiza vsebine in izdelava tematskih modelov

V okviru drugega delotoka je bila za potrebe identifikacije vsebine nestrukturiranega besedila in izdelavo tematskih modelov uporabljena metoda Latent Dirichlet Allocation – LDA (Blei idr. 2003, 993−1022).

Onan (2018) navaja, da je prepoznavanje vsebine pomembna raziskovalna smer, ki vključuje več področij, kot so iskanje informacij, ekstrakcija informacij in kategorizacija besedil. V nadaljevanju Onan (2018) izpostavlja, da LDA (Blei idr. 2003, 993−1022) lahko premaga problem velike dimenzionalnosti modela vektorskega prostora, vendar je prepoznavanje ustreznih parametrov ključnega pomena za delovanje LDA (Blei idr. 2003, 993−1022). V okviru raziskave je bilo posledično primerjalno uporabljenih več različnih parametrov za doseganje čim boljših rezultatov.

Garcia-Pablos idr. (2018) ravno tako izpostavljajo pomembnost izdelave kakovostnih tematskih modelov, predvsem za nadaljnje doseganje dobrih rezultatov pri nadzorovanem ali nenadzorovanem pristopu klasifikacije vsebine.

Izdelava tematskih modelov predstavlja pristop, znotraj katerega lahko dosežemo prepoznavo vsebine skozi določanje tem, iz katerih je možno raztolmačiti osnovni pomen vsebine. Arora idr. (2020, 1) menijo, da je določanje tematskih modelov uporabna oblika za redukcijo dimenzij in analizo vsebine pri obdelavi velikih količin podatkov. Da bi lahko dosegli čim bolj samostojno obdelavo podatkov v kontekstu prepoznave vsebine, je bilo treba pristopiti k razvoju specializiranih algoritmov, katerih cilj je bil izdelati tematske modele z zasledovanjem izhodišč, kjer naj bi se vsebina določala glede na največjo

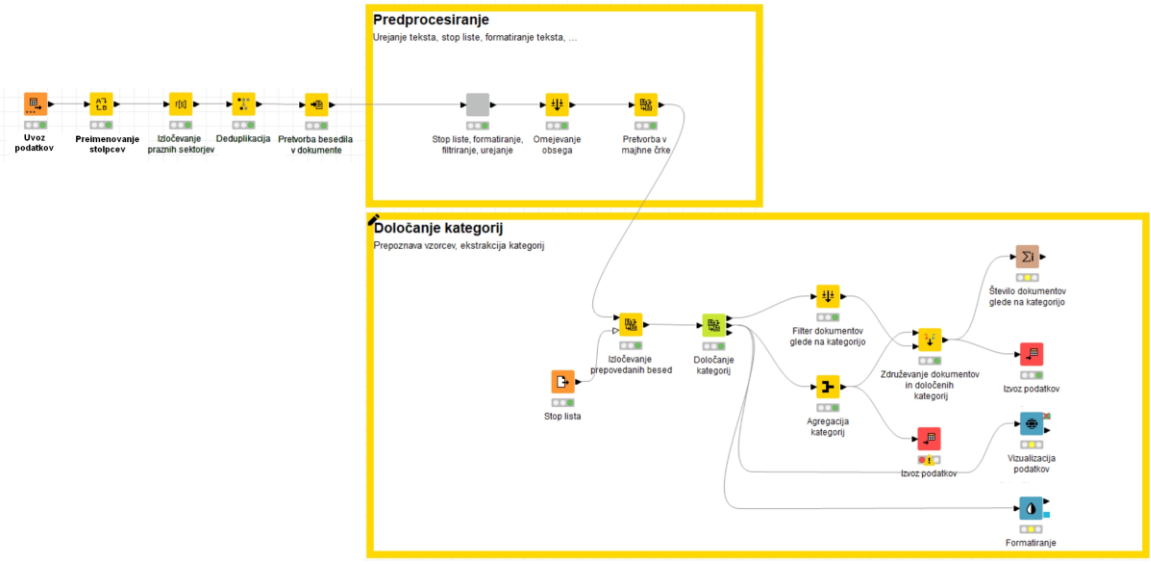
verjetnost v postopku določanja modela. Izdelava tematskih modelov tako predstavlja eno od možnih oblik prepoznave in kategorizacije vsebine, vseeno pa njena učinkovitost še vedno predstavlja določen izziv, s katerim se je treba soočiti.

Asmussen in Møller (2019) metodo izdelave tematskih modelov LDA opredeljujeta kot nenadzorovano verjetnostno metodo modeliranja, ki išče in določi teme iz individualnih besedil. Tema je opredeljena kot gruča, za katero je značilna porazdelitev glede na identificirane ključne besede. LDA analizira besede v vsakem izvirnem zapisu in izračuna skupno porazdelitev verjetnosti med opazovanimi elementi, tj. besedami v zapisu in neopazovanimi, skritimi oziroma latentnimi deli zapisov. Pri metodi LDA se praviloma uporablja dodatna metoda izdelave »vreče besed«, kjer se semantika in pomen stavkov ne ocenjujeta. Namesto tega metoda ocenjuje pogostost besed glede na dani zapis, zato se domneva, da bodo najpogostejše besede znotraj določene teme predstavljale tudi vsebino teme. Podobno opredelitev so podali tudi Blei idr. (2003, 993−1022), ki pravijo, da je metoda Latent Dirichlet Allocation namenjena izdelavi verjetnostnega generativnega modela podatkov, ki jih obdeluje.

Na sliki 21 je prikazan del krovnega delotoka izvedbe klasifikacije nestrukturiranih besedil na podlagi strojnega učenja, in sicer delotok za izdelavo tematskih modelov. Delotok je razdeljen na tri sekcije, ki so izvedbeno povezane:

- uvoz izvirnih zapisov v nestrukturirani obliki (besedila),
- izvedba podprocesa urejanja podatkov ter
- identifikacija ključnih besed, izdelava povezanih kategorij in izdelava tematskih modelov.

Slika 21: Delotok – izdelava tematskih modelov v orodju KNIME



Vir: KNIME 2024.

V tabeli 6 so dokumentirani koraki faze zajema vhodnih podatkov, vključno z vhodno vrednostjo in obliko ter izhodno vrednostjo in obliko ter parametri.

Tabela 6: Aktivnosti pri izdelavi tematskih modelov – zajem izvirnih zapisov

Faza	Opis	Vhodni/izhodni podatki
Zajem podatkov.	Zajem podatkov predstavlja uvoz posameznih nestrukturiranih zapisov (člankov) v digitalni obliki v orodje KNIME (2024). Izbrani format datotek za uvoz podatkov člankov je TXT (tekstovna datoteka) (TXT 2023). Datoteka, ki je bila uvožena, že vključuje agregirano obliko vseh individualni zapisov, in sicer vsak zapis v novi vrstici, vključno z identificiranimi imenskimi entitetami iz prvega krovnega delotoka.	*Vhodna oblika:* individualne datoteke v TXT-formatu (TXT 2023). *Izhodna oblika:* tabela z individualnimi zapisi, ki vključujejo besedilo posamezne tekstovne datoteke, vključno z identificiranimi imenskimi entitetami iz prvega krovnega delotoka.
Preimenovanje stolpca tabele.	Preimenovanje stolpca tabele, v okviru katerega dobi jasno oznako, da shranjuje besedilo posameznega uvoženega zapisa. Namenjeno je kasnejšemu lažjemu sledenju polni vsebini besedila, ki je bil prvotno uvoženo v delotok.	*Vhodna oblika:* tabela z besedili zapisov z generično imenovanim stolpcem. *Izhodna oblika:* tabela z besedili zapisov z jasno poimenovanim stolpcem, kjer se nahaja izvirna oblika besedila.
Izločevanje praznih sektorjev.	Urejanje teksta z namenom izločevanja praznih sektorjev v besedilu zapisa t. i. »white space«. Namen je izločiti dva ali več presledkov skupaj in prelome vrstic.	*Vhodna oblika:* tabela z besedilom zapisa, ki vključuje prazna območja v besedilu. *Izhodna oblika:* tabela z besedilom zapisa, ki ima iz besedila izločena prazna območja.
Deduplikacija vsebine.	Urejanje zapisov z namenom izločevanja identičnih zapisov. Namen je izločiti morebitno podvojeno/identično vsebino individualnih zapisov.	*Vhodna oblika:* tabela z besedili zapisov, ki lahko vključuje podvojene/identične zapise.

		Izhodna oblika: tabela z besedili zapisov, ki ima izločene podvojene/identične zapise.
Pretvorba besedila v dokumente.	V aktivnosti se izvede pretvorba izvirnih zapisov v programski objekt tipa »dokument«, ki omogoča lažjo obdelavo podatkov, posebej v delu možnega dodajanja opisnih vsebin.	*Vhodna oblika:* tabela z izvirnim besedilom zapisov. *Izhodna oblika:* tabela z izvirnim besedilom zapisov in objektno polje tipa »dokument«.

Vir: KNIME 2024.

Slika 22: Izdelava tematskih modelov – faza zajema podatkov

Vir: KNIME 2024.

Zajem vhodnih dokumentov oziroma zapisov (glej sliko 22) je lahko omejen na specifično vrsto datoteke, kot je npr. TXT (2023), lahko pa se uvozi kakršnokoli obliko, ki jo podpira funkcija uvoza in razčlenjevanja besedil v okviru orodja KNIME (2024).

Slika 23: Izdelava tematskih modelov – faza urejanja besedil (predprocesiranje podatkov)

Vir: KNIME 2024.

V tabeli 7 so dokumentirani koraki faze urejanja besedil (glej sliko 23) oziroma predprocesiranja podatkov, vključno z vhodno vrednostjo in obliko in izhodno vrednostjo in obliko ter parametri.

Tabela 7: Aktivnosti pri izdelavi tematskih modelov – urejanje besedil (predprocesiranje podatkov)

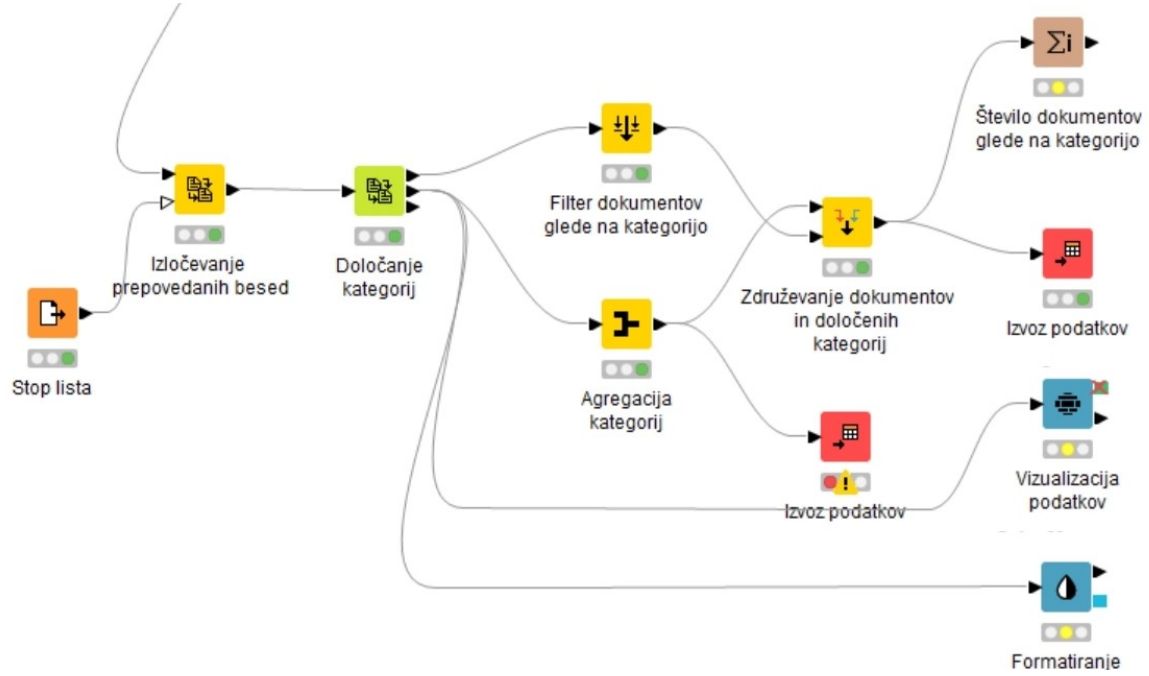
Faza	Opis	Vhodni/izhodni podatki
Urejanje besedil (izločevanje enojnih narekovajev, strukturnih oznak in ločil).	V aktivnosti se izvede obdelava besedila v objektnem polju tipa »dokument«. V besedilu zapisa se izloči identificirane strukturne oznake (npr. strukturne oznake), enojne narekovaje in ločila.	*Vhodna oblika:* tabela z izvirnim besedilom zapisov, vključno z objektnim poljem tipa »dokument«. *Izhodna oblika:* tabela z izvirnim besedilom zapisov, vključno s tabelo z obdelanim objektnim poljem tipa »dokument« in besedilom zapisa, ki ne vključuje identificiranih strukturnih oznak, enojnih narekovajev in ločil.
Izločitev številk.	Iz besedila izvirnega zapisa se izloči vse številke.	*Vhodna oblika:* tabela z izvirnim besedilom zapisov, vključno s tabelo s pretvorjenim objektnim poljem tipa »dokument« in besedilom zapisa, ki ne vključuje identificiranih strukturnih oznak, enojnih narekovajev in ločil. *Izhodna oblika:* tabela z izvirnim besedilom zapisov, vključno s tabelo s pretvorjenim objektnim poljem »dokument« in besedilom zapisa, ki ne vključuje številk.
Izločitev besed glede na njihovo dolžino.	Aktivnost, v okviru katere se iz besedila zapisa izloči besede, ki so krajše od predhodno določene dolžine.	*Vhodna oblika:* tabela z izvirnim besedilom zapisov, vključno s tabelo s pretvorjenim objektnim poljem tipa »dokument« in besedilom zapisa, ki ne vključuje številk. *Izhodna oblika:* tabela z izvirnim besedilom zapisov, vključno s tabelo s pretvorjenim

		objektnim poljem tipa »dokument« in besedilom zapisa, ki ne vključuje besed, ki so krajše od predhodno določene dolžine.
Pretvorba besedila iz velikih v male črke.	Aktivnost, v kateri se izvede pretvorba besedila v male črke.	*Vhodna oblika:* tabela z izvirnim besedilom zapisov, vključno s tabelo s pretvorjenim objektnim poljem tipa »dokument« in besedilom zapisa, ki ne vključuje besed, ki so krajše od predhodno določene, tj. nastavljene dolžine. *Izhodna oblika:* tabela z izvirnim besedilom zapisov, vključno s tabelo s pretvorjenim objektnim poljem tipa »dokument« in besedilom zapisa, kjer imajo vse besede v besedilu male črke.
Izločitev nedovoljenih besed.	Aktivnost, pri kateri se iz besedila zapisa izloči nedovoljene besede, ki niso predvidene kot del dovoljenih besed za identifikacijo in izdelavo tematskih modelov.	*Vhodna oblika:* predhodno pripravljen seznam – datoteka v tekstovnem formatu (tabela), ki vključuje individualne besede za izločevanje, in tabela z izvirnim besedilom zapisov, vključno s tabelo s pretvorjenim objektnim poljem tipa »dokument« in besedilom zapisa, ki ne vključuje besed, ki so krajše od določene dolžine. *Izhodna oblika:* tabela z izvirnim besedilom zapisov, vključno s tabelo s pretvorjenim objektnim poljem »dokument« in besedilom zapisa, ki ne vključuje nedovoljenih besed.
Izločitev nedovoljenih simbolov in znakov.	Aktivnost, v okviru katere se iz besedila izvirnega zapisa izloči nedovoljene znake in simbole.	*Vhodna oblika:* tabela z izvirnim besedilom zapisov, vključno s tabelo s pretvorjenim objektnim poljem tipa »dokument« in besedilom zapisa, ki ne vključuje nedovoljenih besed.

		Izhodna oblika: tabela z izvirnim besedilom zapisov, vključno s tabelo s pretvorjenim objektnim poljem »dokument« in besedilom zapisa, ki ne vključuje nedovoljenih simbolov in znakov.
Omejevanje obsega.	Opcijska aktivnost, v kateri se izloči vse podatke z izjemo pretvorjenega objektnega polja »dokument«.	*Vhodna oblika:* tabela z izvirnim besedilom zapisov, vključno s tabelo s pretvorjenim objektnim poljem »dokument« in besedilom zapisa, ki ne vključuje nedovoljenih simbolov. *Izhodna oblika:* tabela s pretvorjenim objektnim poljem tipa »dokument« in besedilom zapisa.

Vir: KNIME 2024.

Slika 24: Izdelava tematskih modelov – faza identifikacije vsebin in določanja tematskih modelov



Vir: KNIME 2024.

V tabeli 8 so dokumentirani koraki faze identifikacije vsebin in določanja tematskih modelov (glej sliko 24), vključno z vhodno vrednostjo in obliko in izhodno vrednostjo in obliko ter parametri.

Tabela 8: Aktivnosti pri izdelavi tematskih modelov – sklop identifikacije vsebin in določanja tematskih modelov

Faza	Opis	Vhodni/izhodni podatki
Izdelava tematskih modelov.	V tej aktivnosti sistem na podlagi predhodno določenih parametrov izvede analizo vsebine in identificira tematske modele glede na predhodno pripravljeno objektno polje tipa »dokument«. V okviru določitve parametrov je bil določen obseg pričakovanih tematskih modelov v obliki števila predvidenih tematskih modelov in števila ključnih besed, ki naj bi pripadale posameznemu identificiranemu tematskemu modelu. Parametri izdelave tematskih modelov so tako predvideli izdelavo petih tem in osmih besed na posamezno identificirano temo.	*Vhodna oblika:* tabela s pretvorjenim objektnim poljem tipa »dokument« in urejenim besedilom glede na aktivnosti iz faze predprocesiranja besedila. *Izhodna oblika:* tabela s pretvorjenim objektnim poljem tipa »dokument« in vsebino izvirnega zapisa, ki ima za vsak individualen izvirni zapis določen tematski model glede na izračunano verjetnost. Tabela z identificiranimi tematskimi modeli (temami) in naborom besed, ki so bile določene v okviru individualno identificiranega tematskega modela, ter izračunano utežjo verjetnosti ključnih besed, ki so določene v individualnem tematskem modelu.
Združevanje besed posameznega tematskega modela.	V aktivnosti združevanja se izvede izdelovanje gruč ključnih besed v povezavi z individualno identificiranim tematskim modelom.	*Vhodna oblika:* tabela z identificiranimi tematskimi modeli (temami) in naborom besed, ki so bile določene v okviru individualno identificiranega tematskega modela, ter izračunano utežjo verjetnosti ključnih besed, ki so določene v individualnem tematskem modelu. *Izhodna oblika:* tabela z oznakami tematskih modelov in naborom besed v identificiranem zaporedju, ki so določeni, da pripadajo

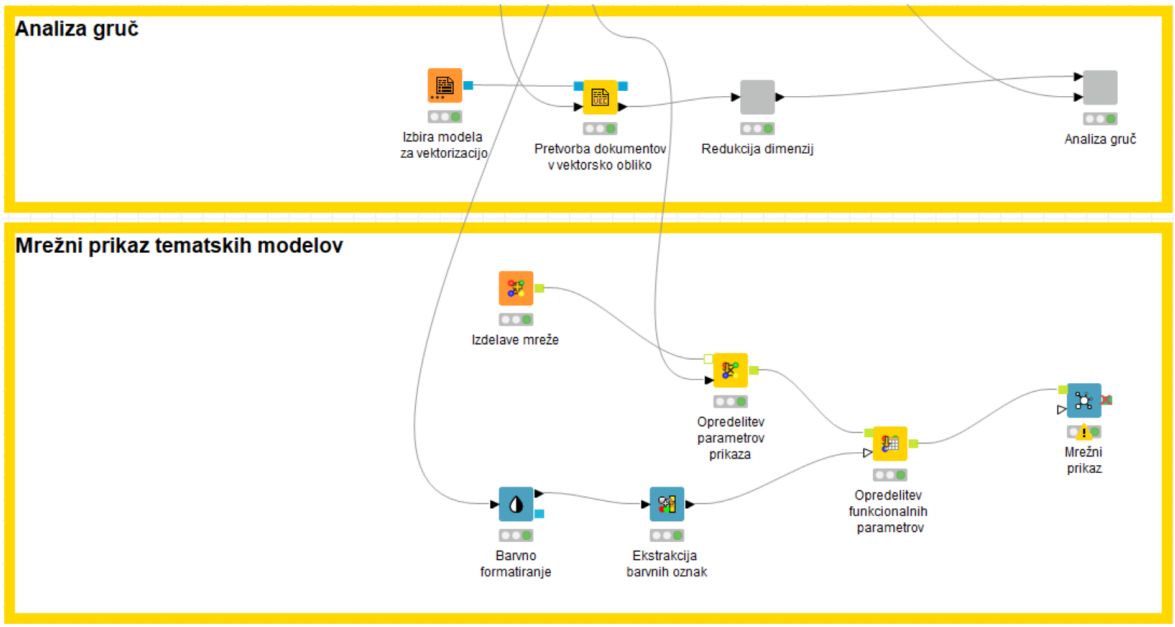
		individualnemu tematskemu modelu.
Grafični prikaz podatkov.	V aktivnosti se izvede vizualizacija določenih tematskih modelov.	*Vhodna oblika:* tabela z oznakami tematskih modelov in naborom besed, ki pripadajo posamezni temi, ter indeks verjetnosti posamezne besede, da pripada posameznemu tematskemu modelu. *Izhodna oblika:* grafični prikaz izdelanih tematskih modelov.
Združevanje izvirnih zapisov z izdelanimi tematskimi modeli.	Združevanje izvirnih zapisov z izdelanimi in določenimi tematskimi modeli.	*Vhodna oblika:* tabela z izvirnimi zapisi in tabela s pretvorjenim objektnim poljem tipa »dokument« in besedilom izvirnega zapisa, ki ima za vsak posamezen dokument/zapis določeno oceno verjetnosti glede na posamezne identificirane teme in določeno končno temo za posamezen dokument/zapis. *Izhodna oblika:* tabela z izvirnimi zapisi in besedili izvirnih zapisov, ki ima za vsak individualen zapis določen tematski model.
Izvoz podatkov izvirnih zapisov.	V aktivnosti se izvede izvoz izvirnih zapisov z identificiranimi tematskimi modeli.	*Vhodna oblika:* tabela z izvirnimi zapisi in besedilom izvirnega zapisa, ki ima za vsak individualen izvirni zapis določen tematski model ter identificirane imenske entitete. *Izhodna oblika:* CSV (2023) – datoteka, ki vključuje rezultat delotoka izdelave tematskih modelov.
Izvoz podatkov identificiranih tematskih modelov.	V aktivnosti se izvede izvoz zapisov določenih tematskih modelov.	*Vhodna oblika:* tabela z oznakami tem in seznamom besed v zaporednem zapisu, ki

		pripadajo individualno identificiranemu tematskemu modelu. *Izhodna oblika:* CSV (2023) – datoteka z vsebino, iz katere so razvidni določeni tematski modeli.

Vir: KNIME 2024.

Na sliki 25 je prikazan delotok za izdelavo analize gruč ter mrežnega prikaza ključnih besed v okviru identificiranih tematskih modelov.

Slika 25: Analiza gruč ter mrežnega prikaza ključnih besed v okviru identificiranih tematskih modelov



Vir: KNIME 2024.

V tabeli 9 so dokumentirani koraki izdelave analize gruč ter mrežnega prikaza ključnih besed v okviru identificiranih tematskih modelov.

Tabela 9: Izdelava analize gruč ter mrežnega prikaza ključnih besed v okviru identificiranih tematskih modelov

Faza	Opis	Vhodni/izhodni podatki
Izbira predizdelanega modela za izvedbo vektorizacije (analiza gruč).	V tej fazi se izvede izbira predizdelanega modela za vektorizacijo dokumentov. Uporabljeni model:	*Vhodna oblika:* / *Izhodna oblika:* izbrani predizdelani model za izvedbo vektorizacije.

	- en_core_web_sm (En-core-web-sm 2023).	
Pretvorba dokumentov v vektorsko obliko (analiza gruč).	V aktivnosti se izvede pretvorba izvirnih zapisov (dokumentov) v vektorsko obliko na podlagi predizbranega modela.	*Vhodna oblika:* izbrani predizdelani model za vektorizacijo dokumentov in tabela z izvirnimi zapisi v obliki objektnega polja »dokument«. *Izhodna oblika:* tabela z izvirnimi zapisi v obliki objektnega polja »dokument« in vektorska oblika dokumentov (zapisov).
Redukcija dimenzij (analiza gruč).	V tej fazi se izvede zmanjševanje dimenzij in predpriprava za vizualizacijo z uporabo naslednjih pristopov: - Principal – component analysis PCA (Jolliffe in Cadima 2016); - t-distributed stochastic neighbor embedding T-SNE (Maaten in Hinton 2008).	*Vhodna oblika:* tabela z izvirnimi zapisi v obliki objektnega polja »dokument« in vektorska oblika dokumentov (zapisov). *Izhodna oblika:* tabela z izvirnimi zapisi v obliki objektnega polja »dokument« in vektorska oblika dokumentov (zapisov) z izračunanimi dimenzijami v obliki določenih gruč.
Vizualizacija analize gruč (analiza gruč).	V tej aktivnosti se izvede prikaz izdelanih gruč tematskih modelov in ključnih besed na podlagi zmanjšanih dimenzij.	*Vhodna oblika:* tabela z izvirnimi zapisi v obliki objektnega polja »dokument« in vektorska oblika dokumentov (zapisov) z izračunanimi dimenzijami v obliki določenih gruč. *Izhodna oblika:* grafični prikaz izdelanih gruč tematskih modelov in ključnih besed na podlagi zmanjšanih dimenzij.

Izdelava mreže (mrežni prikaz tematskih modelov).	Pri izdelavi mreže se izvede inicializacija mreže z osnovnimi parametri za izgradnjo končnega prikaza.	*Vhodna oblika:* / *Izhodna oblika:* osnovna mreža z osnovnimi parametri.
Opredelitev parametrov prikaza (mrežni prikaz tematskih modelov).	V aktivnosti se izvede nastavitev parametrov prikaza za izdelavo mrežne postavitve.	*Vhodna oblika:* seznam ključnih besed znotraj individualno določenega tematskega modela. *Izhodna oblika:* osnovna mreža s posodobljenimi parametri.
Opredelitev funkcionalnih parametrov (mrežni prikaz tematskih modelov).	V aktivnosti se izvede nastavitev funkcionalnih parametrov prikaza za izdelavo mrežne postavitve.	*Vhodna oblika:* seznam ključnih besed znotraj individualno določenega tematskega modela, vključno z osnovno mrežo s posodobljenimi parametri. *Izhodna oblika:* končna mreža, pripravljena za vizualizacijo.
Prikaz rezultatov (mrežni prikaz tematskih modelov).	V aktivnosti se izvede vizualna predstavitev rezultatov delotoka izdelanih tematskih modelov.	*Vhodna oblika:* končna mreža, pripravljena za vizualizacijo. *Izhodna oblika:* grafični prikaz rezultatov – tematskih modelov in pripadajočih ključnih besed.

Vir: KNIME 2024.

3.3.7 Učenje prepoznavanja nestrukturiranih vsebin in izdelava izvedbenega modela klasifikacije vsebin

V drugem delotoku je bilo za potrebe klasifikacije nestrukturiranega besedila uporabljenih več pristopov. Z dvema pristopoma je bil izdelan učeči se prediktivni model, z enim pristopom pa je bil uporabljen večji predizdelan generativni model.

Za vzpostavitev dveh učečih se prediktivnih modelov sta bili uporabljeni naslednji dve metodi:

- metoda odločitvenih dreves (angl. Decision tree learning) (Karabadji idr. 2017) in
- metoda podpornih vektorjev (angl. Support vector machine − SVM) (Cervantes idr. 2020).

Za vzpostavitev pristopa, kjer je bil uporabljen večji predizdelan generativni model, je bil uporabljen pristop z uporabo odprtokodne platforme za strojno učenje Tensorflow (Tensorflow 2023) ter t. i. arhitekture oziroma modela BERT (2023) z uporabo predizdelanega modela iz knjižnice TensorFlow Hub (Tensorflow 2023).

Po besedah Akerkarja (2019, 24) odločitveno drevo deluje kot drevesna struktura z začetkom na vrhu, ki je določen kot osnovno vozlišče in se razteza do listov, medtem ko vsaka veja predstavlja rezultat testiranja. Končna vozlišča predstavljajo klasifikacijske razrede. V primeru, da želimo prepoznati neznan vzorec, sistem preveri lastnosti vzorca v skladu z izbranim odločitvenim drevesom. V nadaljevanju se izvaja obdelava od korena, tj. vrha drevesa, pa vse do končnih vozlišč (listov), ki določajo klasifikacijsko skupino za vzorec, ki je predmet obdelave. Pri uporabi metode odločitvenega drevesa se posledično lahko pojavljajo težave z vidika kakovosti podatkov, ki se rešujejo s pristopom, ki omogoča t. i. obrezovanje. V kontekstu obrezovanja gre za izločevanje kakršnihkoli lastnosti, ki ne nudijo boljšega oziroma bolj optimalnega pristopa k določanju pravil.

Akerkar (24) nadalje navaja, da se pri uporabi metode odločitvenega drevesa lahko uporabi podatke, kot so na primer nestrukturirana besedila, ki so lahko podlaga za izdelavo izvedbenega odločitvenega modela za izvedbo klasifikacije, čemur nato sledi iskanje ključnih terminološko opredeljenih indeksov, kot so identificirani tematski modeli, ki posledično tvorijo osnovo za sprejemanje odločitev v skladu s pričakovanimi cilji.

V primeru, da določen del obdelanih podatkov (izvirni zapis) vključuje določene besede, kot jih v naboru ključnih besed opredeljuje predhodno določen tematski model, se lahko takšna obravnava loči v svojo vejo odločitvenega drevesa vse do zadnje še razdeljive opcije, ki predstavlja končno opredeljeno klasifikacijo izvirnega zapisa.

Akerkar (25) omenja tudi določene pomanjkljivosti oziroma nekonsistentnosti pri uporabi metode odločitvenih dreves, predvsem ko gre za obdelavo velikih količin podatkov, kjer se odzivnost in učinkovitost sistema, ki uporablja metodo odločitvenih dreves, zmanjšata, zaradi česar so tudi rezultati slabši. Ker se vse operacije izvajajo znotraj povezanih aktivnosti, se vsi podatki hranijo neposredno v pomnilniku, kar predstavlja neoptimizirano obliko obdelave podatkov.

Metoda podpornih vektorjev predstavlja algoritem strojnega učenja, ki uporablja modele nadzorovanega učenja za reševanje težav s kompleksno klasifikacijo, regresijo in zaznavanjem izstopnih vrednosti z izvajanjem optimalnih transformacij podatkov, ki določajo meje med podatkovnimi točkami na osnovi vnaprej določenih razredov ali oznak (Kanade 2022). Metoda podpornih vektorjev se pogosto uporablja pri obdelavi naravnega jezika. Podobno Ploj (2013, 9) izpostavlja, da je pri umeščanju zapisov v razrede izjemno koristno, če je identificirana možnost, da se podatki med seboj lahko linearno ločijo. Pri

metodi podpornih vektorjev se podatki transformirajo v obliko, ki omogoča linearno ločljivost in »pri tem upošteva le učne vzorce, ki so blizu odločitvene meje in jih imenuje podporni vektorji. Med preslikavo se prosti parametri jedrnih funkcij izberejo tako, da dobimo čim širši rob (angleško margin) med podatki različnih razredov« (Ploj 2013, 9).

Gonzáles in Garrido (2021) navajata, da je BERT (angl. Bidirectional Encoder Representations from Transformers) model NLP, ki je bil zasnovan za predhodno usposabljanje globokih dvosmernih pomenov iz neoznačenega besedila z možnostjo, da se po usposabljanju izvedejo dodatne natančne nastavitve za izboljševanje modela (angl. Fine tuning) z uporabo označenega besedila za različne naloge NLP.

> »TensorFlow je sistem strojnega učenja, ki deluje v velikem obsegu in v heterogenih okoljih. Ta arhitektura daje razvijalcu aplikacije prilagodljivost ter omogoča eksperimentiranje z novimi metodami optimizacije in algoritmi za učenje. TensorFlow podpira različne aplikacije, s poudarkom na učenju in sklepanju z uporabo globokih nevronskih mrež. Izdan je kot odprtokodni projekt in se pogosto uporablja za raziskave strojnega učenja« (Tensorflow 2023).

»TensorFlow Hub je repozitorij usposobljenih (naučenih) modelov, ki so bili predizdelani z uporabo strojnega učenja in ki so pripravljeni za natančno prilagajanje ter jih je mogoče uporabiti kjerkoli« (Tensorflow 2023). Namen repozitorija je uporaba že izdelanih modelov, kot je npr. BERT (2023), ki omogoča prilagajanje individualnim potrebam glede na zahteve po obdelavi podatkov (angl. Fine tunning).

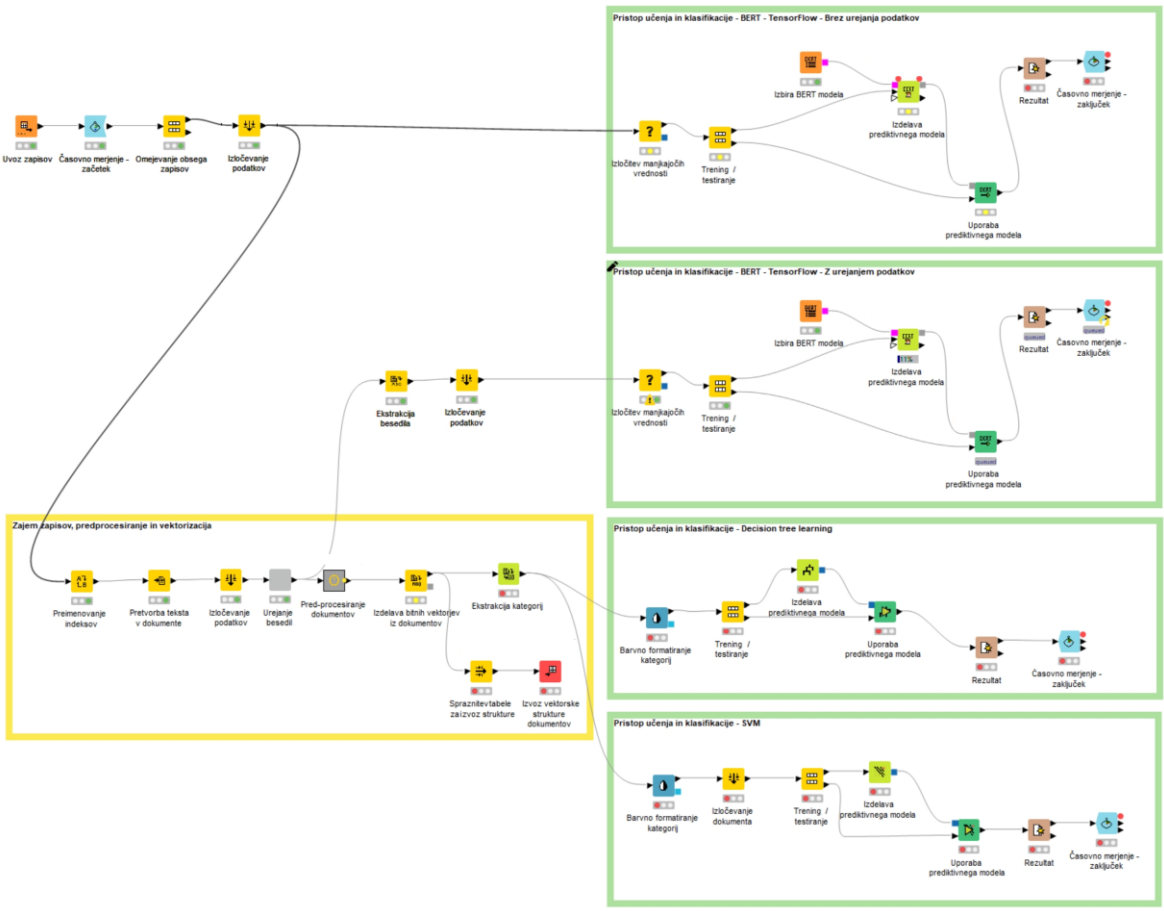
Za uspešno nadaljevanje izvedbe uporabe predizdelanih modelov ali izdelavo novih modelov je bila obvezna izvedba predhodnega delotoka, namenjenega izdelavi tematskih modelov. Izhodna vrednost prvega delotoka je predstavljala vhodno vrednost za obdelavo v drugem delotoku klasifikacije nestrukturiranih vsebin.

Na sliki 26 je predstavljen sklop aktivnosti oziroma delotok učenja prepoznavanja nestrukturirane vsebine in izdelave izvedbenega modela klasifikacije vsebine na podlagi tematskih modelov. Proces oziroma delotok je sestavljen iz naslednjih sklopov:

- zajem vhodnih zapisov,
- predprocesiranje podatkov (podproces je bil vključen tako v delu uporabe predizdelanega modela kot tudi v delu uporabe metode odločitvenih dreves (angl. Decision tree learning) (Karabadji idr. 2017) in metode podpornih vektorjev (angl. Support vector machine − SVM) (Cervantes idr. 2020) ter
- treniranje/učenje in izdelava izvedbenih odločitvenih modelov.

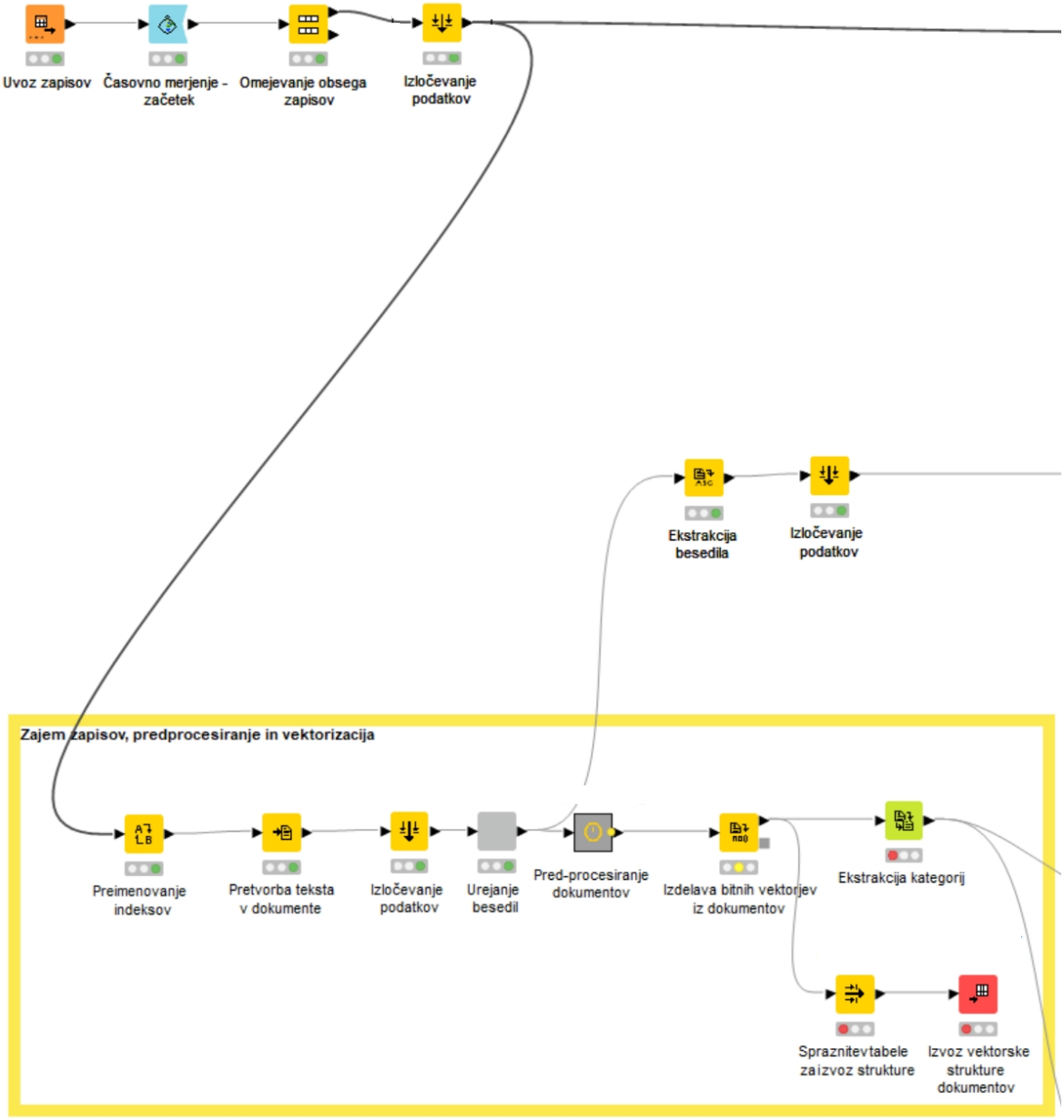
Ker je tretji pristop z uporabo BERT (2023) in Tensorflow (2023) vključeval uporabo predizdelanega modela iz knjižnice TensorHub (Tensorflow 2023), je bil namen, da se preveri, ali bi lahko podproces predprocesiranja podatkov vplival na učinkovitost klasificiranja v primerjavi s pristopom brez izvedenega podprocesa predprocesiranja podatkov.

Slika 26: Celoten delotok vzpostavitve izvedbenega modela za klasifikacijo nestrukturiranih vsebin



Vir: KNIME 2024.

Slika 27: Zajem zapisov z nestrukturirano vsebino ter predprocesiranje in vektorizacija v primeru uporabe metod BERT, Decision tree learning in Support vector machine – SVM



Vir: KNIME 2024.

V tabeli 10 so predstavljene aktivnosti posameznih sklopov v delu zajema izvirnih zapisov, kot so prikazane na sliki 27.

Tabela 10: Zajem zapisov z nestrukturirano vsebino ter predprocesiranje in vektorizacija v primeru uporabe metod BERT, Decision tree learning in Support vector machine – SVM

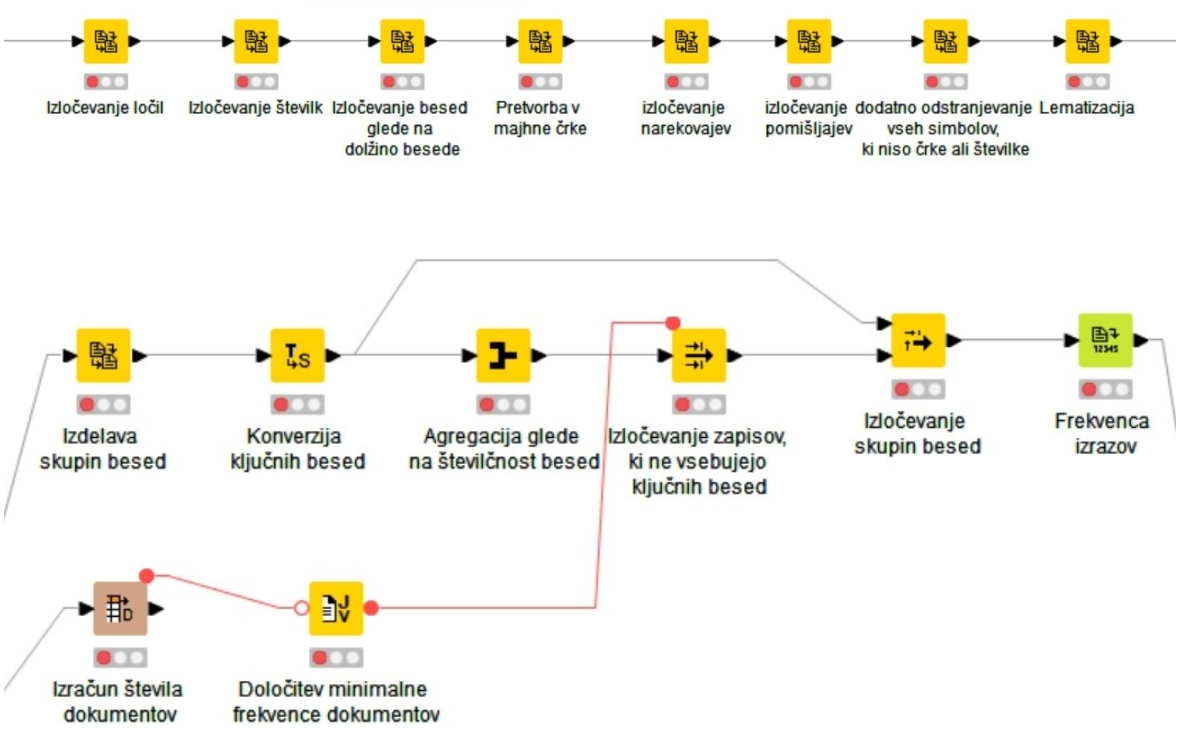
Faza	Opis	Vhodni/izhodni podatki
Zajem zapisov z nestrukturirano vsebino in določenimi tematskimi modeli ter identificiranimi imenskimi entitetami.	Zajem zapisov z nestrukturirano vsebino predstavlja aktivnost, v kateri se izvede uvoz individualnih zapisov z določenimi tematskimi modeli, vključno s seznamom ključnih besed in identificiranimi imenskimi entitetami. Format datotek za uvoz podatkov je CSV (2023), ki je bil izbran in predstavlja rezultat delotoka za izdelavo in določanje tematskih modelov in delotoka za prepoznavo imenskih entitet.	*Vhodna oblika:* individualni zapisi, agregirani v CSV-formatu (CSV 2023). *Izhodna oblika:* tabela z besedilom zapisa, določenimi tematskimi modeli, seznamom ključnih besed, prepoznanimi imenskimi entitetami in ostalimi podatki iz prejšnjih delotokov.
Časovno merjenje – začetek.	Aktivnost je namenjena merjenju časa, ki je potreben za izvedbo delotoka prepoznavanja imenskih entitet. Aktivnost ne vpliva na vsebinsko ali strukturno obdelavo podatkov in ne spreminja obstoječih podatkov.	
Omejevanje obsega zapisov.	V tej fazi je bila izvedena omejitev na 50.000 individualnih zapisov. Aktivnost ni v ničemer vplivala na strukturo ali vsebino podatkov in ne spreminja obstoječih podatkov.	
Izločevanje podatkov.	V okviru aktivnosti je bilo izvedeno izločevanje nepotrebnih podatkov iz prejšnjih delotokov. Omejitev je zajemala le besedilo izvirnega zapisa, izdelane tematske modele, seznam ključnih besed, identificiranih pri individualnem tematskem modelu, ter prepoznane imenske entitete.	*Vhodna oblika:* tabela z besedilom zapisa, določenimi tematskimi modeli, seznamom ključnih besed, prepoznanimi imenskimi entitetami in ostali podatki iz prejšnjih delotokov. *Izhodna oblika:* tabela z besedilom zapisa, določenimi tematskimi modeli, seznamom ključnih besed ter prepoznanimi imenskimi entitetami.

Preimenovanje indeksov.	V aktivnosti je bilo izvedeno preimenovanje določenih indeksov za lažjo nadaljnjo obdelavo podatkov.	*Vhodna oblika:* tabela z besedilom zapisa, določenim tematskim modelom, seznamom ključnih besed ter prepoznanimi imenskimi entitetami. *Izhodna oblika:* tabela s preimenovanimi indeksi, besedilom zapisa, določenimi tematskimi modeli, seznamom ključnih besed ter prepoznanimi imenskimi entitetami.
Pretvorba besedila v dokumente.	Zaradi lažjega programskega obdelovanja besedila so bili zapisi z besedilom pretvorjeni v programski objekt, imenovan »dokument«, ki omogoča lažjo obdelavo podatkov, posebej v delu možnega dodajanja opisnih vsebin.	*Vhodna oblika:* tabela s preimenovanimi indeksi, besedilom zapisa, določenimi tematskimi modeli, seznamom ključnih besed ter prepoznanimi imenskimi entitetami. *Izhodna oblika:* tabela z besedilom zapisa, določenimi tematskimi modeli, seznamom ključnih besed ter prepoznanimi imenskimi entitetami, vključno z objektnim poljem tipa »dokument«.

Vir: KNIME 2024.

V primeru uporabe metod BERT (2023), Decision tree learning (Karabadji idr. 2017) in Support vector machine – SVM (Cervantes idr. 2020) je bil pred izvedbo pretvorbe izvirnih zapisov v vektorsko obliko dodan podpodproces predprocesiranje izvirnih zapisov (glej sliko 28).

Slika 28: Urejanje podatkov – izvedba predprocesiranja za namen zagotavljanja višje kakovosti podatkov



Vir: KNIME 2024.

V tabeli 11 so dokumentirane posamezne aktivnosti znotraj izvedbe podprocesa urejanja podatkov, kot so prikazane na sliki 28.

Tabela 11: Urejanje besedil v delotoku učenja

Faza/-e	Opis	Vhodni/izhodni podatki
Podproces urejanja besedil v prvem delu vključuje aktivnosti, kot so izločevanje ločil, krajšanje besed, izločevanje, številk in pretvorba besedila iz velikih v male črke. Faza je namenjena zagotavljanju višje kakovosti podatkov.	Začetne aktivnosti vključujejo podobne aktivnosti, kot so bile izvedene v okviru delotoka in urejanja besedil pri izdelavi tematskih modelov. Namen urejanja besedil je omejiti nekakovostne informacije/podatke in zagotoviti nabor besed ter urediti besedilo do te mere, da bo lažja izvedba izdelave odločitvenega modela.	*Vhodna oblika:* tabela z besedilom zapisa, določenimi tematskimi modeli, seznamom ključnih besed ter prepoznanimi imenskimi entitetami, vključno z objektnim poljem »dokument«. *Izhodna oblika:* tabela s pretvorjenim objektnim poljem tipa »dokument« in obdelanim besedilom zapisa, ki ne vključuje velikih črk, ločil, številk in daljših besed, kot je dovoljeno.

Izdelovanje lem (kakovost podatkov).	Izvede se krnjenje posameznih besed na njihovo osnovno obliko.	*Vhodna oblika:* tabela s pretvorjenim objektnim poljem tipa »dokument« in obdelanim besedilom zapisa, ki ne vključuje velikih črk, ločil, številk in daljših besed, kot je dovoljeno. *Izhodna oblika:* tabela s pretvorjenim objektnim poljem »dokument« in besedilom zapisa, ki ne vključuje predhodno izločenih znakov, vključuje pa besede v osnovni krnjeni obliki.
Ekstrakcija besedila iz programskega objekta »dokument« (samo v primeru klasifikacije BERT (2023)).	Izvede se ekstrakcija besedila, ki vključuje urejeno besedilo izvirnih zapisov zaradi izvedbe delotoka izdelave modela in klasificiranja z uporabo arhitekture in modela BERT (2023).	*Vhodna oblika:* tabela s pretvorjenim objektnim poljem tipa »dokument« in urejenim besedilom zapisa, ki vključuje predhodno urejene besede v osnovni obliki. *Izhodna oblika:* tabela z izvirnimi zapisi ter prepoznanim in urejenim besedilom zapisa v obliki dokumenta in v obliki navadnega besedila (zapisa).
Izločevanje podatkov (samo v primeru klasifikacije BERT (2023)).	Izvedeno je bilo izločevanje nepotrebnih podatkov iz prejšnjih delotokov, kar pomeni, da omejitev zajema le besedilo, vključeno v izvirnih zapisih, in izdelane tematske modele, ki pripadajo posameznemu individualnemu zapisu.	*Vhodna oblika:* tabela z besedilom zapisa ter določenimi tematskimi modeli in ostalimi podatki iz prejšnjih delotokov. *Izhodna oblika:* tabela z besedilom zapisa in določenimi tematskimi modeli.
Izdelava skupin (vreč) besed.	V aktivnosti se izvede izdelava skupin (vreč) ključnih besed, ki se glede na pogostost pojavljajo v individualnem izvirnem zapisu.	*Vhodna oblika:* tabela s pretvorjenim objektnim poljem tipa »dokument« in urejenim besedilom zapisa, ki vključuje besede v osnovni obliki.

		Izhodna oblika: tabela s pretvorjenim objektnim poljem »dokument« in urejenim besedilom zapisa, vključno z izdelanimi skupinami besed.
Izločevanje skupin (vreč) besed.	Izvede se združevanje identificiranih skupin (vreč) besed glede na število vseh zapisov in pojavnost prepoznanih ključnih besed v posameznih izvirnih zapisih. Dodatno se izvede filtriranje skupin (vreč) besed skladno z relativno pojavnostjo besed v izvirnih zapisih.	*Vhodna oblika:* tabela s pretvorjenim objektnim poljem »dokument« in urejenim besedilom zapisa, vključno z izdelanimi skupinami besed. *Izhodna oblika:* tabela s pretvorjenim objektnim poljem tipa »dokument«, urejenim besedilom zapisa, vključno z izdelanimi skupinami besed, ter indeksom, nanašajočim se na relativnost ugotovljenih skupin (vreč) besed.

Vir: KNIME 2024.

Po zaključenem sklopu predprocesiranja podatkov sta bili izvedeni vektorizacija vsebine in izdelava tematskih modelov kot zadnji del sklopa za urejanje podatkov z namenom izdelave modela in klasifikacije z uporabo arhitekture in modela Decision tree learning (Karabadji idr. 2017) in Support vector machine – SVM (Cervantes idr. 2020) (glej sliko 26).

V primeru izdelave modela in klasifikacije z uporabo arhitekture in modela BERT (2023) s predhodno urejenimi podatki koraki v delotoku sledijo enakim korakom, kot so dokumentirani v delu izvedbe klasifikacije z metodo BERT (2023) brez urejenih podatkov. Veja delotoka je preusmerjena takoj po izvedbi aktivnosti urejanja podatkov (kakovost podatkov).

V tabeli 12 so dokumentirane individualne aktivnosti v prvem sklopu delotoka zajema izvirnih zapisov in izvedbe urejanja podatkov, kot so predstavljene na sliki 26.

Tabela 12: Ekstrakcija tematskih modelov

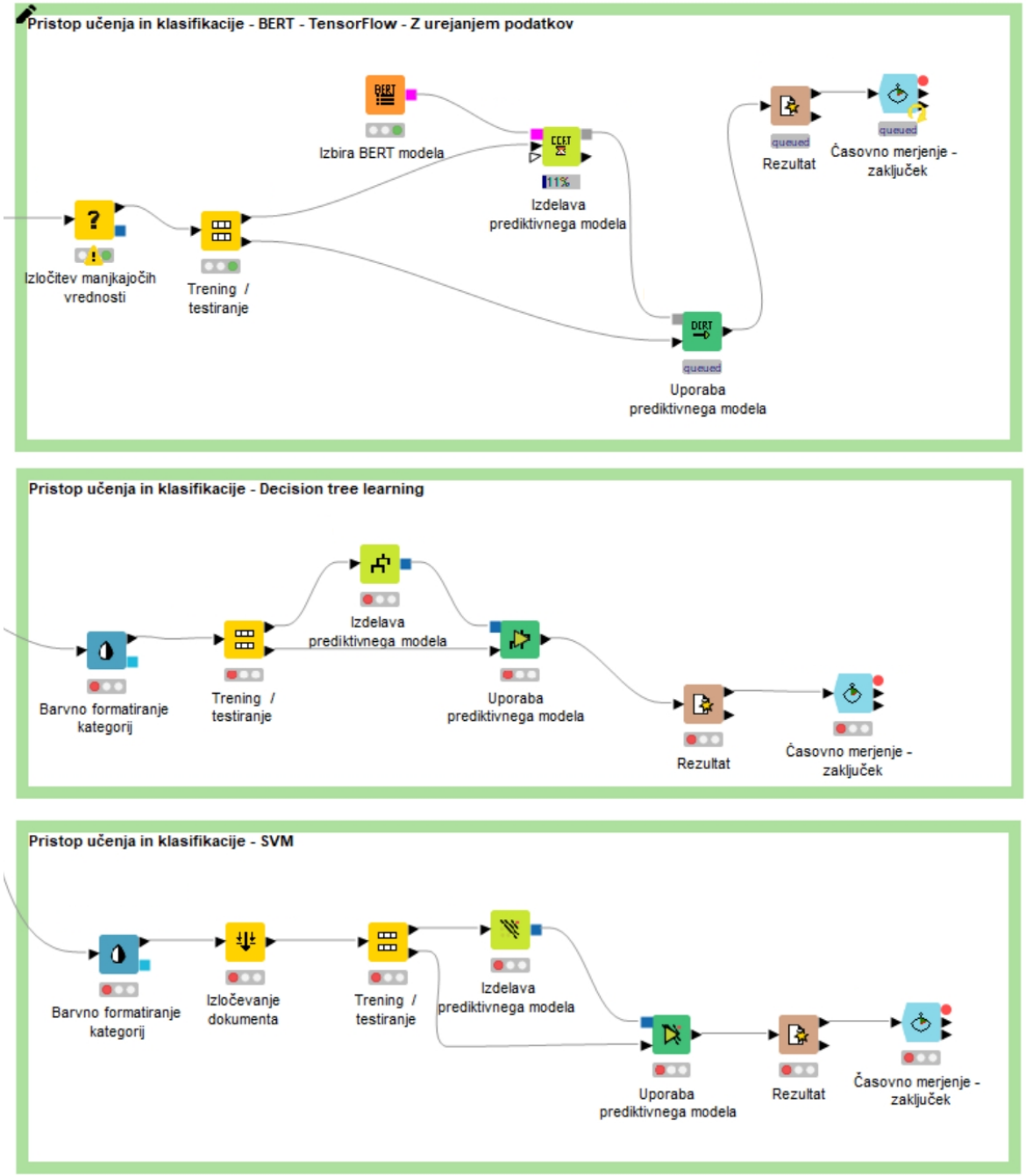
Faza	Opis	Vhodni/izhodni podatki
Vektorizacija podatkov.	V aktivnosti se izvede pretvorba zapisov v večdimenzijsko obliko, tj. bitne vektorje.	*Vhodna oblika:* tabela s pretvorjenim objektnim poljem tipa »dokument«, urejenim besedilom zapisa, vključno z izdelanimi skupinami besed, ter indeksom,

		nanašajočim se na relativnost ugotovljenih skupin (vreč) besed. *Izhodna oblika:* tabela s pretvorjenim objektnim poljem tipa »dokument« in določenimi indeksi identificiranih ključnih besed, ki so bile pretvorjene v vektorsko obliko glede na posamezen zapis in model, uporabljen za izvedbo vektorizacije.
Izvoz tematskih modelov.	V aktivnosti se izvede izvoz tematskih modelov, ki so bili določeni v prvem delotoku, z namenom izdelave izvedbenega modela za klasifikacijo vsebin. Izvoženi tematski modeli predstavljajo predvideno kategorizacijo za izvedbo klasifikacije, torej predstavljajo klasifikacijske oznake za izvedbo končne klasifikacije.	*Vhodna oblika:* tabela s pretvorjenim objektnim poljem tipa »dokument« in določenimi indeksi identificiranih ključnih besed v vektorski obliki glede na posamezen zapis in model, uporabljen za izvedbo vektorizacije, vključno z določenimi parametri za izdelavo tematskega modela. *Izhodna oblika:* tabela s pretvorjenim objektnim poljem tipa »dokument« ter vektoriziranimi podatki in določenimi tematskimi modeli glede na individualen zapis.
Izvoz podatkov v vektorski obliki.	V aktivnosti se izvede izvoz podatkov v tabelarično obliko. Podatki se lahko uporabijo za nadaljnjo obdelavo na področju obdelave naravnega jezika.	*Vhodna oblika:* tabela s pretvorjenim objektnim poljem tipa »dokument« ter vektoriziranimi podatki in identificiranimi tematskimi modeli glede na posamezen zapis. *Izhodna oblika:* Prazna tabela kot programski objekt in strukturirana shema.

Vir: KNIME 2024.

V zadnjem delu krovnega delotoka za izvedbo klasifikacije nestrukturiranih besedil z uporabo strojnega učenja se je za zastavljene tri pristope, ki so del delotoka, izvedel postopek izdelave izvedbenega odločitvenega modela na podlagi učenja in izpeljala klasifikacija nestrukturiranih zapisov z uporabo izdelanega odločitvenega modela glede na predhodno določene tematske modele (glej sliko 29).

Slika 29: Izdelava izvedbenih odločitvenih modelov za klasifikacijo nestrukturiranih besedil



Vir: KNIME 2024.

V tabeli 13 so po individualnih korakih predstavljene aktivnosti za različne tri pristope izdelave odločitvenega modela skladno s predhodno identificiranimi tematskimi modeli in izvirnimi zapisi (glej sliko 29).

Tabela 13: Postopek izdelave odločitvenih modelov

Faza	Opis	Vhodni/izhodni podatki
Ločevanje podatkov na podatke, namenjene učenju, in na podatke, namenjene preizkušanju izdelanega modela.	V aktivnosti se izvede ločevanje podatkov oziroma izvirnih posameznih zapisov na del podatkov, ki je namenjen in uporabljen za učenje iz vsebine in izdelavo modela za izvedbo klasifikacije, in na del podatkov, ki je namenjen in uporabljen za preizkušanje izdelanega odločitvenega modela. Podatki so ločeni v razmerju 70 % za izvedbo učenja ter 30 % za izvedbo preizkušanja – izvirni zapisi so bili izbrani po načelu naključnih zapisov.	*Vhodna oblika:* tabela s pretvorjenim objektnim poljem tipa »dokument« in izdelanimi podatki v vektorski obliki, vključno z določenim tematskim modelom za individualen izvirni zapis. *Izhodna oblika:* tabela s podatki, ki so pripravljeni za izdelavo odločitvenega modela za izvedbo učenja v t. i. ločenem razmerju za izvedbo učenja. Tabela z izvirnimi zapisi, ki so namenjeni testiranju natančnosti in učinkovitosti izdelanih odločitvenih modelov.
Izbira predizdelanega modela – le v primeru uporabe metode BERT (2023).	Korak delotoka je omejen le na obdelavo podatkov z uporabo metode BERT – Tensorflow (BERT 2023; Tensorflow 2023). V fazi se izvede izbira predizdelanega modela iz knjižnice HuggingFace (2024), ki je prosto dostopna na spletu. Uporabljena sta bila modela Roberta-large (2023) in Bert-base-multilingual-cased (2023), ki sta optimizirana modela za naravno jezikovno obdelavo podatkov (NLP).	*Vhodna oblika:* določitev parametra za izbiro predizdelanega modela iz knjižnice HuggingFace (2024). *Izhodna oblika:* uvožen in pripravljen predizdelan model za izvedbo izdelave odločitvenega modela.
Izdelava prediktivnega	V tem koraku oziroma aktivnosti se izvede učenje na obsegu podatkov (izvirnih	*Vhodna oblika:* tabela s podatki, ki so namenjeni učenju in izdelavi odločitvenih

odločitvenega modela.	zapisov), ki so pripravljeni za izvedbo učenja, in na podlagi predhodno določene klasifikacije v obliki identificiranih tematskih modelov. V primeru uporabe metode BERT (2023) se je izdelava odločitvenega modela izvedla z uporabo predizdelanega modela.	modelov za izvedbo klasifikacije in v primeru uporabe metode BERT (2023), in izbrani predizdelan model. *Izhodna oblika:* izdelan izvedbeni odločitveni model, vključno s specifikacijo za izvedbo samostojne klasifikacije vsebine, ki je pripravljen za neposredno izvedbo klasifikacije besedil.
Uporaba izdelanega prediktivnega odločitvenega modela.	V aktivnosti se z uporabo izdelanega prediktivnega odločitvenega modela izvede klasificiranje podatkov v obsegu, kot je bil določen v fazi ločevanja na del, nanašajoč se na izdelavo modela, in na del, nanašajoč se na preizkušanje učinkovitosti izdelanega modela. Namen aktivnosti je pridobiti podatke za izvedbo meritve učinkovitosti in uspešnosti izvedene klasifikacije glede na predhodno razpoložljive podatke o klasifikacijskih oznakah individualnega izvirnega zapisa.	*Vhodna oblika:* izdelani odločitveni model s specifikacijo za izvedbo samostojne klasifikacije vsebine, ki je pripravljen za neposredno izvedbo klasifikacije besedil. Tabela s podatki, tj. izvirnimi zapisi v obliki dokumentov, ki so namenjeni preizkušanju učinkovitosti in uspešnosti izdelanih odločitvenih modelov. *Izhodna oblika:* tabela z izvirnimi zapisi v obliki dokumentov, ki so predstavljali testne podatke, ter predhodno identificirana in dokumentirana klasifikacija v obliki določenega tematskega modela za vsak posamezen izvirni zapis. V tabeli so dodatno dokumentirane uteži in verjetnosti glede na posamezen določen tematski model in individualno določena klasifikacija glede na posamezen izvirni zapis, ki so bili vključeni v nabor podatkov za izvedbo preizkušanja izvedbenega odločitvenega modela.

Poročilo o izvedenem merjenju učinkovitosti in uspešnosti pri uporabi izdelanega odločitvenega modela.	V aktivnosti se opravi pregled in primerjava izvedene klasifikacije na podlagi uporabe izdelanega odločitvenega modela glede na predhodno določene klasifikacijske oznake v obliki določenih tematskih modelov, ki so bili rezultat delotoka določanja tematskih modelov. Poročilo vključuje analitični prikaz dodeljene klasifikacije na podlagi izdelanega odločitvenega modela v povezavi s predhodno znanimi in določenimi tematskimi modeli.	*Vhodna oblika:* tabela z izvirnimi zapisi v obliki dokumentov, ki so predstavljali testne podatke, ter predhodno identificirana in dokumentirana klasifikacija v obliki določenega tematskega modela za vsak posamezen izvirni zapis. V tabeli so dodatno dokumentirane uteži in verjetnosti glede na posamezen določen tematski model in individualno določena klasifikacija glede na posamezen izvirni zapis, ki so bili vključeni v nabor podatkov za izvedbo preizkušanja izvedbenega odločitvenega modela. *Izhodna oblika:* statistični in analitični prikaz določenih klasifikacijskih oznak na podlagi uporabe izdelanega odločitvenega modela glede na predhodno določene klasifikacije v obliki določenih tematskih modelov in odstopanja določenih klasifikacij (tematskih modelov) glede na izhodiščno določene tematske modele v okviru vseh zajetih nestrukturiranih zapisov.
Časovno merjenje – zaključek.	Aktivnost je namenjena merjenju časa, ki je potreben za izvedbo delotoka izdelave modela in izvedbe klasifikacije. Aktivnost ne vpliva na vsebinsko ali strukturno obdelavo podatkov in ne spreminja obstoječih podatkov. Aktivnost ustavi in v sekundah izmeri čas, ki je bil potreben za izvedbo vseh faz in aktivnosti delotoka.	

Vir: KNIME 2024.

3.3.8 Izdelava naslova popisne enote in končne oblike popisne enote z izvozom podatkov v izvirno obliko

V tretjem delotoku je bil za potrebe izdelave naslova popisne enote uporabljen predizdelan model za izvedbo krnjenja besedila oziroma krajšanje in povzemanje vsebine individualnih izvirnih zapisov.

Kot navajajo Allahyari idr. (2017) je naloga samodejnega povzemanja besedila ustvariti jedrnat in smiseln povzetek ob ohranjanju vsebine ključnih informacij in zagotavljanju splošnega pomena. V zadnjih letih so bili razviti številni pristopi za samodejno povzemanje besedila. Danes lahko najpogosteje zasledimo uporabo povzemanja vsebine v obliki povzetkov pri uporabi internetnih brskalnikov in iskanju po vsebinah, ki jih imajo indeksirane. Glavni cilj samodejnega povzemanja je tako ustvariti povzetek, ki vključuje glavne ideje ali sporočila izvirne vsebine v čim krajšem obsegu besedila in s čim manjšim ponavljanjem besedila. Namen je, da končni uporabnik takšnih povzetkov pridobi ključne informacije, ne da bi moral prebrati celoten dokument, s tem pa tudi prihranek na času za preverjanje celotne vsebine izvirnega zapisa (El-Kassas idr. 2020).

Allahyari idr. (2017) podajajo dodatno obrazložitev, da je samodejno povzemanje besedila zelo zahtevno, saj ko ljudje povzemajo del besedila, ga običajno preberejo v celoti, da se praviloma razvije razumevanje vsebine in šele nato izvede pisanje povzetka, v katerem se izpostavlja glavne točke oziroma pomembno vsebino.

Metod za izvedbo povzemanja je več, zadnje čase se večinoma uporabljata dve:

- ekstrakcijsko povzemanje (angl. extractive summarization) in
- abstraktno povzemanje (angl. abstract summarization).

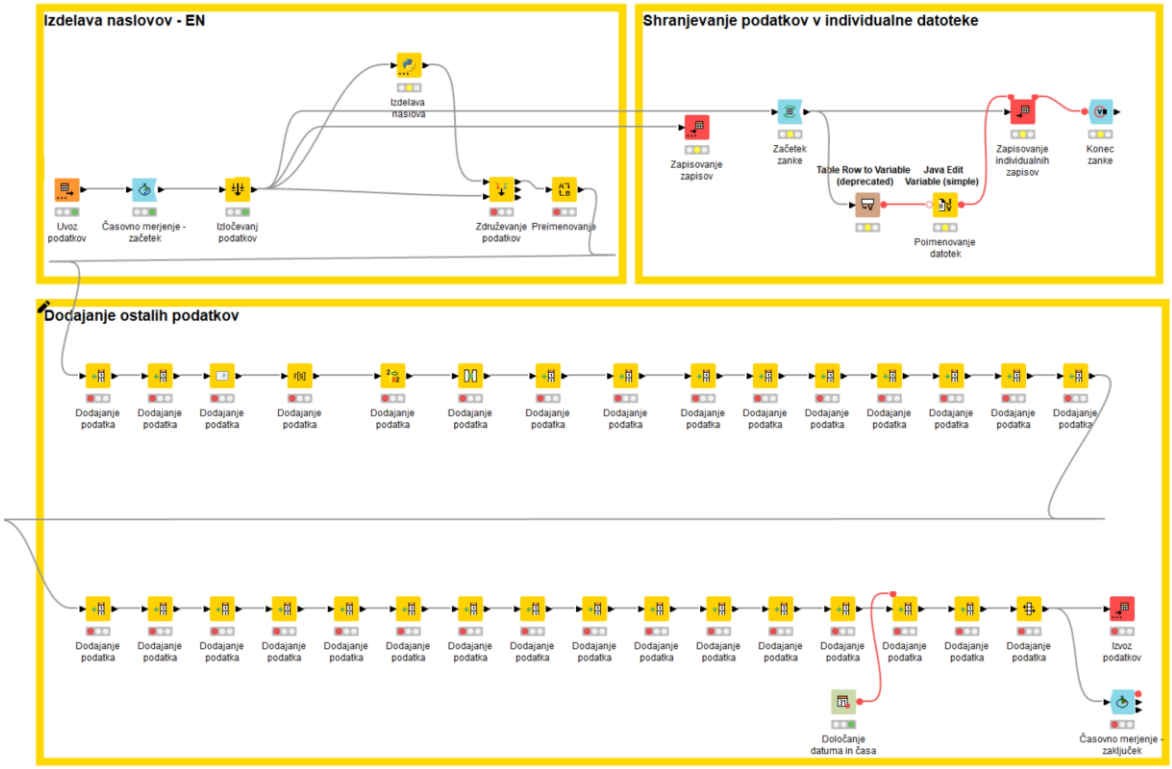
Ekstrakcijsko povzemanje je metoda, kjer povzemanje vključuje izbiranje najbolj ustreznih in relevantnih delov besedila in njihovo sistematično organiziranje. Besedilo, ki sestavlja povzetek, je tako identično prevzeto iz izvornega gradiva. Na splošno je samostojno ekstrakcijsko povzemanje sestavljeno iz treh glavnih faz: predhodne obdelave, ocenjevanja stavkov in izbire stavkov. V fazi predhodne obdelave se običajno izvedejo aktivnosti, kot so tokenizacija (izdelava t. i žetonov) ter segmentacija stavkov in odstavkov. V fazi ocenjevanja stavkov so stavki prednostno razvrščeni glede na določene kriterije, vsak stavek pa se nato individualno ocenjuje. Nazadnje so v fazi izbire stavkov izbrani najboljši oziroma najbolje ocenjeni (relevantni) stavki, ki bodo vključeni v končni povzetek. Poseben vpliv in enega od temeljnih izzivov pri povzemanju večje količine ločenih zapisov predstavlja redundanca, saj je bolj verjetno, da bodo podobni stavki večkrat prikazani v različnih zapisih, zato je namen, da so izbrani najpomembnejši stavki z najmanjšo redundanco (Bidoki idr. 2020).

V nasprotju z ekstrakcijskim povzemanjem predstavlja abstraktno povzemanje v primeru izdelave naslova ustreznejši pristop. Zmožnost ustvarjanja unikatnih povzetkov, ki združujejo pomembne informacije iz izvirnih zapisov, je ena od osnov za uspešno

ustvarjanje naslova popisne enote, predvsem v smeri povzemanja do te mere, da povzetek predstavlja čim krajšo obliko besedila, ki najbolje predstavlja vsebino. Abstraktni povzetek vsebino predstavi v logični, dobro organizirani in slovnično neoporečni obliki. Kakovost povzetka je mogoče bistveno izboljšati, če je vsebina enostavno berljiva oziroma je zagotovljena višja jezikovna kakovost (Shafiq idr. 2022). Suleiman in Awajan (2020) dodajata, da abstraktno povzemanje besedila zahteva razumevanje dokumenta za ustvarjanje povzetka, vzporedno s tem pa so potrebne napredne tehnike strojnega učenja in obsežna obdelava naravnega jezika. Zato je izvedba abstraktnega povzemanje težja kot izvedba ekstrakcijskega povzemanja, saj abstraktno povzemanje zahteva realno poznavanje vsebine in izvedbo ustrezne semantične analize. Z vidika pridobljenih rezultatov pa je abstraktno povzemanje lahko tudi bolj razumljivo oziroma učinkovito od ekstrakcijskega, saj je povzetek približna predstavitev povzetka, kot bi ga ustvaril človek, zaradi česar se šteje, da je glede na analizirano vsebino bolj smiseln.

Za izdelavo naslova popisne enote je bil uporabljen pristop z uporabo odprtokodne platforme za strojno učenje Tensorflow (2023) ter t.i. arhitekture oziroma modela BART (2019) z uporabo predizdelanega modela distilbart-xsum-6-6 (2023).

Slika 30: Izdelava naslova popisne enote z uporabo predizdelanega modela



Vir: KNIME 2024.

V tabeli 14 so dokumentirane aktivnosti za izdelavo naslova popisne enote, kot so predstavljene na sliki 30.

Tabela 14: Izdelava naslova popisne enote z uporabo predizdelanega modela

Faza	Opis	Vhodni/izhodni podatki
Zajem zapisov z nestrukturirano vsebino, določenimi tematskimi modeli in prepoznanimi imenskimi entitetami.	Zajem zapisov z nestrukturirano vsebino predstavlja uvoz posameznih zapisov z določenimi tematskimi modeli oziroma prepoznanimi imenskimi entitetami. Izbrani format datotek za uvoz podatkov je CSV (2023), ki predstavlja izvozno vrednost delotokov za izdelavo tematskih modelov in delotoka za prepoznavo imenskih entitet.	*Vhodna oblika:* individualni zapisi, agregirani v CSV-formatu (CSV 2023). *Izhodna oblika:* tabela z besedilom zapisa, določenimi tematskimi modeli in prepoznanimi imenskimi entitetami ter ostalimi podatki iz prejšnjih delotokov.
Časovno merjenje – začetek.	Aktivnost je namenjena merjenju časa, ki je potreben za izvedbo delotoka prepoznavanja imenskih entitet. Aktivnost ne vpliva na vsebinsko ali strukturno obdelavo podatkov in ne spreminja obstoječih podatkov.	
Izločevanje podatkov.	V okviru aktivnosti je bilo izvedeno izločevanje nepotrebnih podatkov iz prejšnjih delotokov.	*Vhodna oblika:* tabela z besedilom zapisa, določenimi tematskimi modeli, seznamom ključnih besed, prepoznanimi imenskimi entitetami in ostalimi podatki iz prejšnjih delotokov. *Izhodna oblika:* tabela z besedilom izvirnega zapisa.
Uporaba modela za izdelavo naslova popisne enote.	V tej fazi se izvede prepoznava besedila za namen povzemanja vsebine v obliko, primerno za izdelavo naslova popisne enote na podlagi individualnih izvirnih zapisov in predizdelanega modela.	*Vhodna oblika:* predizdelan model za izvedbo povzemanja besedila s specifikacijo za določanje obsega in oblike izvedbe povzemanja ter tabela z izvirnim besedilom zapisov.

	Uporabljen predizdelan model: - za angleški jezik: BART – distilbart-xsum-6-6 (2023), - za slovenski jezik: SloSummarizer – t5-headline (Žagar in Robnik Šikonja 2021).	*Izhodna oblika:* tabela z izdelanimi naslovi popisnih enot oziroma povzetki glede na individualen izvirni zapis.
Združevanje podatkov.	V aktivnosti delotoka se izvede združevanje podatkov izdelanih naslovov popisnih enot glede na individualen izvirni zapis.	*Vhodna oblika:* tabela, ki vključuje izvirno besedilo zapisov, in tabela, kjer se nahajajo izdelani naslovi popisnih enot. *Izhodna oblika:* tabela, kjer so združeni izvirni zapisi in izdelani naslovi popisnih enot.
Shranjevanje podatkov v individualne datoteke.	Namen aktivnosti je vrnitev izvirnih zapisov iz tabelarne oblike v njihovo prvotno tekstno obliko.	*Vhodna oblika:* tabela, ki vključuje izvirno besedilo zapisov. *Izhodna oblika:* individualne tekstovne datoteke za vsak individualen izvirni zapis.
Dodajanje ostalih podatkov.	V tej fazi se poleg že identificiranih podatkov o klasifikaciji, imenskih entitetah in izdelanem naslovu popisne enote izvede še dodajanje fiksnih podatkov za zapolnitev sheme po standardu ISAD(G) (2000).	*Vhodna oblika:* tabela, ki vključuje izvirno besedilo zapisov prepoznane imenske entitete in izdelane naslove popisnih enot. *Izhodna oblika:* tabela kjer so združeni izvirni zapisi, prepoznane imenske entitete, izdelani naslovi popisnih enot in ostali podatki za dopolnitev popisne sheme po standardu ISAD(G) (2000).
Časovno merjenje – zaključek.	Aktivnost je namenjena merjenju časa, ki je potreben za izvedbo delotoka	

	prepoznavanja imenskih entitet. Aktivnost ne vpliva na vsebinsko ali strukturno obdelavo podatkov in ne spreminja obstoječih podatkov. Aktivnost ustavi in v sekundah izmeri čas, ki je bil potreben za izvedbo vseh faz in aktivnosti delotoka.	

Vir: KNIME 2024.

3.4 Rezultati

Ker je namen izvedbe delotokov obdelava vsebine in besedil v nestrukturirani obliki, so besedila in vsebina v enaki obliki, kot so bila prevzeta s portala The Guardian Open Platform (2022) oziroma s portala CLARIN (CLARIN.SI 2022), in predstavljajo izvirno obliko zapisa za izvedbo vseh delotokov. Zaradi primerljivosti obdelane vsebine se je skozi vse faze delotokov ohranila tudi izvirna oblika zapisov in uporabila tudi kot del končnega rezultata izhodnih vrednosti procesa oziroma delotoka, vključno z novo obdelanimi vsebinami in podatki.

V določenih delotokih se je izvedla tudi faza urejanja oziroma »čiščenja podatkov« z namenom preverjanja, ali kakovost podatkov lahko vpliva na končni rezultat.

3.4.1 Delotok za prepoznavanje imenskih entitet v angleškem jeziku

Prva faza izdelave popisne enote je vključevala dva pristopa: pristop z učenjem in izdelavo modela za prepoznavanje imenskih entitet na podlagi slovarjev ter pristop s prepoznavo imenskih entitet na podlagi predizdelanih modelov.

Delotok za izdelavo modela za prepoznavanje imenskih entitet na podlagi slovarjev

Delotok za izdelavo imenskih entitet je bil vzpostavljen kot prvi delotok za izvedbo celotnega procesa urejanja besedil in izdelave popisne enote. Vhodno besedilo je bilo predhodno obdelano na način, da je vsak posamezen vnos (vrstica) v tabeli predstavljal individualen članek oziroma individualno enoto zapisa.

Na sliki 31 je prikazana struktura podatkov za začetek izvedbe delotoka.

Slika 31: Primer zajetega besedila v angleškem jeziku za namen izvedbe prepoznave imenskih entitet

Row0	A day after Barack Obama announced tough new sanctions over what intelligence agencies believe to be Russian attempts to
Row1	A journalist was temporarily banned from Facebook after a post in which he called Trump supporters "a nasty fascistic lot", in
Row2	Technology? Bah humbug: "I think we ought to get on with our lives," said Donald Trump on Wednesday, summing up his take
Row3	While there were some good things in technology released this year, there were also quite a few let downs. From detonating
Row4	The US Department of Homeland Security (DHS) and FBI have released an analysis of the allegedly Russian government-spon
Row5	Samuel Gibbs' piece about the internet of things (How I turned my home into a sci-fi dream, theguardian.com, 24 December) o
Row6	Smart electricity meters, of which there are more than 100m installed around the world, are frequently "dangerously insecure
Row7	Five years ago, the demise of the music industry seemed almost inevitable. Recession, rampant piracy, falling CD sales and a
Row8	My mum is looking to buy a new laptop after Christmas. Her budget is tight: around £200. She uses it for Microsoft Office, bro
Row9	In 2016, 21st-century virtual reality really arrived. From cheap mobile experiences to exuberant desktop machines, if you war
Row10	The signatories to the letter on children's lifestyles (Screen-based lifestyle harms children's health, 26 December) make the usu
Row11	The worldwide system used to coordinate travel bookings between airlines, travel agents, and price comparison websites is ho
Row12	Amazon has refused to hand over data from an Echo smart speaker to US police, who want to access it as part of an investiga
Row13	Over the course of 2016, artificial intelligence made the leap from "science fiction concept" to "almost meaningless buzzword"
Row14	A Facebook safety check for Bangkok, which the company claimed was prompted by a one-man protest near the prime ministe

Vir: KNIME 2024; The Guardian Open Platform 2022.

Delotok za prepoznavanje imenskih entitet z uporabo modela, ki se je učil na podlagi predpripravljenih slovarjev, je vključeval tri različne tipe imenskih entitet (glej sliko 32):

- organizacije,
- lokacije in
- osebe.

Slika 32: Primer zajetega besedila slovarja v angleškem jeziku za namen izdelave modela za prepoznavo imenskih entitet

Row ID	S Name
Row18	Ala
Row19	Al-Aqsa
Row20	Alaska
Row21	Albania
Row22	Alberta
Row23	Aleksandrovac
Row24	Alexandroupolis
Row25	ALEXANDROUPOLIS
Row26	Algeria
Row27	Algiers
Row28	Alicante
Row29	Alkhan-Yurt
Row30	ALKHAN-YURT
Row31	Allenby Bridge
Row32	ALLENTOWN
Row33	Almere
Row34	ALMERE
Row35	AL-MUNTAR
Row36	Alomella Robles
Row37	al-Risala
Row38	Alta
Row39	Altamira
Row40	Amazon
Row41	America
Row42	Amman
Row43	AMMAN
Row44	Amsterdam
Row45	AMSTERDAM

Vir: KNIME 2024; Finkel idr. 2005.

Pri izdelavi modela za prepoznavo imenskih entitet je bilo treba ločiti slovarje glede na posamezni tip imenskih entitet ter izdelati model za prepoznavo imenskih entitet glede na

posamezni tip. V okviru delotoka je bilo nato izvedeno združevanje pridobljenih rezultatov posameznega odločitvenega modela glede na posamezni tip entitete. Ob tem je bila zagotovljena konsistentnost podatkov glede na uporabljene izvirne zapise.

Na sliki 33 je prikazan rezultat prepoznave imenskih entitet v primeru uporabe modela, ki je bil izdelan na podlagi predpripravljenih slovarjev, kjer so združeni rezultati prepoznave imenskih entitet, ki so bili predmet obdelave.

Slika 33: Rezultat prepoznave imenskih entitet v angleškem jeziku v primeru uporabe modela, ki je bil izdelan na podlagi slovarjev

#ThankYourMP trended on Twitter on Friday, following the killing of the Labour MP Jo Cox. Thousands of voters used the platfor…	[labour,express]	[or]	[little]
(World rankings as of Mon 25 June; ups and downs calculated from Mon 28 May) 1 New Zealand ◄► Lethal in the second halves …	[sydney]	[new,zealand,france,ireland,sydney,series,austr…	[ross]
… we must proceed according to a phased approach giving priority to an orderly withdrawal. In these negotiations the union will …	[union,eu,house]	[britain,british]	?
1 A United Kingdom (12A) (Amma Asante, 2016, US/UK/Cze) 111 mins. The telling is conventional but the true story told here is a…	[united,london,resistance,new,y…	[united,kingdom,botswana,london,britain,new,yo…	[david,jim]
1 Abstract Expressionism Prepare to be awed and exhilarated by some of the greatest art of the modern age. In the 1940s, a ge…	[new,york,disney,london]	[new,york,european,mexico,london]	[jackson,pollock,mark]
1 Adding Machine: A Musical US playwright Elmer Rice looked at the effects of big business and mechanisation on the lives of ordi…	[liverpool,house,birmingham]	[us,people,black,birmingham,rep,uk]	[black,hill,david]
1 Afrobeats music festival Between Drake's funky flirtations, dancehall crew Mixpak triumphing at Culture Clash, and acclaim for …	[house,birmingham]	[uk,europe,nigerian,singer,england,euro,manche…	?
1 Alvin Ailey American Dance Theater In the third programme for this company of fabulous dancers, Robert Battle's Awakening is …	[american,ipswich,national]	[wales]	[barton]
1 American Honey (15) (Andrea Arnold, 2016, UK/US) 164 mins British auteur Arnold hits the open road for a dreamy, dishevelled…	[american,florida]	[british,open,us,midwest,florida,uk]	[arnold,cruise,jim]
1 Anselm Kiefer The art of Anselm Kiefer is thick with history – and thick with paint, which he heaps up on his vast canvases along…	?	[german,nazi,era]	[hall]

Vir: KNIME 2024; Finkel idr. 2005.

Iz prikazanega rezultata je razvidno, da je imel model precejšnje težave s prepoznavanjem celotnega obsega imenskih entitet, kar se nanaša predvsem na obseg in kakovost slovarjev, ki so bili uporabljeni v primeru izdelave modela. Ker se je model učil iz razpoložljivih podatkov, je bila tudi učinkovitost prepoznave imenskih entitet v precejšnji meri odvisna od obsega besed, ki so bile podlaga za izdelavo modela. Ker je bila izdelava modelov ločena na prepoznavo različnih tipov entitet, se tudi rezultati prepoznave oziroma učinkovitost lahko razlikuje glede na uporabljeni model in vrsto tipa imenskih entitet, ki je bila predmet prepoznave.

Na sliki 34 je na enem izvirnem zapisu prikazana primerjava prepoznave treh tipov imenskih entitet na podlagi osebnega (fizičnega) prepoznavanja in prepoznavanja, izvedenega na osnovi izdelanega modela. Iz slike ja razvidno, da je izdelani model glede na razpoložljive slovarje dobro prepoznal le tip imenske entitete »lokacija«, medtem ko je pri ostalih dveh tipih entitet rezultat precej slabši. Razlog za to je pomanjkljiva vsebina slovarjev oziroma premajhen vsebinski obseg slovarjev, ki so bili uporabljeni za izdelavo modela.

Slika 34: Primerjava prepoznave treh tipov imenskih entitet v angleškem jeziku na podlagi osebnega (fizičnega) prepoznavanja in prepoznavanja, izvedenega na osnovi izdelanega modela

Personal anotation		NER model	
New Zealand	LOCATION	New Zealand	LOCATION
France	LOCATION	France	LOCATION
Julian Savea	LOCATION		
Toulon	PERSON		
Scott Barrett	PERSON		
Damian McKenzie	PERSON		
Gerard Meagher	PERSON		
Ireland	LOCATION	Ireland	LOCATION
Wallabies	ORGANIZATION		
Sydney	LOCATION	Sydney	LOCATION
Joe Schmidt	PERSON		
Ross Byrne	PERSON		
Australia	LOCATION	Australia	LOCATION
Conor Murray	PERSON	Conor Murray	PERSON
Johnny Sexton	PERSON	Johnny Sexton	PERSON
Twickenham	LOCATION	Twickenham	LOCATION
Wales	LOCATION	Wales	LOCATION
England	LOCATION	England	LOCATION
Warren Gatland	PERSON		
South Africa	LOCATION	South Africa	LOCATION
Argentina	LOCATION	Argentina	LOCATION
Paul Rees	PERSON		
Jonny May	PERSON		
Danny Cipriani	PERSON		
Cape Town	LOCATION	Cape Town	LOCATION
Japan	LOCATION	Japan	LOCATION
Brisbane	LOCATION	Brisbane	LOCATION
Kurtley Beale	LOCATION		
Melbourne	LOCATION	Melbourne	LOCATION
Will Genia	PERSON		
Michael Cheika	PERSON		
David Pocock	PERSON		
Faf de Klerk	PERSON	de Klerk	PERSON
Duane Vermeulen	PERSON		
Rassie Erasmus	PERSON		
Springboks	LOCATION	Springboks	LOCATION
Siya Kolisi	PERSON		
Robert Kitson	PERSON		
Scotland	LOCATION	Scotland	LOCATION
United States	LOCATION	United States	LOCATION
Fiji	LOCATION		
Anthony Belleau	PERSON		
Wesley Fofana	PERSON		
Lautoka	LOCATION		
Tonga	LOCATION	Tonga	LOCATION
Georgia	LOCATION	Georgia	LOCATION
Daniel Hourcade	PERSON		
Italy	LOCATION	Italy	LOCATION
Russia	LOCATION	Russia	LOCATION
Canada	LOCATION	Canada	LOCATION
Adam Byrne	PERSON	Adam Byrne	PERSON

Vir: KNIME 2024.

Celoten delotok je za predhodno obdelavo podatkov, izdelavo modela za prepoznavanje imenskih entitet ter izvedbo prepoznavanja potreboval slabih 16 ur časa brez uporabe specializiranih jeder grafične kartice (glej sliko 35).

Slika 35: Čas, potreben za predhodno obdelavo podatkov, izdelavo modela za prepoznavanje imenskih entitet ter izvedbo prepoznavanja v angleškem jeziku

Mean Execution Time (s)	57007.342
Worst Execution Time (s)	57007.342
Best Execution Time (s)	57007.342
Total Execution Time (s)	57007.342
Last Execution time (s)	57007.342

Vir: KNIME 2024.

Delotok za prepoznavanje imenskih entitet na podlagi predizdelanih modelov

Delotok za izdelavo imenskih entitet na podlagi predizdelanih modelov je bil vzpostavljen kot drugi delotok za izvedbo celotnega procesa urejanja besedil in izdelave popisne enote. Vhodno besedilo z vidika zagotavljanja višje kakovosti podatkov predhodno ni bilo obdelano, ravno tako je vsak posamezen vnos (vrstica) v tabeli predstavljal individualen članek oziroma individualno enoto zapisa. Edina razlika v okviru predobdelave podatkov je bila pri metodi Stanford Named Entity Recognizer (Finkel idr. 2005), kjer je bilo dodatno izvedeno izdelovanje lem.

Zaradi primerjanja učinkovitosti prepoznavanja imenskih entitet so bili preverjeni trije različni predizdelani modeli in povezane arhitekture:

- Stanford Named Entity Recognizer (Finkel idr. 2005),
- BERT Named Entity Recognizer – bert-large-NER (Devlin idr. 2018; BERT-large-NER 2023),
- SPACY EntityRecognizer – en-core-web-lg (En-core-web-lg 2023).

Delotok za prepoznavanje imenskih entitet z uporabo predizdelanih modelov je vključeval štiri različne tipe imenskih entitet:

- organizacije,
- lokacije,
- osebe in
- datum.

Na sliki 36 je prikazan rezultat prepoznave imenskih entitet z uporabo Stanford Named Entity Recognizer (Finkel idr. 2005), kjer so poleg identificiranih imenskih entitet ob vsaki imenski entiteti razvidni tudi tipi entitet.

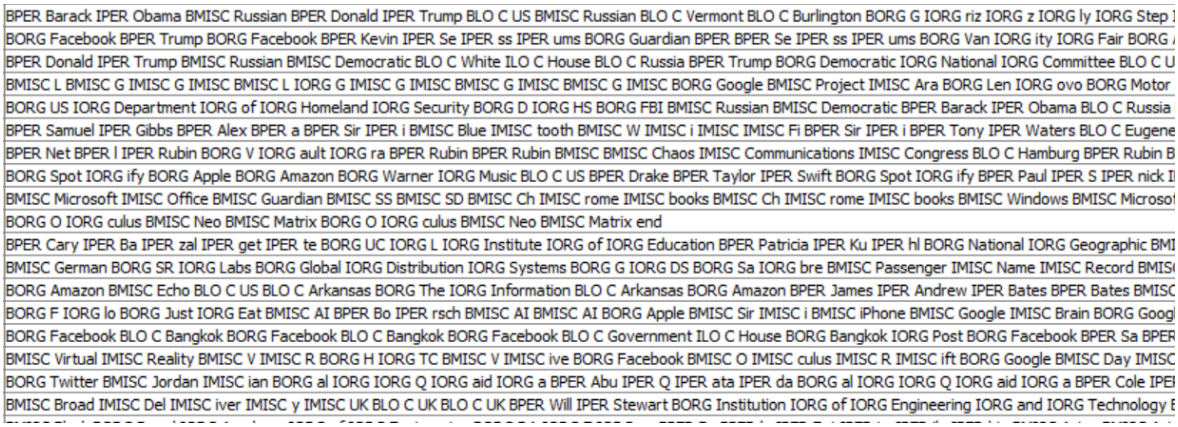
Slika 36: Rezultat delotoka prepoznave imenskih entitet z uporabo modela Stanford Named Entity Recognizer v angleškem jeziku

Barack Obama[ORGANIZATION(NE)];Donald Trump[PERSON(NE)];US[LOCATION(NE)];Vermont[LOCATION(NE)];Burlington[LOCATION(NE)];Friday[DATE(
Facebook[ORGANIZATION(NE)];Kevin Sessums[PERSON(NE)];Vanity Fair[ORGANIZATION(NE)];ABC[ORGANIZATION(NE)];Matthew Dowd[ORGANIZATIO
Donald Trump[PERSON(NE)];Wednesday[DATE(NE)];2016[DATE(NE)];Russia[LOCATION(NE)];Trump[PERSON(NE)];Christmas[DATE(NE)];morning[TIME(N
this year[DATE(NE)];2016[DATE(NE)];Google[ORGANIZATION(NE)];Motorola[LOCATION(NE)];MacBook[ORGANIZATION(NE)];2015[DATE(NE)];USB[ORG/
US Department of Homeland Security[ORGANIZATION(NE)];FBI[ORGANIZATION(NE)];2016[DATE(NE)];Thursday[DATE(NE)];Barack Obama[ORGANIZATI
December[DATE(NE)];Alexa[LOCATION(NE)];Tony Waters Eugene[ORGANIZATION(NE)];Oregon[LOCATION(NE)];USA[ORGANIZATION(NE)];2016[DATE(
Netanel Rubin[PERSON(NE)];Vaultra[ORGANIZATION(NE)];"Rubin[PERSON(NE)];Rubin[PERSON(NE)];33rd Chaos Communications Congress[ORGANIZAT
2016[DATE(NE)];Amazon[LOCATION(NE)];December[DATE(NE)];Warner Music[ORGANIZATION(NE)];$1bn[MONEY(NE)];$100m[MONEY(NE)];first half of 2
Christmas[DATE(NE)];Microsoft Office[ORGANIZATION(NE)];SSD[ORGANIZATION(NE)];USB[ORGANIZATION(NE)];DVD[ORGANIZATION(NE)];Chromeboo
2016[DATE(NE)];Kansas[LOCATION(NE)]
December[DATE(NE)];UCL Institute of Education[ORGANIZATION(NE)];Patricia Kuhl[PERSON(NE)];National Geographic[ORGANIZATION(NE)];January 201
SR Labs[ORGANIZATION(NE)];1960s[DATE(NE)];Sabre[LOCATION(NE)];PNR[ORGANIZATION(NE)];Karsten Nohl[PERSON(NE)];Nemanja Nikodijevic[PERS
US[LOCATION(NE)];Arkansas[LOCATION(NE)];Amazon[LOCATION(NE)];James Andrew Bates[PERSON(NE)];November 2015[DATE(NE)];2014[DATE(NE)];
2016[DATE(NE)];Borsch[LOCATION(NE)];2017[DATE(NE)];2011[DATE(NE)];Google Brain[LOCATION(NE)];Waymo[LOCATION(NE)];Google[LOCATION(NE
Bangkok[LOCATION(NE)];December[DATE(NE)];Facebook[ORGANIZATION(NE)];Government House[ORGANIZATION(NE)];Bangkok Post[ORGANIZATION(
Oculus Rift[PERSON(NE)];Daydream[LOCATION(NE)];VR[ORGANIZATION(NE)];AAA[ORGANIZATION(NE)];today[DATE(NE)];Eve Valkyrie[ORGANIZATIO
Abu Qatada[PERSON(NE)];Cole Bunzel[PERSON(NE)];Princeton University[ORGANIZATION(NE)];Syria[LOCATION(NE)];Islamic State[ORGANIZATION(NE)

Vir: KNIME 2024.

Na sliki 37 je prikazan rezultat prepoznave imenskih entitet z uporabo modela in arhitekture BERT (Devlin idr. 2018; BERT-large-NER (2023), kjer so poleg identificiranih imenskih entitet ob vsaki imenski entiteti razvidni tudi tipi entitet.

Slika 37: Rezultat delotoka prepoznave imenskih entitet z uporabo modela in arhitekture BERT v angleškem jeziku



Vir: KNIME 2024.

Na sliki 38 je prikazan rezultat prepoznave imenskih entitet z uporabo modela in knjižnice SPACY (2023), kjer so poleg identificiranih imenskih entitet ob vsaki imenski entiteti razvidni tudi tipi entitet.

Slika 38: Rezultat delotoka prepoznave imenskih entitet z uporabo SPACY v angleškem jeziku

Vir: KNIME 2024.

Za preverjanje učinkovitosti je bila na enem izvirnem zapisu izvedena primerjava rezultatov prepoznave imenskih entitet z uporabo vseh treh modelov. Rezultati vseh treh modelov so bili dodatno primerjani s predhodnim ročnim pregledom in indeksacijo izvirnega zapisa. Za preverjanje pravilnosti prepoznanih imenskih entitet je bila pri ročnem pregledu in indeksaciji dodana opredelitev tipa prepoznanih imenskih entitet.

Na sliki 39 je prikazana primerjava rezultatov prepoznave vseh treh modelov, vključno z ročno indeksiranimi vrednostmi. Iz rezultatov je razvidno, da sta na podlagi izbranih predizdelanih modelov boljše rezultate zagotovila pristopa z uporabo modela in arhitekture Stanford NER (Finkel idr. 2005) in SPACY NER (SPACY 2023) oziroma sta bila vsaj glede na preverjeni izvirni zapis enako učinkovita. Model in arhitektura BART (Devlin idr. 2018; BERT-large-NER 2023) je ravno tako zagotovil dobre rezultate z možnostjo dodatnega podrobnejšega učenja modelov (angl. Fine tuining). Ob uporabi dodatnih virov bi mogoče lahko zagotovili še boljše rezultate, kar pa je predvsem odvisno od kakovosti in obsega vira za dodatno učenje. Pri rezultatu prepoznanih imenskih entitet z uporabo modela BART (Devlin idr. 2018; BERT-large-NER 2023) izstopa predvsem slabša identifikacija tipa imenskih entitet »oseba«. Pri vseh treh modelih ni bilo treba izvesti dodatnega učenja (angl. Fine tuning), saj so že obstoječi modeli zagotovili dobre rezultate pri prepoznavanju vseh predvidenih tipov imenskih entitet.

Slika 39: Primerjava rezultatov prepoznave vseh treh modelov, vključno z ročno indeksiranimi vrednostmi v angleškem jeziku

Personal anotation		Stanford		BERT		SPACY	
Mon 25 June	DATE	Mon 25 June	DATE			Mon 25 June	DATE
Mon 28 May	DATE	Mon 28 May	DATE			Mon 28 May	DATE
New Zealand	LOCATION	New Zealand	LOCATION	New Zealand	LOCATION	New Zealand	LOCATION
France	LOCATION	France	LOCATION	France	LOCATION	France	LOCATION
Julian Savea	LOCATION		LOCATION	Julian Savea	PERSON	Julian Savea	LOCATION
Toulon	PERSON	Toulon	PERSON	Toulon	LOCATION		
Scott Barrett	PERSON	Scott Barrett	PERSON	Scott Barrett	PERSON	Scott Barrett	PERSON
Damian McKenzie	PERSON	Damian McKenzie	PERSON	Damian McKenzie	PERSON	Damian McKenzie	PERSON
Gerard Meagher	PERSON			Gerard Meagher	PERSON	Gerard Meagher	PERSON
Ireland	LOCATION	Ireland	LOCATION	Ireland	LOCATION	Ireland	LOCATION
Wallabies	ORGANIZATION			Wallabies	ORGANIZATION		
Sydney	LOCATION	Sydney	LOCATION	Sydney	LOCATION	Sydney	LOCATION
Joe Schmidt	PERSON	Joe Schmidt	PERSON	Joe Schmidt	PERSON	Joe Schmidt	PERSON
Ross Byrne	PERSON	Ross Byrne	PERSON	Ross Byrne	PERSON	Ross Byrne	PERSON
Australia	LOCATION	Australia	LOCATION	Australia	LOCATION	Australia	LOCATION
Conor Murray	PERSON	Conor Murray	PERSON	Conor Murray	PERSON	Conor Murray	PERSON
Johnny Sexton	PERSON	Johnny Sexton	PERSON	Johnny Sexton	PERSON	Johnny Sexton	PERSON
Twickenham	LOCATION	Twickenham	LOCATION	Twickenham	LOCATION	Twickenham	LOCATION
Wales	LOCATION			Wales	LOCATION	Wales	
England	LOCATION	England	LOCATION	England	LOCATION	England	LOCATION
Warren Gatland	PERSON	Warren Gatland	PERSON	Warren Gatland	PERSON	Warren Gatland	PERSON
South Africa	LOCATION	South Africa	LOCATION	South Africa	LOCATION	South Africa	LOCATION
Argentina	LOCATION	Argentina	LOCATION	Argentina	LOCATION	Argentina	LOCATION
Paul Rees	PERSON	Paul Rees	PERSON	Paul Rees	PERSON	Paul Rees	PERSON
Jonny May	DATE	Jonny May	DATE			Jonny May	DATE
Danny Cipriani	PERSON	Danny Cipriani	PERSON			Danny Cipriani	PERSON
Cape Town	LOCATION	Cape Town	LOCATION	Cape Town	LOCATION		
Japan	LOCATION	Japan	LOCATION	Japan	LOCATION	Japan	LOCATION
November	DATE	November	DATE			November	DATE
Brisbane	LOCATION	Brisbane	LOCATION	Brisbane	LOCATION	Brisbane	LOCATION
Kurtley Beale	LOCATION	Kurtley	LOCATION			Kurtley Beale	LOCATION
Melbourne	LOCATION	Melbourne	LOCATION	Melbourne	LOCATION	Melbourne	LOCATION
Will Genia	PERSON	Genia	PERSON			Will Genia	PERSON
Michael Cheika	PERSON	Michael Cheika	PERSON			Michael Cheika	PERSON
David Pocock	PERSON	David Pocock	PERSON	David Pocock	PERSON	David Pocock	PERSON
Faf de Klerk	PERSON	Klerk	PERSON			Faf de Klerk	PERSON
Duane Vermeulen	PERSON	Duane Vermeulen	PERSON			Duane Vermeulen	PERSON
Rassie Erasmus	PERSON	Rassie Erasmus	PERSON			Rassie Erasmus	PERSON
Springboks	LOCATION	Springboks	LOCATION				
Siya Kolisi	PERSON	Siya Kolisi	PERSON			Siya Kolisi	LOCATION
Robert Kitson	PERSON	Robert Kitson	PERSON	Robert Kitson	PERSON	Robert Kitson	PERSON
Scotland	LOCATION	Scotland	LOCATION	Scotland	LOCATION	Scotland	LOCATION
United States	LOCATION	United States	LOCATION	United States	LOCATION	United States	LOCATION
Fiji	LOCATION	Fiji	LOCATION	Fiji	LOCATION	Fiji	LOCATION
Anthony Belleau	PERSON	Anthony Belleau	PERSON			Anthony Belleau	PERSON
Wesley Fofana	PERSON	Wesley Fofana	ORGANIZATION			Wesley Fofana	ORGANIZATION
Lautoka	LOCATION	Lautoka	LOCATION	Lautoka	LOCATION	Lautoka	LOCATION
Tonga	LOCATION	Tonga	LOCATION	Tonga	LOCATION	Tonga	LOCATION
Georgia	LOCATION	Georgia	LOCATION	Georgia	LOCATION	Georgia	LOCATION
Daniel Hourcade	PERSON	Daniel Hourcade	PERSON	Daniel Hourcade	PERSON	Daniel Hourcade	PERSON
Italy	LOCATION	Italy	LOCATION	Italy	LOCATION	Italy	LOCATION
Russia	LOCATION	Russia	LOCATION	Russia	LOCATION	Russia	LOCATION
Canada	LOCATION	Canada	LOCATION	Canada	LOCATION	Canada	LOCATION
June 2018	DATE	June 2018	DATE			26 June 2018	DATE
Adam Byrne	PERSON	Adam Byrne	PERSON	Adam Byrne	PERSON	Adam Byrne	PERSON

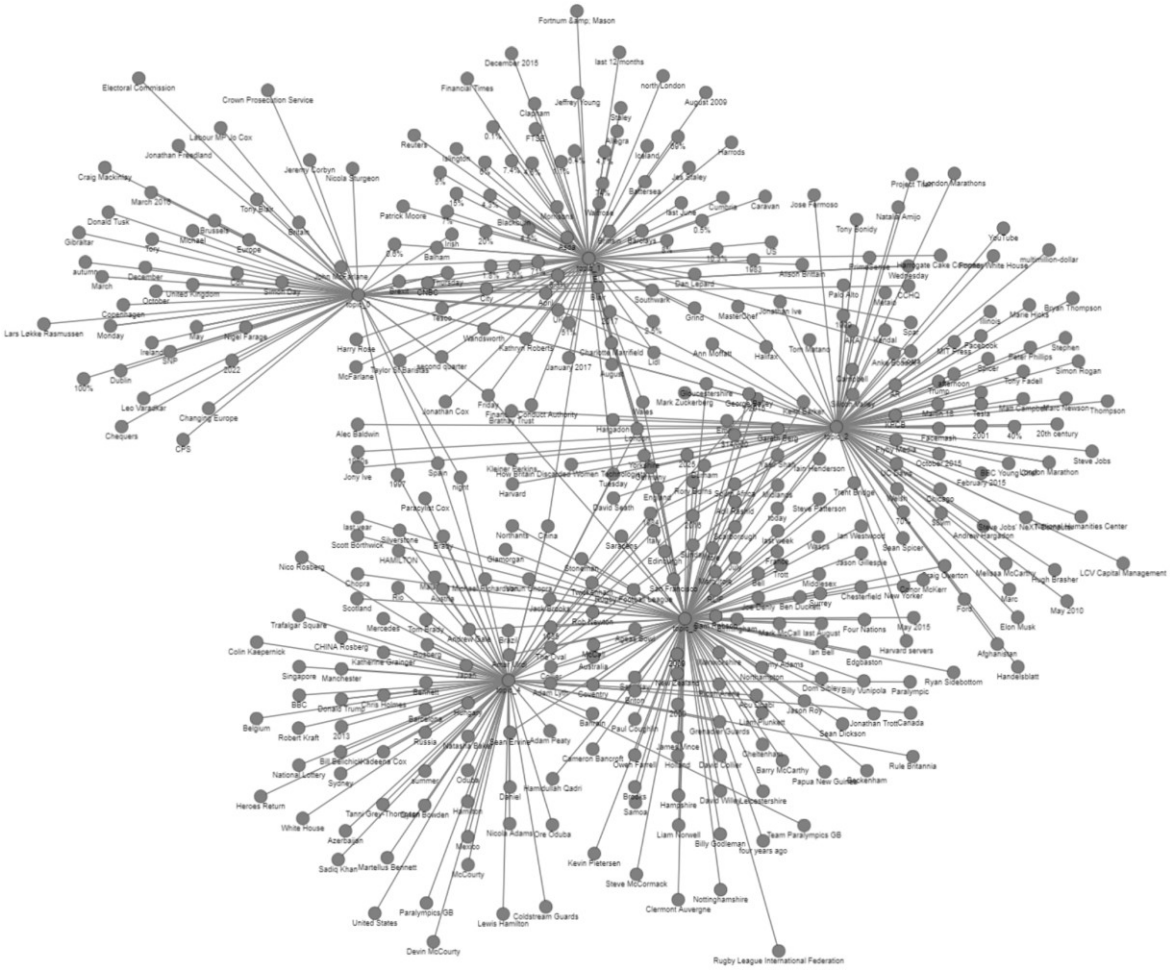
Vir: Lastni 2024.

Za nadaljnjo obdelavo podatkov so bili uporabljeni rezultati modela in pristopa Stanford Named Entity Recognizer (Finkel idr. 2005). Glede na pridobljene rezultate bi lahko bili uporabljeni tudi rezultati, pridobljeni z modelom SPACY NER (SPACY 2023).

Na sliki 40 je prikazan primer koncentracije individualno prepoznanih imenskih entitet glede na pogostost pojavljanja v gruči izvirnih zapisov. Prepoznane imenske entitete, ki se

pojavljajo bliže izvorni klasifikacijski oznaki oziroma določenemu tematskemu modelu (glej tabelo 8), so pogostejše pri identifikaciji entitet glede na gručo posameznih izvirnih zapisov

Slika 40: Primer koncentracije individualno prepoznanih imenskih entitet glede na pogostost pojavljanja v izvirnih zapisih



Vir: KNIME 2024.

Celoten delotok je za prepoznavanje imenskih entitet v postopku uporabe modela in pristopa Stanford Named Entity Recognizer (Finkel idr. 2005) potreboval 55 minut, brez uporabe specializiranih jeder grafične kartice (glej sliko 41).

Slika 41: Čas v sekundah, potreben za izvedbo delotoka prepoznavanja imenskih entitet uporabe modela in pristopa Stanford Named Entity Recognizer v angleškem jeziku

Name	Value
Mean Execution Time (s)	3293.056
Worst Execution Time (s)	3293.056
Best Execution Time (s)	3293.056
Total Execution Time (s)	3293.056
Last Execution time (s)	3293.056

Vir: KNIME 2024.

3.4.2 Delotok za prepoznavanje nestrukturirane vsebine in klasifikacijo vsebine v angleškem jeziku

Druga faza izdelave popisne enote je vključevala izdelavo tematskih modelov v okviru prepoznavanja nestrukturirane vsebine ter izdelavo modela za samostojno izvedbo klasifikacije vsebine na podlagi prepoznanih tematskih modelov. Krovni delotok je bil tako razdeljen na dva dela zaradi večje preglednosti nad izvajanjem procesa in nadzora nad doseganjem ciljev v vsakem od delotokov.

Namen izdelave prvega delotoka je bil v prvi vrsti prepoznava nestrukturirane vsebine na način, da je bilo možno določiti skupne lastnosti individualnim zapisom, ki bi predstavljali osnovo za izvedbo drugega delotoka, tj. izdelavo modela za klasifikacijo vsebine in nato na podlagi izdelanega modela tudi njeno izvedbo.

Izdelava tematskih modelov

Za prepoznavo skupnih lastnosti je bil uporabljen pristop izdelave tematskih modelov z uporabo metode LDA (Blei idr. 2003, 993−1022). Del delotoka za izdelavo tematskih modelov je bilo predhodno tudi urejanje besedila z namenom zagotavljanja čim višje kakovosti podatkov, ki so bili predmet obdelave. Ker v tej fazi ni bilo vzpostavljenega pristopa za uporabo predizdelanih modelov oziroma ta ni bila potrebna, je bila kakovost podatkov ključnega pomena za učinkovito izdelavo tematskih modelov.

Postopek izdelave tematskih modelov je bil prek vhodnih parametrov LDA (Blei idr. 2003, 993−1022) nastavljen na način, da zasleduje cilj izdelave petih različnih tematskih modelov.

Slika 42: Primer zajetih vhodnih zapisov v angleškem jeziku z izvirnimi zapisi in prepoznanimi imenskimi entitetami

S fields.bodyText	S Term
#ThankYourMP trended on Twitter on Friday, following the killing of the Labour MP Jo Cox. Th...	Friday[DATE(NE)];Labour MP Jo Cox[ORGANIZATION(NE)];Jonathan Freedland[PERSON(NE)];Cox[ORG
(World rankings as of Mon 25 June; ups and downs calculated from Mon 28 May) 1 New Zeala...	Mon 25 June[DATE(NE)];Mon 28 May[DATE(NE)];New Zealand[LOCATION(NE)];France[LOCATION(NE)];
... we must proceed according to a phased approach giving priority to an orderly withdrawal. ...	EU[ORGANIZATION(NE)];Lancaster House[ORGANIZATION(NE)];United Kingdom[LOCATION(NE)];Europ
1 A United Kingdom (12A) (Amma Asante, 2016, US/UK/Cze) 111 mins. The telling is conventio...	United Kingdom[ORGANIZATION(NE)];Amma Asante[PERSON(NE)];2016[DATE(NE)];Botswana[LOCATIC
1 Abstract Expressionism Prepare to be awed and exhilarated by some of the greatest art of t...	1940s[DATE(NE)];New York[LOCATION(NE)];Mexico[LOCATION(NE)];Jackson Pollock[PERSON(NE)];Mar
1 Adding Machine: A Musical US playwright Elmer Rice looked at the effects of big business an...	1923[DATE(NE)];October 2[DATE(NE)];2015[DATE(NE)];The Haunting Of Hill House[ORGANIZATION(NE
1 Afrobeats music festival Between Drake's funky flirtations, dancehall crew Mixpak triumphing...	Mixpak[LOCATION(NE)];UK[ORGANIZATION(NE)];NSG[ORGANIZATION(NE)];2016[DATE(NE)];Shaun[LO
1 Alvin Ailey American Dance Theater In the third programme for this company of fabulous da...	Alvin Ailey American Dance[ORGANIZATION(NE)];Wells[LOCATION(NE)];Gary Avis & Friends Avis[O
1 American Honey (15) (Andrea Arnold, 2016, UK/US) 164 mins British auteur Arnold hits the o...	Andrea Arnold[PERSON(NE)];2016[DATE(NE)];US midwest[LOCATION(NE)];Shia[LOCATION(NE)];Kate P
1 Anselm Kiefer The art of Anselm Kiefer is thick with history – and thick with paint, which he h...	Anselm Kiefer[PERSON(NE)];Nazi[ORGANIZATION(NE)];Walhalla[LOCATION(NE)];Germanic[LOCATION(
1 Anthony Braxton Former chief conductor Ilan Volkov returns to the BBC Scottish Symphony ...	Anthony Braxton Former[ORGANIZATION(NE)];Ilan Volkov[PERSON(NE)];BBC[ORGANIZATION(NE)];Jan
1 Arrival (12A) (Denis Villeneuve, 2016, US) 116 mins. A familiar alien invasion scenario yields ...	Denis Villeneuve[PERSON(NE)];2016[DATE(NE)];Earth[LOCATION(NE)];Amy Adams[ORGANIZATION(NE
1 August 1961: Britain applies to join EEC Mr Macmillan, a weary-looking father figure, at last ...	August 1961[DATE(NE)];Britain[LOCATION(NE)];EEC Mr Macmillan[ORGANIZATION(NE)];Europe[LOCAT
1 Australian Chamber Orchestra There should be a memorable start to the Edinburgh internati...	Edinburgh[LOCATION(NE)];Hall[PERSON(NE)];Siegfried Idyll[PERSON(NE)];Dutilleux[LOCATION(NE)];An
1 Barenboim v Thielemann Two orchestras that epitomise the central European symphonic tra...	Berlin Staatskapelle[PERSON(NE)];Daniel Barenboim[PERSON(NE)];Dresden Staatskapelle[ORGANIZATIC
1 Betroffenheit This collaboration between choreographer Crystal Pite and director Jonathon ...	Jonathon Young[PERSON(NE)];Wells[LOCATION(NE)];Royal Opera House[ORGANIZATION(NE)];Tue[LO
1 Beyoncé As the saying goes: when life gives you lemonade, go and see Beyoncé in Sunderla...	Beyoncé[LOCATION(NE)];Sunderland[LOCATION(NE)];Tue[LOCATION(NE)];Cardiff[LOCATION(NE)];Thu
1 Beyond Caring Alexander Zeldin's devised piece is set in a meat-packing factory where a qu...	Sheffield[LOCATION(NE)];Sea Helen McCrory[ORGANIZATION(NE)];Hester Collyer[PERSON(NE)];Terenc

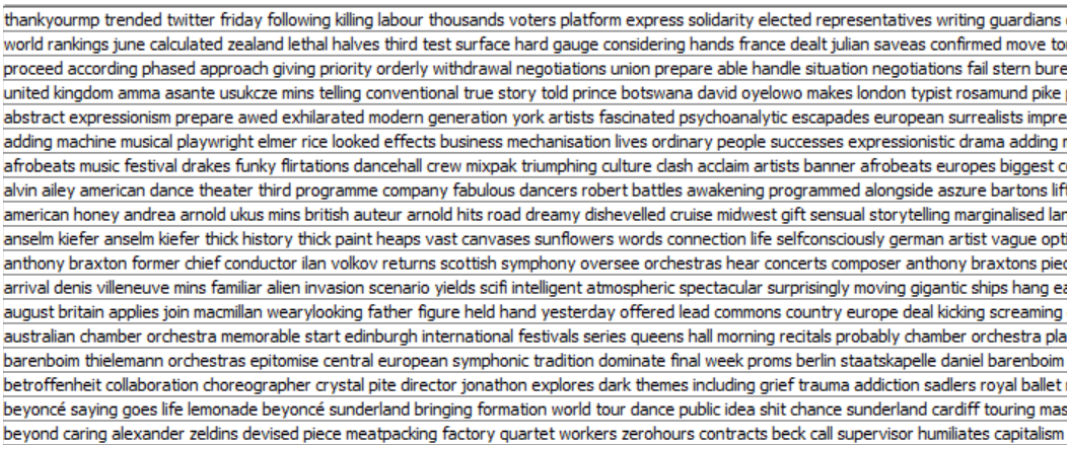
Vir: KNIME 2024.

Slika 42 prikazuje začetno obliko delotoke za izdelavo tematskih modelov. Enako obliko je predstavljal izhodni rezultat delotoka za prepoznavanje imenskih entitet. Pri tem se je ohranila tudi izvirna oblika zapisov.

Da bi omogočili čim boljšo prepoznavo vsebine, je bilo treba izvesti podproces urejanja podatkov posebej v delu odstranjevanja nepotrebnih delov besedila, ki bi lahko vplivali na kakovost izdelave tematskih modelov. Pri tem se je izvedlo tudi filtriranje nedovoljenih besed in pretvorba v obliko, ki je olajšala nadaljnjo obdelavo podatkov.

Na sliki 43 je viden rezultat podprocesa urejanja podatkov, tj. zagotavljanja višje kakovosti podatkov za nadaljnjo obdelavo.

Slika 43: Rezultat podprocesa urejanja podatkov v angleškem jeziku



Vir: KNIME 2024.

Na slikah 44 in 45 je prikazan rezultat delotoka za identifikacijo in izdelavo tematskih modelov, iz katerega so razvidni določeni tematski modeli, ključne besede, ki so bile določene individualnemu tematskemu modelu, izvirni zapisi ter prepoznane imenske entitete, poleg tega pa tudi uteži posameznih ključnih besed. Rezultati so razvrščeni glede na vsebino izvirnih zapisov. Iz uteži je možno razbrati, kakšno težo pri oblikovanju posameznega tematskega modela je imela posamezna ključna beseda, ki je bila uvrščena v nabor besed, ki so tvorile posamezen tematski model.

Slika 44: Rezultat izdelave tematskih modelov v angleškem jeziku

Topic id	Concatenate(Term)	Content	Term
topic_0	brexit, labour, government, minister, vote, deal, corbyn, secretary	#ThankYourMP trended on Twitter on Friday, following the killing of the Labour MP Jo Co…	Friday[DATE(NE)];Labour MP Jo Cox[ORGANIZATION(NE)];Jonathan Freedlan
topic_0	brexit, labour, government, minister, vote, deal, corbyn, secretary	… we must proceed according to a phased approach giving priority to an o…	EU[ORGANIZATION(NE)];Lancaster House[ORGANIZATION(NE)];United Kingdo
topic_0	brexit, labour, government, minister, vote, deal, corbyn, secretary	1 August 1961: Britain applies to join EEC Mr Macmillan, a weary-looking father figure, at…	August 1961[DATE(NE)];Britain[LOCATION(NE)];EEC Mr Macmillan[ORGANIZAT
topic_0	brexit, labour, government, minister, vote, deal, corbyn, secretary	1 Traditional marginals With a sizeable swing to the Conservatives expected, much atten…	Conservatives[ORGANIZATION(NE)];England[LOCATION(NE)];Wales[LOCATIC
topic_0	brexit, labour, government, minister, vote, deal, corbyn, secretary	1 What the royal prerogative is Fundamental to the government case is this prerogative …	2015[DATE(NE)];2011[DATE(NE)];Lord Neuberger[PERSON(NE)];Scotland[LOC
topic_0	brexit, labour, government, minister, vote, deal, corbyn, secretary	1) That exit poll, which was just too much for some of you. That shock exit poll at 10pm …	Thursday[DATE(NE)];George Osborne[PERSON(NE)];mixed night[TIME(NE)];En
topic_0	brexit, labour, government, minister, vote, deal, corbyn, secretary	1) Theresa May says she has a plan for Brexit, will see it through and has "just about" m…	May[DATE(NE)];Brexit[LOCATION(NE)];Boris Johnson[PERSON(NE)];Johnson[P
topic_0	brexit, labour, government, minister, vote, deal, corbyn, secretary	1. Making the radical mainstream Jeremy Corbyn set the radical policies announced in Liv…	Jeremy Corbyn[PERSON(NE)];Liverpool[LOCATION(NE)];John McDonnell[PERS
topic_0	brexit, labour, government, minister, vote, deal, corbyn, secretary	1. The cabinet will meet next week to agree its Brexit negotiating position. Theresa May …	May[DATE(NE)];Tuesday[DATE(NE)];Monday[DATE(NE)];Ireland[LOCATION(NE
topic_0	brexit, labour, government, minister, vote, deal, corbyn, secretary	1. The next prime minister will be a woman Apologies for stating the obvious, but the Co…	Friday 9 September[DATE(NE)];UK[ORGANIZATION(NE)];Northern Ireland[LOC
topic_0	brexit, labour, government, minister, vote, deal, corbyn, secretary	1. Voters notice most when the campaign confirms existing views While commentators fo…	LBC[ORGANIZATION(NE)];May[DATE(NE)];Jeremy Corbyn[ORGANIZATION(NE
topic_0	brexit, labour, government, minister, vote, deal, corbyn, secretary	1. Voters were making a protest about Brexit – but were not necessarily voting against …	Heathrow[LOCATION(NE)];Lib Dems[PERSON(NE)];May[DATE(NE)];David Davis
topic_0	brexit, labour, government, minister, vote, deal, corbyn, secretary	1. What compensation will those wrongly targeted receive? Asked about this at prime mi…	Wednesday[DATE(NE)];Theresa May[ORGANIZATION(NE)];Windrush[PERSON
topic_0	brexit, labour, government, minister, vote, deal, corbyn, secretary	1. What will be done to calm the markets? With the pound in freefall, will the Bank of Eng…	Bank of England[ORGANIZATION(NE)];Number 10[DATE(NE)];this morning[TIM
topic_0	brexit, labour, government, minister, vote, deal, corbyn, secretary	1. 'Get the band back together' Corbyn will have to reach out to members of the parliam…	Iain McNicol[PERSON(NE)];Emily Thornberry[PERSON(NE)];Dan Jarvis[PERSON
topic_0	brexit, labour, government, minister, vote, deal, corbyn, secretary	10 Tony Blair The politician who just won't lie down. Even the publication of the Chilcot r…	Brexit[LOCATION(NE)];Neil Hamilton OK[PERSON(NE)];Ukip[ORGANIZATION(NE
topic_0	brexit, labour, government, minister, vote, deal, corbyn, secretary	10pm The polls closed with the shock prediction that Theresa May was on course to lose …	Theresa May[ORGANIZATION(NE)];Hastings[LOCATION(NE)];Amber Rudd[PER
topic_0	brexit, labour, government, minister, vote, deal, corbyn, secretary	10pm: Polls close It will be difficult to beat the drama of the shock 10pm survey that pre…	Nuneaton[LOCATION(NE)];Nuneaton & Bedworth[ORGANIZATION(NE)];V
topic_0	brexit, labour, government, minister, vote, deal, corbyn, secretary	11 September 2001. 15 September 2008. To that list of huge stock market plunges, it lo…	September 2001[DATE(NE)];September 2008[DATE(NE)];June 2016[DATE(NE)

Vir: KNIME 2024.

Slika 45: Pregled uteži posameznih ključnih besed v angleškem jeziku glede na posamezen tematski model

Row ID	Topic id	Term	Weight
Row0	topic_0	brexit	55,875
Row1	topic_0	labour	54,361
Row2	topic_0	government	45,520
Row3	topic_0	minister	33,018
Row4	topic_0	vote	29,698
Row5	topic_0	deal	28,749
Row6	topic_0	corbyn	24,088
Row7	topic_0	secretary	23,277
Row8	topic_1	growth	25,944
Row9	topic_1	market	23,224
Row10	topic_1	bank	23,134
Row11	topic_1	business	21,072
Row12	topic_1	economy	20,610
Row13	topic_1	company	19,989
Row14	topic_1	financial	17,918
Row15	topic_1	markets	15,856
Row16	topic_2	company	13,993
Row17	topic_2	facebook	11,315
Row18	topic_2	women	9,369
Row19	topic_2	media	8,333
Row20	topic_2	data	7,745
Row21	topic_2	google	7,568
Row22	topic_2	social	6,975
Row23	topic_2	trump	6,918
Row24	topic_3	england	46,110
Row25	topic_3	ball	35,804
Row26	topic_3	australia	21,241
Row27	topic_3	match	19,715
Row28	topic_3	players	16,799
Row29	topic_3	cricket	14,594
Row30	topic_3	final	13,159
Row31	topic_3	team	12,607
Row32	topic_4	team	22,618
Row33	topic_4	race	18,085
Row34	topic_4	season	12,692
Row35	topic_4	final	11,975
Row36	topic_4	games	10,767
Row37	topic_4	sport	10,349
Row38	topic_4	gold	9,625
Row39	topic_4	champion	8,952

Vir: KNIME 2024.

Na sliki 46 je prikazano število izvirnih zapisov, ki pripadajo določenim identificiranim tematskim modelom.

Slika 46: Število izvirnih zapisov v angleškem jeziku, ki pripadajo določenim identificiranim tematskim modelom

topic_0	11933
topic_1	11064
topic_2	10832
topic_3	8497
topic_4	10081

Vir: KNIME 2024.

Predhodno je bil izpostavljen proces urejanja podatkov predvsem v okviru pristopa zagotavljanja višje kakovosti podatkov za izdelovanje tematskih modelov. Na podlagi urejenosti podatkov je možno doseči boljše rezultate in učinkovitejšo izvedbo delotoka izdelovanja tematskih modelov.

Iz slik 47 in 48 je razvidna razlika v primeru izvedenega podprocesa urejanja podatkov in brez izvedenega podprocesa urejanja podatkov.

Slika 47: Izdelani tematski modeli brez predhodno izvedenega podprocesa urejanja podatkov v angleškem jeziku

topic_0	,, ., a, ", –, :, ?, 's
topic_1	,, ., a, %, :, ", UK, growth
topic_2	,, ., a, ", :, Brexit, EU, Labour
topic_3	., ,, a, :, ",), (, –
topic_4	,, ., a, (,), :, England, ball

Vir: KNIME 2024.

Slika 48: Izdelani tematski modeli s predhodno izvedenim podprocesom urejanja podatkov v angleškem jeziku

topic_0	brexit, labour, government, minister, vote, deal, corbyn, secretary
topic_1	growth, market, bank, business, economy, company, financial, m...
topic_2	company, facebook, women, media, data, google, social, trump
topic_3	england, ball, australia, match, players, cricket, final, team
topic_4	team, race, season, final, games, sport, gold, champion

Vir: KNIME 2024.

Iz slike 48 je razvidno, da podproces urejanja podatkov ključno vpliva na učinkovitost in preglednost izdelanih tematskih modelov, saj je vsebinska opredelitev možna že na podlagi identificiranih ključnih besed, ki se nanašajo na posamezen izdelan tematski model. Obseg tematskih modelov se lahko poljubno spremeni glede na pridobljene rezultate. Za nadaljnjo obdelavo podatkov je bil obseg petih tematskih modelov ustrezno določen in tudi uporabljen.

Ker je bil namen osnovno razumeti predhodno neznane zapise v nestrukturirani obliki v njihovi kategorični obliki, je bil z izdelavo tematskih modelov ta cilj dosežen.

Za razliko od delotoka, kjer je bil izveden podproces urejanja vsebine, je rezultat, kjer omenjeni podproces ni bil izveden, vidno slabši oziroma ima kakovost podatkov velik vpliv na učinkovito izdelavo tematskih modelov. Iz slike 47 je razvidno, da je brez izvedbe podprocesa urejanja podatkov za izdelavo tematskih modelov zajetih preveč nepotrebnih podatkov, zaradi česar je tudi rezultat primerljivo mnogo slabši. Osnovna vsebinska opredelitev posameznega tematskega model in povezanih izvirnih zapisov je tako problematična oziroma niti ni možna.

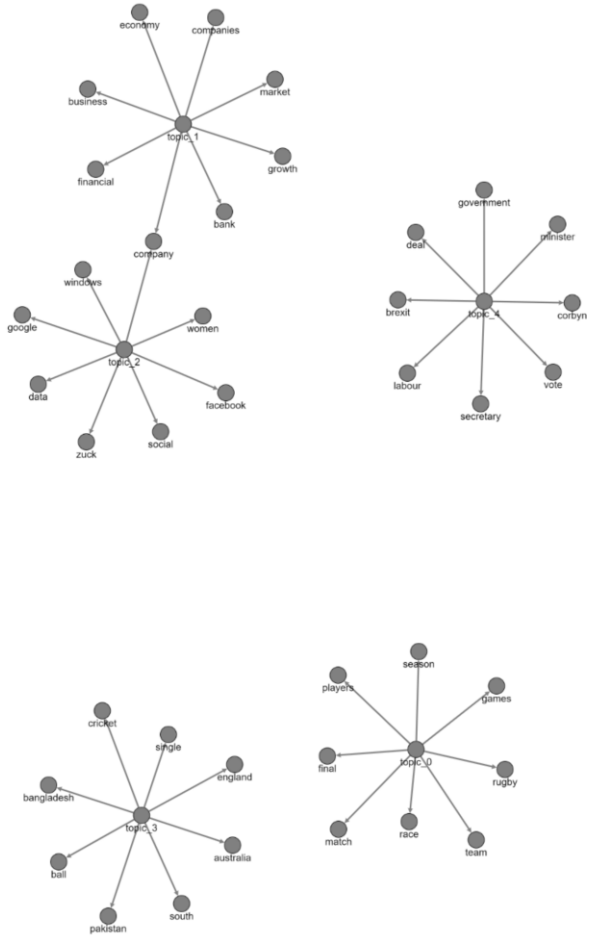
Na sliki 49 je prikazana analiza gruč izvirnih zapisov, ki so bili vključeni v izdelavo tematskih modelov, in njihova umeščenost pri izdelavi tematskih modelov glede na uporabljene pristope za izvedbo analize. Iz analize je razvidno, da so vse gruče približno enakovredno dimenzionirane.

Slika 49: Analiza gruč vseh izvirnih zapisov, ki so bili predmet izdelave tematskih modelov

Vir: KNIME 2024.

Na sliki 50 je prikazan primer povezanosti določenih individualnih tematskih modelov glede na identificirane ključne besede. Iz slike je razvidno, da gre za unikatno razporeditev ključnih besed v primeru treh izdelanih tematskih modelov, v primeru dveh izdelanih tematskih modelov pa gre za povezljivost le na eni skupni ključni besedi, kar predstavlja kakovostno distribucijo in izvirnost posameznega izdelanega tematskega modela.

Slika 50: Povezanost določenih individualnih tematskih modelov glede na identificirane ključne besede



Vir: KNIME 2024.

Celoten delotok za izdelavo tematskih modelov z izvedbo podprocesa urejanja podatkov je potreboval 36 minut, brez uporabe specializiranih jeder grafične kartice (glej sliko 51).

Slika 51: Čas v sekundah, potreben za izvedbo delotoka izdelave tematskih modelov v angleškem jeziku z izvedbo podprocesa urejanja podatkov

Mean Execution Time (s)	2160.804
Worst Execution Time (s)	2160.804
Best Execution Time (s)	2160.804
Total Execution Time (s)	2160.804
Last Execution time (s)	2160.804

Vir: KNIME 2024.

Izdelava modela za samostojno klasifikacijo vsebine

Drugi delotok v postopku prepoznave vsebine in klasifikacije zapisov je vključeval postopek za izdelavo treh različnih odločitvenih modelov. Vsi trije odločitveni modeli so bili izdelani neposredno iz zapisov, ki so bili del izdelave tematskih modelov, eden od modelov pa je bil izdelan tudi na podlagi predizdelanega modela, vključno s podrobnejšim izboljševanjem modela (angl. Fine tuning), tj. na podlagi klasificiranih izvirnih zapisov.

Zaradi primerjanja učinkovitosti izvedbe klasifikacije izvirnih zapisov so bili izvedeni trije različni pristopi:

- uporaba metode odločitvenih dreves (angl. Decision tree learning) (Karabadji idr. 2017),
- uporaba metode podpornih vektorjev (angl. Support vector machine − SVM) (Cervantes idr. 2020) in
- uporaba arhitekture in modela BERT (2023).

Pri vseh metodah je bil dodatno izveden podproces urejanja podatkov. Pri uporabi arhitekture in modela BERT (2023) je bil proces vzporedno izveden tudi brez izvedbe podprocesa urejanja podatkov.

Vhodni podatki za izvedbo delotoka so bili podatki, kot so bili pripravljeni v okviru delotoka izdelave predhodnih delotokov, ki so vključevali izvirne zapise, identificirane imenske entitete ter določene tematske modele kot obliko klasifikacije. Oblika zapisa se je ohranila skozi vse korake obdelav podatkov.

V postopku klasifikacije je bilo pri vseh izdelanih modelih uporabljeno enako razmerje razdelitve izvirnih zapisov za učenje in preverjanje pravilnosti izvedene klasifikacije. Obseg izvirnih zapisov je bil razdeljen na naključnih 70 % vseh zapisov za izdelavo odločitvenega modela in naključnih 30 % vseh zapisov za izvedbo klasifikacije in preverjanja pravilnosti izvedene klasifikacije.

Na sliki 52 sta prikazana rezultat in pravilnost izvedene klasifikacije na podlagi modela, ki je bil izdelan z metodo odločitvenih dreves (Karabadji idr. 2017).

Slika 52: Rezultat in pravilnost izvedene klasifikacije v angleškem jeziku na podlagi modela, ki je bil izdelan z metodo odločitvenih dreves

Prediction ...	topic_0	topic_1	topic_2	topic_4	topic_3
topic_0	2827	254	294	54	21
topic_1	303	2504	397	55	20
topic_2	235	371	2023	251	130
topic_4	41	32	286	2344	172
topic_3	9	15	100	175	2087

Correct classified: 11.785 Wrong classified: 3.215

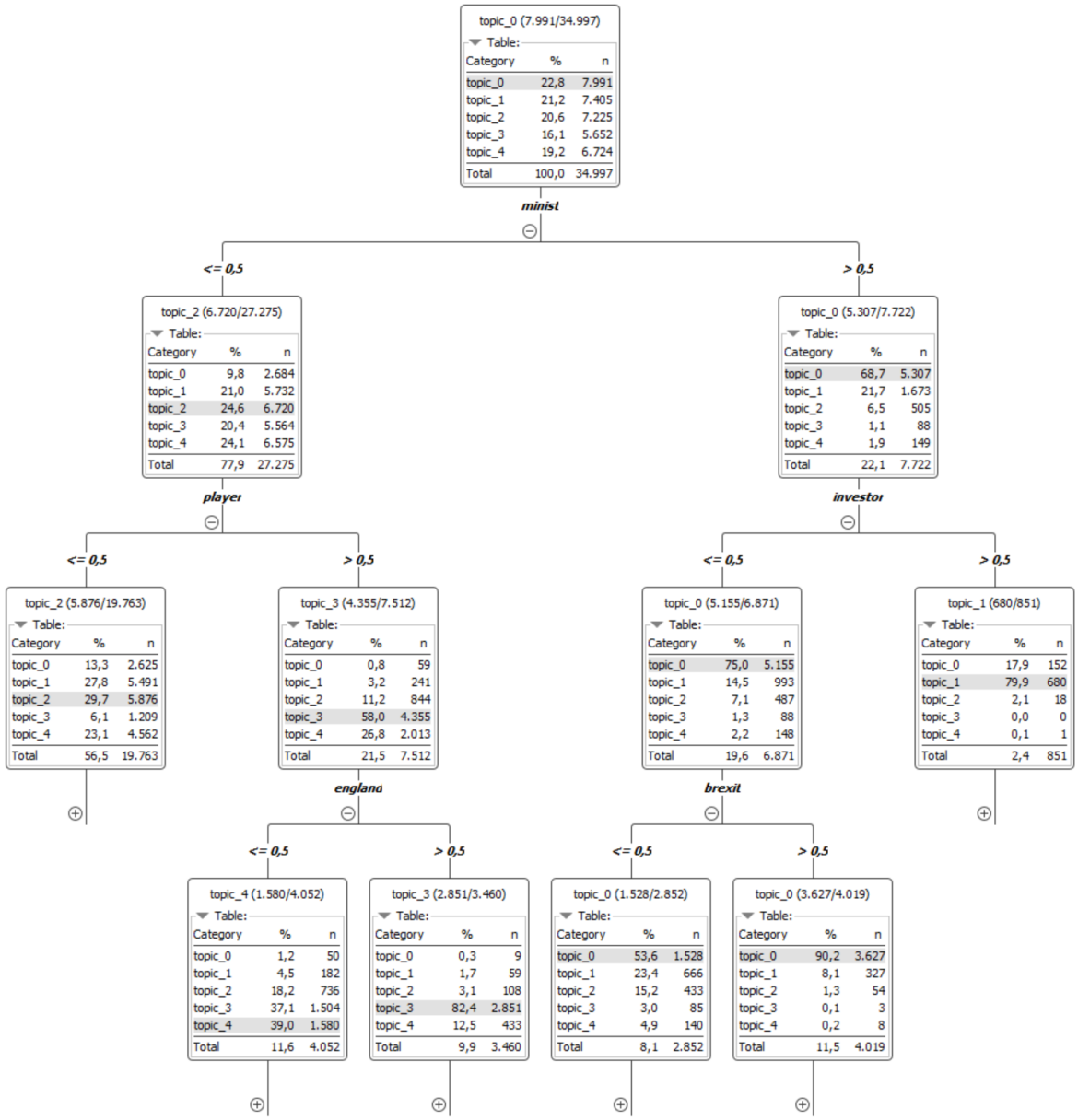
Accuracy: 78,567% Error: 21,433%

Cohen's kappa (κ): 0,731%

Vir: KNIME 2024.

Slika 53 prikazuje prerez in hierarhijo izvedene klasifikacije na podlagi modela metode Decision tree learning (Karabadji idr. 2017) s pregledom vrednotenja uteži posameznih ključnih besed glede na identificirane tematske modele.

Slika 53: Prikaz hierarhije odločitvenega modela metode »odločitveno drevo« v angleškem jeziku



Vir: KNIME 2024.

Slika 54 prikazuje rezultat in pravilnost izvedene klasifikacije na podlagi modela, ki je bil izdelan z metodo podpornih vektorjev (angl. Support vector machine − SVM) (Cervantes idr. 2020).

Slika 54: Rezultat in pravilnost izvedene klasifikacije v angleškem jeziku na podlagi modela, ki je bil izdelan z metodo SVM

Prediction ...	topic_0	topic_2	topic_3	topic_1	topic_4
topic_0	3184	154	8	90	18
topic_2	115	2639	28	142	109
topic_3	12	52	2292	9	87
topic_1	93	134	9	2913	14
topic_4	15	125	93	13	2652

Correct classified: 13.680 Wrong classified: 1.320

Accuracy: 91,2% Error: 8,8%

Cohen's kappa (κ): 0,89%

Vir: KNIME 2024.

Slika 55 prikazuje rezultat in pravilnost izvedene klasifikacije na podlagi modela, ki je bil izdelan z arhitekturo in modelom BERT (2023), brez izvedbe podprocesa urejanja podatkov.

Slika 55: Rezultat in pravilnost izvedene klasifikacije na podlagi modela, ki je bil izdelan z arhitekturo in metodo BERT v angleškem jeziku, brez izvedbe podprocesa urejanja podatkov

Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specificity	D F-meas...	D Accuracy	D Cohen'...
topic_0	3263	118	11356	263	0.925	0.965	0.925	0.99	0.945	?	?
topic_1	3032	188	11584	196	0.939	0.942	0.939	0.984	0.94	?	?
topic_2	2566	428	11803	203	0.927	0.857	0.927	0.965	0.891	?	?
topic_4	2795	115	11954	136	0.954	0.96	0.954	0.99	0.957	?	?
topic_3	2457	38	12416	89	0.965	0.985	0.965	0.997	0.975	?	?
Overall	?	?	?	?	?	?	?	?	?	0.941	0.926

Vir: KNIME 2024.

Slike 52, 54 in 55 prikazujejo rezultat izvedbe delotoka klasifikacije vsebine. Iz rezultatov je razvidno, da je bila natančnost klasifikacije izmerjena skladno z uporabo individualnega odločitvenega modela glede na določen obseg za izvedbo izdelave odločitvenega modela in izvedbo klasifikacije na podlagi izdelanega odločitvenega modela.

Natančnost klasifikacije si sledi v naslednjem vrstnem redu od najnatančnejšega odločitvenega modela do najmanj natančnega:

- 94 % natančnost – uporaba arhitekture in modela BERT (2023),
- 91 % natančnost – uporaba metode podpornih vektorjev – SVM (Cervantes idr. 2020),
- 79 % natančnost – uporaba metode odločitvenih dreves (Karabadji idr. 2017).

Pri merjenju natančnosti je bila najbolj natančna izvedba klasifikacije z uporabo arhitekture in metode BERT (2023), ki je predstavljala izdelavo odločitvenega modela na podlagi predizdelanega modela. Visoko natančnost je med ostalim možno pripisati tudi uporabi predizdelanega modela, ki je bil izdelan/učen na podlagi izjemno velikega obsega podatkov.

Izjemno dober rezultat je bil izmerjen tudi v primeru klasifikacije z uporabo metode podpornih vektorjev – SVM (Cervantes idr. 2020), ki pa za razliko od predhodnega metoda, kjer je bila izmerjena najvišja natančnost, ni bil na voljo predizdelan model, ampak je bil

odločitveni model izdelan izključno na podlagi izvirnih zapisov in izdelanih tematskih modelov.

V primerjavi z rezultati prvih dveh metod in odločitvenih modelov je bil precej slabši rezultat izmerjen v primeru klasifikacije z uporabo metode odločitvenih dreves (Karabadji idr. 2017), ki je bil za 12 do 15 % manj natančen.

V okviru raziskave je bilo pri najbolj natančnem pristopu z uporabo arhitekture in metode BERT (2023) preverjeno tudi, ali bi bila natančnost klasifikacije višja s predhodno izvedenim podprocesom urejanja podatkov glede na to, da je bil predizdelani model izdelan na podlagi polnih besedil.

Slika 56 prikazuje rezultat in pravilnost izvedene klasifikacije na podlagi modela, ki je bil izdelan z arhitekturo in modelom BERT (2023) z izvedbo podprocesa urejanja podatkov.

Slika 56: Rezultat in pravilnost izvedene klasifikacije na podlagi modela, ki je bil izdelan z arhitekturo in metodo BERT v angleškem jeziku z izvedbo podprocesa urejanja podatkov

Prediction ...	topic_0	topic_2	topic_1	topic_4	topic_3
topic_0	3201	165	179	7	0
topic_2	84	2405	113	26	4
topic_1	48	115	2906	3	3
topic_4	36	246	21	2788	150
topic_3	5	55	6	96	2338

Correct classified: 13.638 Wrong classified: 1.362

Accuracy: 90,92% Error: 9,08%

Cohen's kappa (κ): 0,886%

Vir: KNIME 2024.

Iz rezultatov je razvidno, da je natančnost izvedene klasifikacije na podlagi modela, ki je bil izdelan z arhitekturo in modelom BERT (2023) z izvedbo podprocesa urejanja podatkov, manjša (91 %) od natančnosti v okviru postopka klasifikacije, ki je bil izveden brez izvedbe podprocesa urejanja podatkov (94 %), kar je možno pripisati arhitekturi predizdelanega modela BERT (2023), ki je v osnovi »maskirani jezikovni model« (angl. Masked language model), pri izdelavi katerega je bil uporabljen postopek prekrivanja določenega dela podatkov, da bi se lahko doseglo uporabno predvidevanje za pravilnost podatkov. Glede na to, da je bil predizdelani model učen/treniran na velikih količinah podatkov, je vključeval tudi veliko nepotrebnih podatkov, ki pa pri izdelavi odločitvenega modela niso ključno vplivali na učinkovitost izvedene klasifikacije.

Celoten delotok za izdelavo odločitvenega modela in izvedbo klasifikacije izvirnih zapisov je glede na individualen pristop zahteval naslednji čas (slike 57, 58 in 59):

- 20 ur in 37 minut – uporaba arhitekture in model BERT (2023) – z uporabo procesorskih jeder grafične kartice,
- 15 ur in 35 minut – uporaba metode podpornih vektorjev – SVM (Cervantes idr. 2020) – brez uporabe procesorskih jeder grafične kartice,

- 1 ura in 6 minut – uporaba metode odločitvenih dreves (Karabadji idr. 2017) – brez uporabe procesorskih jeder grafične kartice.

Slika 57: Čas v sekundah, potreben za izvedbo delotoka izdelave odločitvenega modela in izvedbo klasifikacije izvirnih zapisov v angleškem jeziku – arhitektura in model BERT

Mean Execution Time (s)	73340.547
Worst Execution Time (s)	73340.547
Best Execution Time (s)	73340.547
Total Execution Time (s)	73340.547
Last Execution time (s)	73340.547

Vir: KNIME 2024.

Slika 58: Čas v sekundah, potreben za izvedbo delotoka izdelave odločitvenega modela in izvedbo klasifikacije izvirnih zapisov v angleškem jeziku – metoda podpornih vektorjev – SVM

Mean Execution Time (s)	55259.593
Worst Execution Time (s)	55259.593
Best Execution Time (s)	55259.593
Total Execution Time (s)	55259.593
Last Execution time (s)	55259.593

Vir: KNIME 2024.

Slika 59: Čas v sekundah, potreben za izvedbo delotoka izdelave odločitvenega modela in izvedbo klasifikacije izvirnih zapisov v angleškem jeziku – metoda odločitvenih dreves

Mean Execution Time (s)	3839.645
Worst Execution Time (s)	3839.645
Best Execution Time (s)	3839.645
Total Execution Time (s)	3839.645
Last Execution time (s)	3839.645

Vir: KNIME 2024.

Uvodoma je bila predstavljena omejitev uporabe procesorskih jeder grafične kartice v primeru, da aktivnosti v posamezni veji delotoka ne bi bile izvedljive v času do 72 ur. V primeru izdelave modela in izvedbe klasifikacije izvirne vsebine so bila grafična procesorska jedra uporabljena le v primeru arhitekture in modela BERT (2023), kar kaže na to, da je bil za izdelavo končnega modela uporabljen obsežen predizdelan model, ki je za izvedbo celotnega delotoka zahteval veliko več procesorske moči kot pri ostalih pristopih.

Vsi izmerjeni časi tudi na splošno kažejo, kako zelo procesorsko obremenjujoč je proces izdelave odločitvenih modelov, ki pa se še naprej veča z velikostjo obsega izvirnih zapisov, na podlagi katerih se izvaja učenje in posledično klasifikacija vsebine.

3.4.3 Delotok za izdelavo naslova popisne enote v angleškem jeziku

Tretja faza izdelava popisne enote je vključevala izdelavo naslova popisne enote in dodajanje ostalih fiksnih podatkov skladno za zapolnitev sheme po standardu ISAD(G) (2000).

Na sliki 60 je prikazan rezultat izdelave naslova popisne enote glede na pristop abstraktnega povzemanja. Parametri izdelave naslova popisne enote so bili nastavljeni na način, da izdela naslov v obliki ene povedi.

Slika 60: Rezultat izdelave naslova popisne enote v angleškem jeziku

Title	Classification	Content	Entity
The House of Lords has voted in favour of an amendment that would allow MPs to vote on a final Brexit deal.	topic_0	A much-debated and much-amended section of the EU withdraw...	EU[ORGANIZATION(NE)];Wednesday[DATE(NE)];Lords[LOCATION(NE
A Swedish company that claims to be the world's most effective contraceptive has come under fire.	topic_2	A much-hyped birth control app has been reported to Swedish au...	Europe[LOCATION(NE)];Cern[ORGANIZATION(NE)];Elina Berglund[PE
What does it mean for Leeds Rhinos and their hopes of reaching the Super League play-offs?	topic_4	A much-improved performance? Without question. But with a gap...	Leeds[LOCATION(NE)];Brian McDermott[PERSON(NE)];18-12[DATE(NE
The last remaining public phone box in the UK is to be removed from the streets on Monday.	topic_2	A much-loved but almost obsolete feature is set to vanish from t...	20%[PERCENT(NE)];90%[PERCENT(NE)];last decade[DATE(NE)];193
We asked you to send in your photos of what you think has changed the view of London.	topic_2	A much-loved view of St Paul's Cathedral has been affected by a...	London[LOCATION(NE)];Sadiq Khan[PERSON(NE)];Manhattan Loft Ga
Australia beat New Zealand by eight wickets in a rain-affected Twenty20 international in Hobart.	topic_3	A much-needed T20 win for Australia, and thoroughly deserved t...	Australia[LOCATION(NE)];Stanlake[LOCATION(NE)];Black Caps[ORGA
Australia's government has delivered a budget deficit of $4.4bn (Â£3.2bn), less than expected.	topic_1	A multibillion-dollar cut to spending on social services and higher-t...	Scott Morrison[PERSON(NE)];Australia[LOCATION(NE)];$4bn[MONEY(
Wasps ended a run of five straight Premiership defeats with a hard-fought victory over Leicester at the Ricoh Arena.	topic_3	A murky, wintry evening in Coventry but this was a performance ...	Coventry[LOCATION(NE)];Saturday[DATE(NE)];night[TIME(NE)];Kear
The Ragged School, a Victorian school that closed in 1907, is to get a £4.3m makeover.	topic_2	A museum commemorating the canalside warehouses that once p...	London[LOCATION(NE)];Ragged School Museum[ORGANIZATION(NE)

Vir: KNIME 2024.

Celoten delotok za izdelavo naslovov popisne enote na podlagi vsebine izvirnih zapisov je potreboval 63 ur in 11 minut, brez uporabe specializiranih jeder grafične kartice (glej sliko 61).

Slika 61: Čas v sekundah, potreben za izvedbo delotoka izdelave naslovov popisne enote v angleškem jeziku

Mean Execution Time (s)	227222.573
Worst Execution Time (s)	227222.573
Best Execution Time (s)	227222.573
Total Execution Time (s)	227222.573
Last Execution time (s)	227222.573

Vir: KNIME 2024.

Iz časa, ki je bil potreben za izdelavo naslovov popisne enote, je možno zaključiti, da gre za najbolj procesorsko obremenjen del aktivnosti izmed vseh aktivnosti v vseh delotokih, ki so bili predmet raziskave.

V zadnjem delu delotoka izdelave popisne enote so bili dodani fiksni podatki (glej sliko 62), ki so skupaj s podatki, ki so bili ustvarjeni v predhodnih delotokih, vključno z izdelanimi naslovi popisne enote, tvorili končno obliko popisne enote.

Slika 62: Dodajanje fiksnih podatkov za izdelavo končne oblike popisne enote v angleškem jeziku

Refere...	Accumu...	LevelOf...	ExtentAnd...	NameOfCreator	AdministrativeAndBiographicalHistory
SI AM 00000	01.maj.2022	Item	1 electronic record	Guardian News & Media Limited	The Guardian is a British daily newspaper. It was founded in 1821 as The
SI AM 00001	01.maj.2022	Item	1 electronic record	Guardian News & Media Limited	The Guardian is a British daily newspaper. It was founded in 1821 as The
SI AM 00002	01.maj.2022	Item	1 electronic record	Guardian News & Media Limited	The Guardian is a British daily newspaper. It was founded in 1821 as The
SI AM 00003	01.maj.2022	Item	1 electronic record	Guardian News & Media Limited	The Guardian is a British daily newspaper. It was founded in 1821 as The
SI AM 00004	01.maj.2022	Item	1 electronic record	Guardian News & Media Limited	The Guardian is a British daily newspaper. It was founded in 1821 as The

Vir: KNIME 2024.

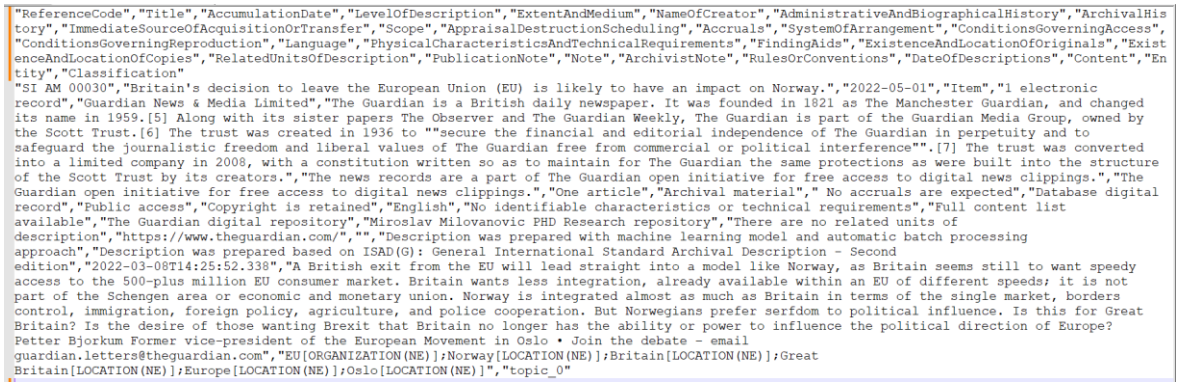
Izhodna vrednost delotoka za izdelavo naslova popisne enote, ki je hkrati predstavljala tudi končno oziroma izhodno vrednost vseh delotokov, je tekstovna datoteka, ki je vključevala naslednje podatke:

- individualen izvirni zapis,
- prepoznane imenske entitet, nanašajoče se na individualen izvirni zapis,

- določena klasifikacija na podlagi izdelanih tematskih modelov – v povezavi z individualnim izvirnim zapisom,
- izdelani naslov popisne enote – nanašajoč se na individualen izvirni zapis,
- fiksni podatki za dopolnitev sheme ISAD(G) (2000) – z izjemo oznake popisne enote, enako za vse izvirne zapise.

Na sliki 63 je prikazan individualen zapis v izvoženi tekstovni datoteki, ki vključuje vse podatke za izvedbo predogleda popisne enote.

Slika 63: Prikaz individualnega zapisa v angleškem jeziku v izvoženi tekstovni datoteki, ki vključuje vse podatke za izvedbo predogleda popisne enote



"ReferenceCode","Title","AccumulationDate","LevelOfDescription","ExtentAndMedium","NameOfCreator","AdministrativeAndBiographicalHistory","ArchivalHistory","ImmediateSourceOfAcquisitionOrTransfer","Scope","AppraisalDestructionScheduling","Accruals","SystemOfArrangement","ConditionsGoverningAccess","ConditionsGoverningReproduction","Language","PhysicalCharacteristicsAndTechnicalRequirements","FindingAids","ExistenceAndLocationOfOriginals","ExistenceAndLocationOfCopies","RelatedUnitsOfDescription","PublicationNote","Note","ArchivistNote","RulesOrConventions","DateOfDescriptions","Content","Entity","Classification"
"SI AM 00030","Britain's decision to leave the European Union (EU) is likely to have an impact on Norway.","2022-05-01","Item","1 electronic record","Guardian News & Media Limited","The Guardian is a British daily newspaper. It was founded in 1821 as The Manchester Guardian, and changed its name in 1959.[5] Along with its sister papers The Observer and The Guardian Weekly, The Guardian is part of the Guardian Media Group, owned by the Scott Trust.[6] The trust was created in 1936 to ""secure the financial and editorial independence of The Guardian in perpetuity and to safeguard the journalistic freedom and liberal values of The Guardian free from commercial or political interference"".[7] The trust was converted into a limited company in 2008, with a constitution written so as to maintain for The Guardian the same protections as were built into the structure of the Scott Trust by its creators.","The news records are a part of The Guardian open initiative for free access to digital news clippings.","The Guardian open initiative for free access to digital news clippings.","One article","Archival material"," No accruals are expected","Database digital record","Public access","Copyright is retained","English","No identifiable characteristics or technical requirements","Full content list available","The Guardian digital repository","Miroslav Milovanovic PHD Research repository","There are no related units of description","https://www.theguardian.com/","","Description was prepared with machine learning model and automatic batch processing approach","Description was prepared based on ISAD(G): General International Standard Archival Description - Second edition","2022-03-08T14:25:52.338","A British exit from the EU will lead straight into a model like Norway, as Britain seems still to want speedy access to the 500-plus million EU consumer market. Britain wants less integration, already available within an EU of different speeds; it is not part of the Schengen area or economic and monetary union. Norway is integrated almost as much as Britain in terms of the single market, borders control, immigration, foreign policy, agriculture, and police cooperation. But Norwegians prefer serfdom to political influence. Is this for Great Britain? Is the desire of those wanting Brexit that Britain no longer has the ability or power to influence the political direction of Europe? Petter Bjorkum Former vice-president of the European Movement in Oslo • Join the debate - email guardian.letters@theguardian.com","EU[ORGANIZATION(NE)];Norway[LOCATION(NE)];Britain[LOCATION(NE)];Great Britain[LOCATION(NE)];Europe[LOCATION(NE)];Oslo[LOCATION(NE)]","topic_0"

Vir: Lastni 2024.

3.4.4 Delotok za prepoznavanje imenskih entitet v slovenskem jeziku

Namen obdelave izvirnih zapisov v slovenskem jeziku je bil, da se preveri ustreznost vseh izdelanih delotokov in njihovo uporabnost v primeru, da gre pri obdelavi za izvirne zapise, ki so v drugih jezikih in ne zgolj v angleškem. Vsi nadaljnji opisi rezultatov bodo osredotočeni na opis pridobljenih rezultatov in morebitna odstopanja ali spremembe od izdelanih delotokov, ki so bili izvedeni za doseganje pridobljenih rezultatov.

Prva faza izdelave popisne enote, tj. prepoznava imenskih entitet, je vključevala samo en postopek za razliko od delotoka, v katerem so bili obdelani izvirni zapisi v angleškem jeziku, in sicer pristop s prepoznavo imenskih entitet na podlagi predizdelanih modelov. Iz delotoka za prepoznavo imenskih entitet na podlagi slovarjev v angleškem jeziku je bilo razvidno, da pristop ni učinkovit, če ni na voljo dovolj obsežnih slovarjev za izdelavo modela, kar jih v primeru slovenskega jezika tudi ni bilo. Posledično je bila izvedena prepoznava imenskih entitet na podlagi predizdelanih modelov.

Delotok za prepoznavanje imenskih entitet na podlagi predizdelanih modelov

V delotoku za prepoznavanje imenskih entitet v slovenskem jeziku sta bila z vidika učinkovitosti prepoznave preverjena dva pristopa, ki sta že pri prepoznavanju imenskih entitet v angleškem jeziku zagotovila najboljše rezultate. Uporabljen je bil pristop z uporabo

Stanford Named Entity Recognizer (Finkel idr. 2005) ter pristop z uporabo SPACY EntityRecognizer – sl-core-news-lg (SL_core__news_lg 2023).

V primeru pristopa z uporabo Stanford Named Entity Recognizer (Finkel idr. 2005) ni bilo izvedenih nobenih sprememb v primerjavi s prepoznavo v angleškem jeziku, tudi model za prepoznavo imenskih entitet je ostal enak.

V primeru pristopa z uporabo SPACY EntityRecognizer – sl-core-news-lg (SL_core__news_lg 2023) je bila izvedena sprememba pri izbiri predizdelanega modela, kjer je bil uporabljen model, treniran/učen na podlagi besedila v slovenskem jeziku.

Delotok za prepoznavanje imenskih entitet z uporabo predizdelanih modelov je za razliko od prepoznave imenskih entitet v angleškem jeziku vključeval tri različne tipe imenskih entitet:

- organizacije,
- lokacije in
- osebe.

Na sliki 64 je prikazana primerjava rezultatov prepoznave dveh modelov, vključno z ročno indeksiranimi vrednostmi. Rezultati prikazujejo, da je izmed obeh predizdelanih modelov večjo natančnost prepoznave dosegel pristop z uporabo SPACY EntityRecognizer – sl-core-news-lg (SL_core__news_lg 2023). Pri uporabi Stanford Named Entity Recognizer (Finkel idr. 2005) poleg manjše učinkovitosti prepoznavanja imenskih entitet izstopa predvsem neustrezno določanje tipov entitet, kar je v prvi vrsti možno pripisati dejstvu, da gre za širok predizdelan model, ki ni bil treniran/učen izključno na besedilih, ki so bila v slovenskem jeziku.

Iz prikazanih rezultatov je razvidno, da je za večjo učinkovitost priporočljivo, da se uporabi predizdelane modele, ki so bili trenirani/učeni na osnovi besedila, ki je v jeziku, v katerem je namen izvajati prepoznavo imenskih entitet.

Slika 64: Primerjava rezultatov prepoznave dveh modelov, vključno z ročno indeksiranimi vrednostmi v slovenskem jeziku

Personal anotation		StanfordNLP	SPACY
Mejra Festič			
Sloveniji	LOCATION	Sloveniji[LOCATION(NE)]	Sloveniji[LOC(SPACY_NE)]
Unicredit	ORGANIZATION		banke Unicredit[ORG(SPACY_NE)]
Banke Slovenije	ORGANIZATION	Banka Slovenije[PERSON(NE)	Banke Slovenije[ORG(SPACY_NE)]
Merkurju	ORGANIZATION	Merkurju[ORGANIZATION(NE)	Merkurja[LOC(SPACY_NE)]
Salonit Anhovo	ORGANIZATION	Salonit Anhovo[PERSON(NE)	Salonit Anhovo[ORG(SPACY_NE)]
Hidria	ORGANIZATION	Hidria[LOCATION(NE)	Hidria[ORG(SPACY_NE)]
Iskratel	ORGANIZATION	Iskratel[LOCATION(NE)	Iskratel[ORG(SPACY_NE)]
Marko Kranjec	PERSON	Marko Kranjec[PERSON(NE)	Marko Kranjec[PER(SPACY_NE)]
Božo Jašovič	PERSON	Božo Jašovič[PERSON(NE)]	Božo Jašovič[PER(SPACY_NE)]
Istrabenza	ORGANIZATION	Istrabenza[LOCATION(NE)]	Istrabenza[ORG(SPACY_NE)]
Pivovarne Laško	ORGANIZATION	Pivovarne Laško[LOCATION(NE)]	Pivovarne Laško[ORG(SPACY_NE)]
Zvon Ena	ORGANIZATION	Zvon Ena Holdingu[ORGANIZATION(NE)]	Zvon[ORG(SPACY_NE)]
Banke Celje	ORGANIZATION	Banke Celje[ORGANIZATION(NE)]	Banke Celje Zvon[ORG(SPACY_NE)]
Infond Holding	ORGANIZATION		Infond Holding[ORG(SPACY_NE)]
Romano Pajenk	PERSON	Romano Pajenk[PERSON(NE)];	Romano Pajenk[PER(SPACY_NE)]
Dnevniku	ORGANIZATION	Dnevniku[LOCATION(NE)]	Dnevniku[ORG(SPACY_NE)]
Islandiji	LOCATION	Islandiji[LOCATION(NE)];	Islandiji[LOC(SPACY_NE)]
Kaupthing	ORGANIZATION		banki Kaupthing[ORG(SPACY_NE)]
Probanke	ORGANIZATION		Probanke[ORG(SPACY_NE)]
Factor banki	ORGANIZATION		Factor banki[ORG(SPACY_NE)]
KD Banke	ORGANIZATION		KD Banke[ORG(SPACY_NE)]
Janezu Janši	PERSON	Janezu Janši[PERSON(NE)];	Janezu Janši[PER(SPACY_NE)]
NKBM	ORGANIZATION	NKBM[ORGANIZATION(NE)];	NKBM[ORG(SPACY_NE)]
Igor Bavčar	PERSON		Igor Bavčar[PER(SPACY_NE)]
Jugoslavije	LOCATION		Jugoslavije[LOC(SPACY_NE)]
Hrvaškem	LOCATION		Hrvaškem[LOC(SPACY_NE)]
Primož Britovšek	PERSON	Primož Britovšek[PERSON(NE)];	Primož Britovšek[PER(SPACY_NE)]
Aleš Hauc	PERSON	Aleš Hauc[PERSON(NE)];	Aleš Hauc[PER(SPACY_NE)]
SDS	ORGANIZATION		SDS[ORG(SPACY_NE)]
Matjaž Kovačič	PERSON		Matjaž Kovačič[PER(SPACY_NE)]
Republika Slovenija	LOCATION		Republika Slovenija[LOC(SPACY_NE)]
NLB	ORGANIZATION	NLB[ORGANIZATION(NE)];	NLB[ORG(SPACY_NE)];DUTB[ORG(SPACY_NE)]
ministrstva za gospodarski razvoj in tehnologijo	ORGANIZATION		ministrstva za gospodarski razvoj in tehnologijo[ORG(SPACY_NE)]
Družba za upravljanje terjatev bank	ORGANIZATION		
Bruslju	LOCATION		Bruslju[LOC(SPACY_NE)]
Frankfurtu	LOCATION	Frankfurtu[LOCATION(NE)];	Frankfurtu[LOC(SPACY_NE)]
Ivana Ribnikarja	PERSON	Ivana Ribnikarja[LOCATION(NE)];	Ivana Ribnikarja[PER(SPACY_NE)]

Vir: Lastni 2024.

Za nadaljnjo obdelavo podatkov so bili uporabljeni rezultati modela in pristopa SPACY (2023).

Celoten delotok je za prepoznavanje imenskih entitet v postopku uporabe modela in pristopa SPACY (SPACY 2023) potreboval 11 minut, brez uporabe specializiranih jeder grafične kartice (glej sliko 65).

Slika 65: Čas v sekundah, potreben za izvedbo delotoka prepoznavanja imenskih entitet uporabe modela in pristopa SPACY v slovenskem jeziku

Mean Execution Time (s)	678.922
Worst Execution Time (s)	678.922
Best Execution Time (s)	678.922
Total Execution Time (s)	678.922
Last Execution time (s)	678.922

Vir: KNIME 2024.

3.4.5 Delotok za prepoznavanje nestrukturirane vsebine in klasifikacijo vsebine v slovenskem jeziku

Druga faza izdelave popisne enote v slovenskem jeziku je vključevala enake korake in metode, kot so bili opredeljeni pri obdelavi izvirnih zapisov v angleškem jeziku. Vhodni podatki so bili pripravljeni na enak način (glej sliko 66).

Postopek izdelave tematskih modelov v slovenskem jeziku je bil prek vhodnih parametrov LDA (Blei idr. 2003, 993−1022) nastavljen na način, da zasleduje cilj izdelave petih različnih tematskih modelov.

Slika 66: Primer zajetih vhodnih zapisov z izvirnimi zapisi in prepoznanimi imenskimi entitetami v slovenskem jeziku

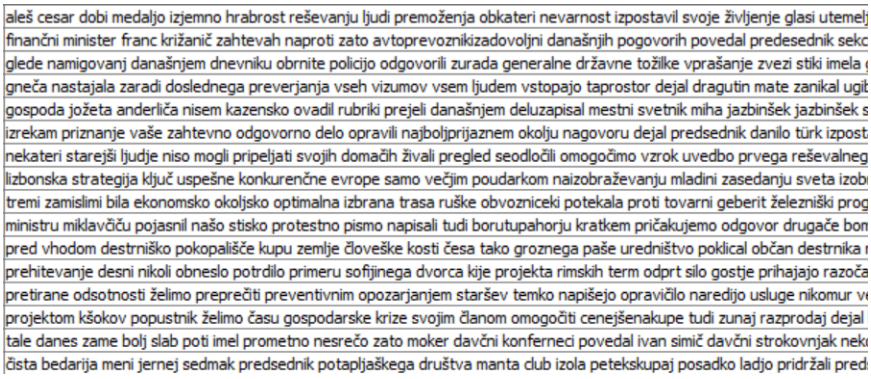
" Aleš Cesar dobi medaljo za izjemno hrabrost pri reševanju ljudi in premoženja, obkateri je v nevarnost izpo...	Aleš Cesar[PER(SPACY_NE_sl_core_news_lg-3.7.0)];Danila Türka[PER(SPACY_NE_sl_core_news_lg-3.7.0)]
" Finančni minister Franc Križanič nam je šel pri zahtevah naproti, zato so avtoprevoznikizadovoljni," je po da...	Franc Križanič[PER(SPACY_NE_sl_core_news_lg-3.7.0)];Obrtnizbornici Slovenije[ORG(SPACY_NE_sl_core_n
" Glede namigovanj v današnjem Dnevniku se obrnite na policijo," so za STA odgovorili zurada generalne drž...	Dnevniku[ORG(SPACY_NE_sl_core_news_lg-3.7.0)];STA[ORG(SPACY_NE_sl_core_news_lg-3.7.0)];Barbara
" Gneča je nastajala zaradi doslednega preverjanja vseh vizumov vsem ljudem, ki pač vstopajo v taprostor,"...	Dragutin Mate[PER(SPACY_NE_sl_core_news_lg-3.7.0)];poMatejevih[DERIV_PER(SPACY_NE_sl_core_news
" Gospoda Jožeta Anderliča nisem kazensko ovadil," je v rubriki Prejeli smo v današnjem Deluzapisal mestni s...	Jožeta Anderliča[PER(SPACY_NE_sl_core_news_lg-3.7.0)];Miha Jazbinšek[PER(SPACY_NE_sl_core_news_lg
" Izrekam vam priznanje za vaše zahtevno in odgovorno delo, ki ste ga opravili v ne najboljprijaznem okolju,"...	Danilo Türk[PER(SPACY_NE_sl_core_news_lg-3.7.0)];Onkološki inštitutprepoznal velik[MISC(SPACY_NE_sl_c
" Ker nekateri starejši ljudje niso mogli pripeljati svojih domačih živali na pregled, smo seodločili, da jim to omo...	vSloveniji[LOC(SPACY_NE_sl_core_news_lg-3.7.0)];zurnal24.si[ORG(SPACY_NE_sl_core_news_lg-3.7.0)];N
" Lizbonska strategija bo ključ do uspešne in konkurenčne Evrope samo z večjim poudarkom naizobraževanju...	Lizbonska strategija[MISC(SPACY_NE_sl_core_news_lg-3.7.0)];Evrope[LOC(SPACY_NE_sl_core_news_lg-3.
" Med tremi zamislimi je bila kot ekonomsko in okoljsko optimalna izbrana trasa ruške obvoznice,ki bo potekala...	Ruš[LOC(SPACY_NE_sl_core_news_lg-3.7.0)];tovarni Geberit[ORG(SPACY_NE_sl_core_news_lg-3.7.0)];Lim
" Ministru Miklavčiču sem pojasnil našo stisko. Protestno pismo pa smo napisali tudi BorutuPahorju. V kratkem ...	Miklavčiču[PER(SPACY_NE_sl_core_news_lg-3.7.0)];BorutuPahorju[PER(SPACY_NE_sl_core_news_lg-3.7.0)
" Pred vhodom na destrniško pokopališče so na kupu zemlje človeške kosti. Česa tako groznega paše ne!" je ...	Destrnika[PER(SPACY_NE_sl_core_news_lg-3.7.0)];Jančič[PER(SPACY_NE_sl_core_news_lg-3.7.0)];Lenart
" Prehitevanje po desni se nikoli ni obneslo, kar se je potrdilo v primeru Sofijinega dvorca, kije le del projekta ...	Sofijinega dvorca[LOC(SPACY_NE_sl_core_news_lg-3.7.0)];Drago Zupan[PER(SPACY_NE_sl_core_news_lg-
" Pri nas pretirane odsotnosti želimo preprečiti z preventivnim opozarjanjem staršev, da s tem,ko napišejo op...	Gimnazije Šentvid Jaka Erker[ORG(SPACY_NE_sl_core_news_lg-3.7.0)];Gimnaziji Šiška[LOC(SPACY_NE_sl_c
" S projektom Kšokov Popustnik želimo v času gospodarske krize svojim članom omogočiti cenejšenakupe tudi...	Klemen Zonta[PER(SPACY_NE_sl_core_news_lg-3.7.0)];Klubu študentov občine Koper[ORG(SPACY_NE_sl_c
" Tale dan je danes zame bolj slab, na poti sem se imel prometno nesrečo in sem zato ves moker," je na davč...	Ivan Simič[PER(SPACY_NE_sl_core_news_lg-3.7.0)];Dursa[ORG(SPACY_NE_sl_core_news_lg-3.7.0)];Simič[
" To je čista bedarija," meni Jernej Sedmak, predsednik potapljaškega društva Manta Club Izola, ki so ga v p...	Jernej Sedmak[PER(SPACY_NE_sl_core_news_lg-3.7.0)];potapljaškega društva[ORG(SPACY_NE_sl_core_n
" To vsi ugotavljamo in vidimo, in napačno bi bilo, če bi si zatiskali oči," je danes pozasedanju ministrov EU, pr...	EU[ORG(SPACY_NE_sl_core_news_lg-3.7.0)];Žerjav[ORG(SPACY_NE_sl_core_news_lg-3.7.0)];DARS[ORG(

Vir: KNIME 2024.

Na podlagi delotoka je bil za izdelavo tematskih modelov v slovenskem jeziku ravno tako izveden podproces urejanja podatkov.

Iz prikaza na sliki 67 je razviden rezultat podprocesa urejanja podatkov, tj. zagotavljanja višje kakovosti podatkov za nadaljnjo obdelavo.

Slika 67: Rezultat podprocesa urejanja podatkov v slovenskem jeziku



aleš cesar dobi medaljo izjemno hrabrost reševanju ljudi premoženja obkateri nevarnost izpostavil svoje življenje glasi utemelj
finančni minister franc križanič zahtevah naproti zato avtoprevoznikizadovoljni današnjih pogovorih povedal predesednik sekc
glede namigovanj današnjem dnevniku obrnite policijo odgovorili zurada generalne državne tožilke vprašanje zvezi stiki imela
gneča nastajala zaradi doslednega preverjanja vseh vizumov vsem ljudem vstopajo taprostor dejal dragutin mate zanikal ugit
gospoda jožeta anderliča nisem kazensko ovadil rubriki prejeli današnjem deluzapisal mestni svetnik miha jazbinšek jazbinšek s
izrekam priznanje vaše zahtevno odgovorno delo opravili najboljprijaznem okolju nagovoru dejal predsednik danilo türk izpost
nekateri starejši ljudje niso mogli pripeljati svojih domačih živali pregled seodločili omogočimo vzrok uvedbo prvega reševalneg
lizbonska strategija ključ uspešne konkurenčne evrope samo večjim poudarkom naizobraževanju mladini zasedanju sveta izob
tremi zamislimi bila ekonomsko okoljsko optimalna izbrana trasa ruške obvozniceki potekala proti tovarni geberit železniški prog
ministru miklavčiču pojasnil našo stisko protestno pismo napisali tudi borutupahorju kratkem pričakujemo odgovor drugače bor
pred vhodom destrniško pokopališče kupu zemlje človeške kosti česa tako groznega paše uredništvo poklical občan destrnika
prehitevanje desni nikoli obneslo potrdilo primeru sofijinega dvorca kije projekta rimskih term odprt silo gostje prihajajo razoča
pretirane odsotnosti želimo preprečiti preventivnim opozarjanjem staršev temko napišejo opravičilo naredijo usluge nikomur v
projektom kšokov popustnik želimo času gospodarske krize svojim članom omogočiti cenejšenakupe tudi zunaj razprodaj dejal
tale danes zame bolj slab poti imel prometno nesrečo zato moker davčni konferneci povedal ivan simič davčni strokovnjak nek
čista bedarija meni jernej sedmak predsednik potapljaškega društva manta club izola petekskupaj posadko ladjo pridržali pred

Vir: KNIME 2024.

Na slikah 68 in 69 je prikazan rezultat delotoka za izdelavo tematskih modelov, iz katerega so razvidni določeni tematski modeli, ključne besede, ki pripadajo posameznemu tematskemu modelu, izvirni zapisi ter prepoznane imenske entitete, poleg tega pa tudi uteži posameznih ključnih besed.

Slika 68: Rezultat delotoka izdelave tematskih modelov v slovenskem jeziku

Topic id	Concatenate(Term)	Content	Term
topic_0	sodišče, namreč, sodišča, uprave, sodišču, evrov, policisti, postopek	" Glede namigovanj v današnjem Dnevniku se obrnite na policijo," so za...	Dnevniku[ORG(SPACY_NE_sl_core_news_lg-3.7.0)];STA[ORG(SPACY_N
topic_0	sodišče, namreč, sodišča, uprave, sodišču, evrov, policisti, postopek	" Gospoda Jožeta Anderliča nisem kazensko ovadil," je v rubriki Prejeli s...	Jožeta Anderliča[PER(SPACY_NE_sl_core_news_lg-3.7.0)];Mlha Jazbinš
topic_0	sodišče, namreč, sodišča, uprave, sodišču, evrov, policisti, postopek	"15. novembra sem od Iva Matoše iz Celja za 300 evrov brez računa k...	Iva Matoše[PER(SPACY_NE_sl_core_news_lg-3.7.0)];Celja[LOC(SPACY
topic_0	sodišče, namreč, sodišča, uprave, sodišču, evrov, policisti, postopek	"Ah, pa ona (Suzaana Gale, op. p.) ni nič kriva, to je bilo vse malo čud...	Suzaana Gale[PER(SPACY_NE_sl_core_news_lg-3.7.0)];Žurnal24[ORG(
topic_0	sodišče, namreč, sodišča, uprave, sodišču, evrov, policisti, postopek	"Andrej Rezar je vedel, da je podjetje Mebles prezadolženo in ima nepr...	Andrej Rezar[PER(SPACY_NE_sl_core_news_lg-3.7.0)];Mebles[PER(SP
topic_0	sodišče, namreč, sodišča, uprave, sodišču, evrov, policisti, postopek	"Andrej Verlič je kot podžupan izrabil položaj in posredoval pri direktorj...	Andrej Verlič[PER(SPACY_NE_sl_core_news_lg-3.7.0)];Zdravstvenega
topic_0	sodišče, namreč, sodišča, uprave, sodišču, evrov, policisti, postopek	"Andrej Vizjak je naš minister za gospodarstvo. Slišali smo, da ni želel p...	Andrej Vizjak[PER(SPACY_NE_sl_core_news_lg-3.7.0)];Janez Janša[PE
topic_0	sodišče, namreč, sodišča, uprave, sodišču, evrov, policisti, postopek	"Anonimna ovadba zoper ministra za infrastrukturo in prostor Zvoneta ...	Anonimna[ORG(SPACY_NE_sl_core_news_lg-3.7.0)];Zvoneta Černača[
topic_0	sodišče, namreč, sodišča, uprave, sodišču, evrov, policisti, postopek	"Anonimnih pisem v Uradu predsednika Republike Slovenije ne komentir...	Anonimnih pisem[MISC(SPACY_NE_sl_core_news_lg-3.7.0)];Uradu[ORG
topic_0	sodišče, namreč, sodišča, uprave, sodišču, evrov, policisti, postopek	"Anonimno ovadbo smo zaradi vsebinskih pomanjkljivosti, ki niso omogo...	SDT[ORG(SPACY_NE_sl_core_news_lg-3.7.0)];Zvoneta Černača[PER(S
topic_0	sodišče, namreč, sodišča, uprave, sodišču, evrov, policisti, postopek	"Avstrijski strokovnjak, ki smo ga prosili za strokovno mnenje, se je izlo...	Zdravstvene zbornice Slovenije[ORG(SPACY_NE_sl_core_news_lg-3.7.(
topic_0	sodišče, namreč, sodišča, uprave, sodišču, evrov, policisti, postopek	"Avto smo peljali na servis v Beljak, in ko smo se vračali, smo na avtoc...	Beljak[LOC(SPACY_NE_sl_core_news_lg-3.7.0)];Dumitru Alexandru[PE
topic_0	sodišče, namreč, sodišča, uprave, sodišču, evrov, policisti, postopek	"Banki Raiffeisen se ne nameravamo opravičiti, saj smo javnosti le posr...	Banki Raiffeisen[ORG(SPACY_NE_sl_core_news_lg-3.7.0)];Združenja g
topic_0	sodišče, namreč, sodišča, uprave, sodišču, evrov, policisti, postopek	"Barbara Verdnik je imela sklenjene pogodbe s kar 17 novinarji, ni pa ji...	Barbara Verdnik[PER(SPACY_NE_sl_core_news_lg-3.7.0)];Primorske no
topic_0	sodišče, namreč, sodišča, uprave, sodišču, evrov, policisti, postopek	"Begić je po pogovoru z delavcem podjetja na javni povrĺ ˇini fotografir...	Begić[PER(SPACY_NE_sl_core_news_lg-3.7.0)];Rem[PER(SPACY_NE_sl
topic_0	sodišče, namreč, sodišča, uprave, sodišču, evrov, policisti, postopek	"Bil sem dober gospodar. In če bi še vedno opravljal vodstveno funkcij...	Stanovanjskega sklada RS[ORG(SPACY_NE_sl_core_news_lg-3.7.0)];E
topic_0	sodišče, namreč, sodišča, uprave, sodišču, evrov, policisti, postopek	"Bil sem na policijski postaji v Šiški in sem vlagal ovadbo. Medtem me je ...	Šiški[LOC(SPACY_NE_sl_core_news_lg-3.7.0)];Hilda Tovšak[PER(SPAC
topic_0	sodišče, namreč, sodišča, uprave, sodišču, evrov, policisti, postopek	"Bil sem v zmoti in Matjaž Ivartnik je grdo izkoristil mene in moje uslužb...	Matjaž Ivartnik[PER(SPACY_NE_sl_core_news_lg-3.7.0)];Matic Tasič[PE
topic_0	sodišče, namreč, sodišča, uprave, sodišču, evrov, policisti, postopek	"Bil sem zaslišan na policiji. Tega nisem storil. Solzivec lahko dam v anali...	Solzivec[PER(SPACY_NE_sl_core_news_lg-3.7.0)];Zgornjem Dupleku[LO

Vir: KNIME 2024.

Slika 69: Pregled uteži posameznih ključnih besed v slovenskem jeziku glede na posamezen tematski model

Row ID	Topic id	Term	Weight
Row0	topic_0	sodišče	3,963
Row1	topic_0	namreč	2,025
Row2	topic_0	sodišča	1,744
Row3	topic_0	uprave	1,718
Row4	topic_0	sodišču	1,612
Row5	topic_0	evrov	1,472
Row6	topic_0	policisti	1,449
Row7	topic_0	postopek	1,391
Row8	topic_1	predsednik	5,660
Row9	topic_1	vlade	5,040
Row10	topic_1	vlada	4,268
Row11	topic_1	minister	3,848
Row12	topic_1	slovenije	3,103
Row13	topic_1	zakona	2,968
Row14	topic_1	stranke	2,936
Row15	topic_1	slovenija	2,878
Row16	topic_2	občine	3,142
Row17	topic_2	občina	2,897
Row18	topic_2	ljubljana	2,871
Row19	topic_2	občini	2,842
Row20	topic_2	evrov	2,788
Row21	topic_2	župan	2,581
Row22	topic_2	promet	2,330
Row23	topic_2	tisoč	2,246
Row24	topic_3	evrov	15,175
Row25	topic_3	odstotkov	5,223
Row26	topic_3	odstotka	4,568
Row27	topic_3	milijonov	4,073
Row28	topic_3	tisoč	3,995
Row29	topic_3	milijona	3,615
Row30	topic_3	evra	2,971
Row31	topic_3	družbe	2,684
Row32	topic_4	sloveniji	2,850
Row33	topic_4	ljudi	1,957
Row34	topic_4	otrok	1,839
Row35	topic_4	slovenije	1,823
Row36	topic_4	stopinj	1,396
Row37	topic_4	temperature	1,302
Row38	topic_4	imajo	1,149
Row39	topic_4	vreme	1,107

Vir: KNIME 2024.

Na sliki 70 je prikazano število izvirnih zapisov, ki pripadajo določenim identificiranim tematskim modelom.

Slika 70: Število izvirnih zapisov v slovenskem jeziku, ki pripadajo določenim identificiranim tematskim modelom

topic_0	7502
topic_1	12276
topic_2	9369
topic_3	10848
topic_4	8869

Vir: KNIME 2024.

Predhodno je bil izpostavljen proces urejanja podatkov predvsem z namenom zagotavljanja višje kakovosti podatkov za izdelavo tematskih modelov. Na podlagi urejenosti podatkov je možno doseči boljše rezultate in učinkovitejšo izvedbo delotoka izdelave tematskih modelov.

Iz slik 71 in 72 je razvidna razlika v primeru izvedenega podprocesa urejanja podatkov in brez izvedenega podprocesa urejanja podatkov.

Slika 71: Izdelani tematski modeli v slovenskem jeziku brez predhodno izvedenega podprocesa urejanja podatkov

topic_0	,, ., ", (, več, še, že,)
topic_1	,, ., evrov, (,), še, ...
topic_2	,, ., ", še, že, (, sodi...
topic_3	,, ., ", še, že, (,), -
topic_4	,, ., ", še, vlade, že,...

Vir: KNIME 2024.

Slika 72: Izdelani tematski modeli v slovenskem jeziku s predhodno izvedenim podprocesom urejanja podatkov

topic_0	sodišče, namreč, sodišča, uprave, sodišču, evrov, policisti, postopek
topic_1	predsednik, vlade, vlada, minister, slovenije, zakona, stranke, slovenija
topic_2	občine, občina, ljubljana, občini, evrov, župan, promet, tisoč
topic_3	evrov, odstotkov, odstotka, milijonov, tisoč, milijona, evra, družbe
topic_4	sloveniji, ljudi, otrok, slovenije, stopinj, temperature, imajo, vreme

Vir: KNIME 2024.

Iz slike 72 je razvidno, da podproces urejanja podatkov ključno vpliva na učinkovitost in preglednost izdelanih tematskih modelov, saj je vsebinska opredelitev možna že na podlagi identificiranih ključnih besed, ki se nanašajo na posamezen izdelan tematski model. Obseg tematskih modelov se lahko poljubno spremeni glede na pridobljene rezultate. Za nadaljnjo obdelavo podatkov je bil obseg petih tematskih modelov ustrezno določen in tudi uporabljen.

Ker je bil namen razumeti predhodno neznane zapise v nestrukturirani obliki v njihovi kategorični obliki, je bil z izdelavo tematskih modelov ta cilj dosežen.

Za razliko od delotoka, kjer je bil izveden podproces urejanja vsebine, je rezultat, kjer omenjeni podproces ni bil izveden, vidno slabši oziroma ima kakovost podatkov velik vpliv na učinkovito izdelavo tematskih modelov. Iz slike 71 je razvidno, da je bilo brez izvedbe podprocesa urejanja podatkov v izdelavo tematskih modelov zajetih preveč nepotrebnih podatkov, zaradi česar je tudi rezultat primerljivo mnogo slabši.

Izdelava modela za samostojno klasifikacijo vsebine

Drugi delotok v postopku prepoznave vsebine in klasifikacije zapisov je vključeval postopek za izdelavo treh različnih odločitvenih modelov. Dva od odločitvenih modelov sta bila izdelana neposredno iz zapisov, ki so bili del izdelave tematskih modelov, eden od modelov pa je bil izdelan na osnovi predizdelanega modela, vključno s podrobnejšim izboljševanjem modela (angl. Fine tuning) na podlagi klasificiranih izvirnih zapisov.

Zaradi primerjanja učinkovitosti izvedbe klasifikacije izvirnih zapisov so bili izvedeni trije različni pristopi:

- uporaba metode odločitvenih dreves (Karabadji idr. 2017),
- uporaba metode podpornih vektorjev (Cervantes idr. 2020) in
- uporaba arhitekture in modela BERT (2023).

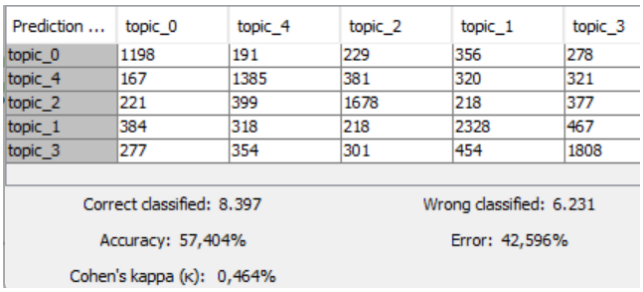
Pri vseh metodah je bil dodatno izveden podproces urejanja podatkov, pri uporabi arhitekture in modela BERT (2023) je bil proces vzporedno izveden tudi brez izvedbe podprocesa urejanja podatkov.

Vhodni podatki za izvedbo delotoka so bili podatki, kot so bili pripravljeni v okviru delotoka izdelave predhodnih delotokov, ki so vključevali izvirne zapise, identificirane imenske entitete in določene tematske modele kot obliko klasifikacije.

V postopku klasifikacije je bilo pri vseh izdelanih modelih uporabljeno enako razmerje razdelitve izvirnih zapisov za učenje in preverjanje pravilnosti izvedene klasifikacije. Obseg izvirnih zapisov je bil razdeljen na 70 % vseh zapisov za izdelavo odločitvenega modela in 30 % vseh zapisov za izvedbo klasifikacije in preverjanja pravilnosti izvedene klasifikacije.

Slika 73 prikazuje rezultat in pravilnost izvedene klasifikacije na podlagi modela, ki je bil izdelan z metodo odločitvenih dreves (angl. Decision tree learning) (Karabadji idr. 2017).

Slika 73: Rezultat in pravilnost izvedene klasifikacije v slovenskem jeziku na podlagi modela, ki je bil izdelan z metodo odločitvenih dreves



Prediction ...	topic_0	topic_4	topic_2	topic_1	topic_3
topic_0	1198	191	229	356	278
topic_4	167	1385	381	320	321
topic_2	221	399	1678	218	377
topic_1	384	318	218	2328	467
topic_3	277	354	301	454	1808

Correct classified: 8.397 Wrong classified: 6.231

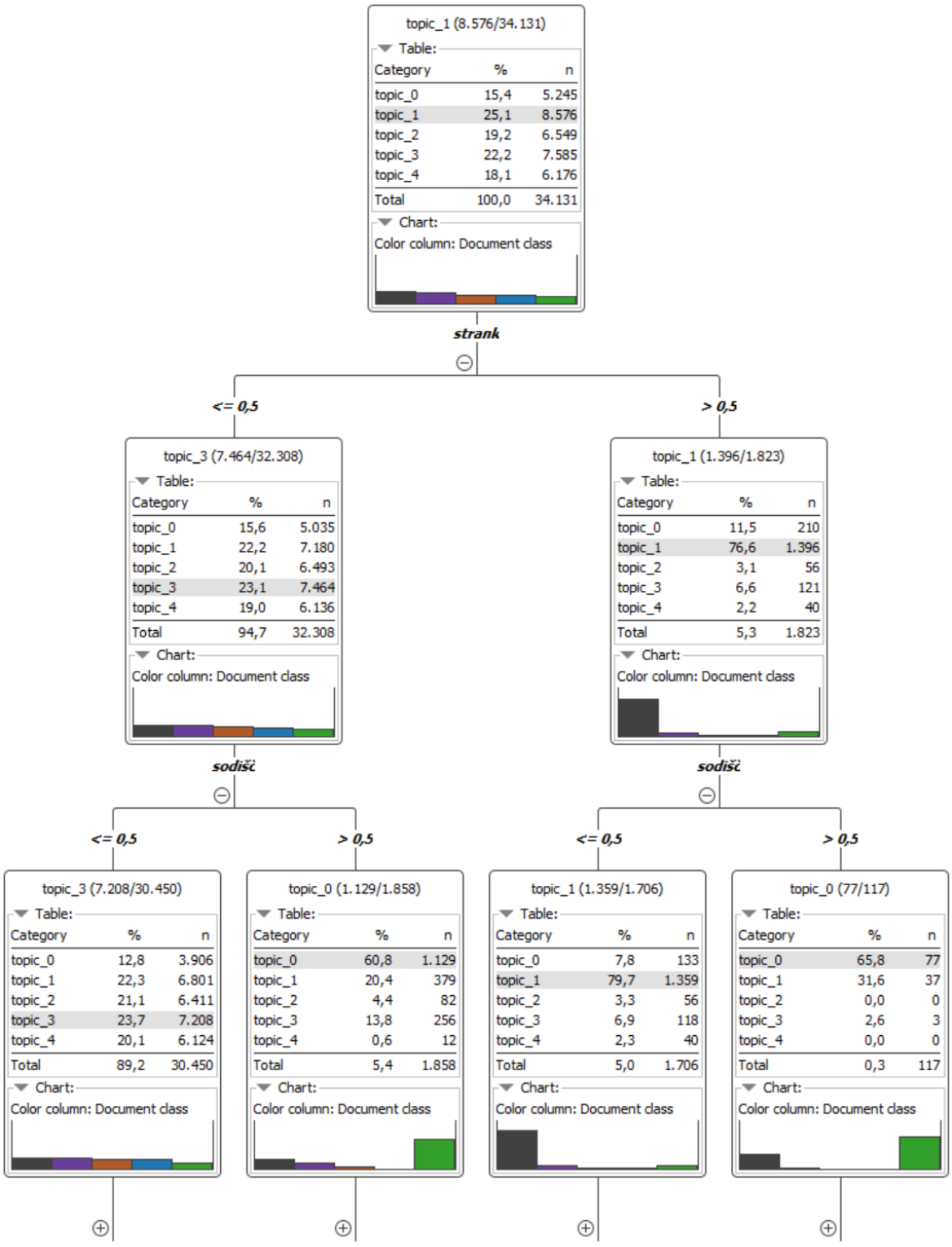
Accuracy: 57,404% Error: 42,596%

Cohen's kappa (κ): 0,464%

Vir: KNIME 2024.

Slika 74 prikazuje prerez in hierarhijo izvedene klasifikacije na podlagi modela metode odločitvenih dreves (Karabadji idr. 2017).

Slika 74: Prikaz hierarhije odločitvenega modela metode odločitvenih dreves v slovenskem jeziku



Vir: KNIME 2024.

Slika 75 prikazuje rezultat in pravilnost izvedene klasifikacije na podlagi modela, ki je bil izdelan z metodo podpornih vektorjev (angl. Support vector machine − SVM) (Cervantes idr. 2020).

Slika 75: Rezultat in pravilnost izvedene klasifikacije v slovenskem jeziku na podlagi modela, ki je bil izdelan z metodo SVM

Prediction ...	topic_0	topic_1	topic_4	topic_2	topic_3
topic_0	1566	157	79	130	125
topic_1	244	3038	186	109	285
topic_4	126	165	1956	268	242
topic_2	149	103	229	2113	196
topic_3	162	213	197	187	2403

Correct classified: 11.076 Wrong classified: 3.552

Accuracy: 75,718% Error: 24,282%

Cohen's kappa (κ): 0,694%

Vir: KNIME 2024.

Slika 76 prikazuje rezultat in pravilnost izvedene klasifikacije na podlagi modela, ki je bil izdelan z arhitekturo in modelom BERT (2023), brez izvedbe podprocesa urejanja podatkov.

Slika 76: Rezultat in pravilnost izvedene klasifikacije v slovenskem jeziku na podlagi modela, ki je bil izdelan z arhitekturo in metodo BERT brez izvedbe podprocesa urejanja podatkov

Prediction ...	topic_0	topic_2	topic_1	topic_3	topic_4
topic_0	1708	127	161	109	81
topic_2	93	2008	45	73	158
topic_1	228	87	3121	115	113
topic_3	149	320	323	2845	204
topic_4	60	225	73	100	2134

Correct classified: 11.816 Wrong classified: 2.844

Accuracy: 80,6% Error: 19,4%

Cohen's kappa (κ): 0,755%

Vir: KNIME 2024.

Slike 73, 75 in 76 prikazujejo rezultat izvedbe delotoka klasifikacije vsebine v slovenskem jeziku. Z izjemo uporabe drugačnega predizdelanega večjezičnega modela v primeru pristopa arhitekture in modela BERT (2023) so rezultati glede na natančnost podobni, kot so bili v primeru obdelave izvirnih zapisov v angleškem jeziku.

Natančnost klasifikacije pri obdelavi izvirnih zapisov v slovenskem jeziku si sledi v naslednjem vrstnem redu od najnatančnejšega odločitvenega modela do najmanj natančnega:

- 81 % natančnost – uporaba arhitekture in modela BERT (2023),
- 76 % natančnost – uporaba metode podpornih vektorjev – SVM (Cervantes idr. 2020),
- 57 % natančnost – uporaba metode odločitvenih dreves (Karabadji idr. 2017).

Podobno kot v primeru obdelave izvirnih zapisov v angleškem jeziku je možno tudi v primeru obdelave izvirnih zapisov v slovenskem jeziku pripisati boljše rezultate pristopu z uporabo predizdelanega modela na podlagi tega, da je bila osnova za izdelavo odločitvenega modela ravno tako predizdelan model, ki je bil treniran/učen na večjem obsegu podatkov.

Iz pridobljenih rezultatov obdelave izvirnih zapisov v slovenskem jeziku na podlagi izvedbe delotoka in korakov, kot so bili uporabljeni pri obdelavi izvirnih zapisov v angleškem jeziku, lahko zaključimo, da je izvedba predvidene oblike delotoka in povezane aktivnosti učinkovita tudi v primeru izvirnih besedil, ki so v različnih jezikih.

3.4.6 Delotok za izdelavo naslova popisne enote v slovenskem jeziku

Tretja faza izdelava popisne enote v slovenskem jeziku je enako kot v primeru obdelave izvirnih zapisov v angleškem jeziku vključevala izdelavo naslova popisne enote in dodajanje ostalih fiksnih podatkov skladno za zapolnitev sheme po standardu ISAD(G) (2000).

Na sliki 77 je prikazan rezultat izdelave naslova popisne enote v slovenskem jeziku z uporabo pristopa abstraktnega povzemanja.

Slika 77: Rezultat izdelave naslova popisne enote v slovenskem jeziku

Title	Classification	Content	Entity
Piranski župan Peter Bossman: Življenje ob morju privilegij, življenje ob morju privilegij	topic_4	„Če še niste opazili, tudi jaz sem „forešt" (tujec), se pa strinjam, da je življenje...	Piran[LOC(SPACY_NE_sl_core_news_lg-3.7.0)];Peter Bossman[PER(SPACY_NE
Raziskava: Raziskava o zlorabi drog med mladino ne kaže večjih odstopanj med mladost...	topic_4	„Zadnjič sem se ob enih zjutraj sprehodil čez prizorišče v BTC-ju, kjer se je tisti...	BTC-ju[ORG(SPACY_NE_sl_core_news_lg-3.7.0)];Festival mladih[MISC(SPACY_
Kampos: Televizijske oddaje o Slovencih v Italiji bodo kršene (dopolnjeno)	topic_4	„Vladna prepoved sklepanja avtorskih in svetovalnih pogodb bo močno prizade...	TV Koper[ORG(SPACY_NE_sl_core_news_lg-3.7.0)];Primorsko kroniko[LOC(SPA
Vreme: Ponoči bo nekaj ploh in neviht in nevihti (dopolnjeno)	topic_4	„V ponedeljek bo nekaj ploh in nevihti, v torek bo sončno, nato pa v sredo pop...	Jure Cedilnik[PER(SPACY_NE_sl_core_news_lg-3.7.0)];Agencije Republike Slov
Lenarčič: Prelet Atlantika po pristanku na Irsko (dopolnjeno)	topic_4	„Slabo vreme po pristanku na letališču Gander je narekovalo, da bo njegov čez...	letališču Gander[ORG(SPACY_NE_sl_core_news_lg-3.7.0)];Matevž[PER(SPACY
Žurnal24: Iz klavnice TMI Košaki pobegnil velik bik (dopolnjeno)	topic_4	„Saj nisem verjel, da prav vidim, ko je mimo mene med vožnjo stekel bik," nas j...	Maribora[LOC(SPACY_NE_sl_core_news_lg-3.7.0)];Tovarne mesnih izdelkov[O
Žurnalovi bralci navdušeni nad uspehom Baumgartnerja (dopolnjeno)	topic_4	„Ne razumem fascinacije z Baumgartnerjem, sploh pa ne primerjav z dosežki N...	Baumgartnerjem[PER(SPACY_NE_sl_core_news_lg-3.7.0)];Nase[ORG(SPACY_
Mori na Belvederju nad Izolo priseli 20 odstotkov pridelka (dopolnjeno)	topic_4	„Naključni mimoidoči in tudi ne ravno naključni veselo in brez sramu pobirajo pri...	Boris Mori[PER(SPACY_NE_sl_core_news_lg-3.7.0)];Belvederju[LOC(SPACY_NE

Vir: KNIME 2024.

V zadnjem delu delotoka izdelave popisne enote v slovenskem jeziku so bili dodani fiksni podatki (glej sliko 78), ki so skupaj s podatki, ki so bili ustvarjeni v predhodnih delotokih, vključno z izdelanimi naslovi popisne enote, tvorili končno obliko popisne enote.

Slika 78: Dodajanje fiksnih podatkov za izdelavo končne oblike popisne enote v slovenskem jeziku

Refere...	Accumu...	LevelOfDe...	ExtentAnd...
SI AM 00	01.maj.2022	Individualni zapis	1 elektronski zapis
SI AM 01	01.maj.2022	Individualni zapis	1 elektronski zapis
SI AM 02	01.maj.2022	Individualni zapis	1 elektronski zapis
SI AM 03	01.maj.2022	Individualni zapis	1 elektronski zapis
SI AM 04	01.maj.2022	Individualni zapis	1 elektronski zapis
SI AM 05	01.maj.2022	Individualni zapis	1 elektronski zapis

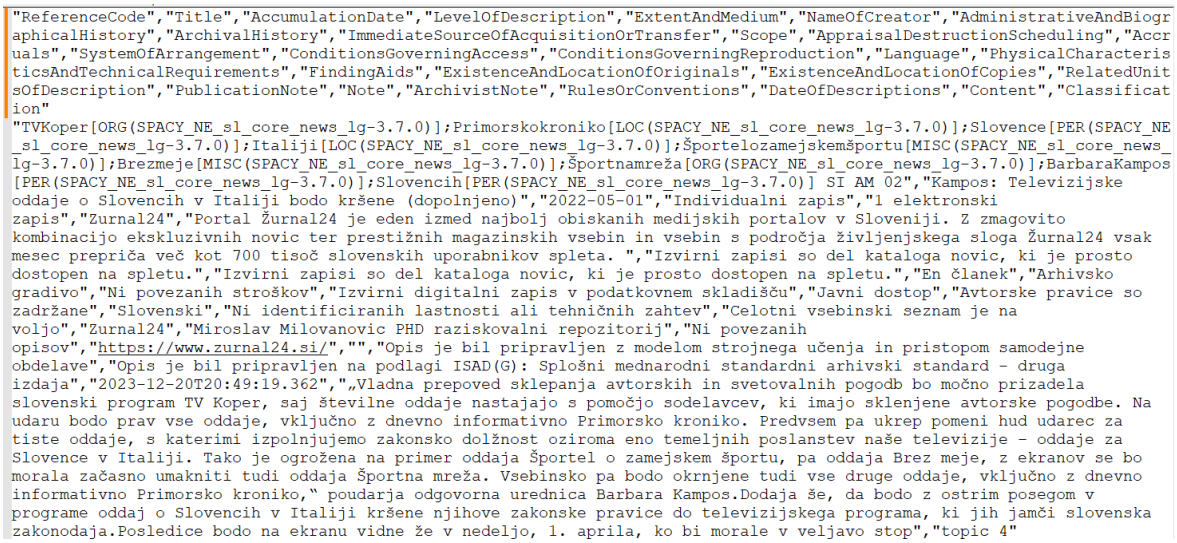
Vir: KNIME 2024.

Izhodna vrednost delotoka za izdelavo naslova popisne enote v slovenskem jeziku, ki je hkrati predstavljala tudi končno oziroma izhodno vrednost vseh delotokov, je tekstovna datoteka, ki je enako kot v primeru obdelave izvirnih zapisov v angleškem jeziku vključevala naslednje podatke:

- individualen izvirni zapis,
- prepoznane imenske entitete, nanašajoče se na individualen izvirni zapis,
- določeno klasifikacijo na podlagi izdelanih tematskih modelov – nanašajoč se na individualen izvirni zapis,
- izdelan naslov popisne enote – nanašajoč se na individualen izvirni zapis,
- fiksne podatke za dopolnitev sheme ISAD(G) (2000) – z izjemo oznake popisne enote, enako za vse izvirne zapise.

Na sliki 79 je prikazan individualen zapis v slovenskem jeziku v izvoženi tekstovni datoteki, ki vključuje vse podatke za izvedbo predogleda popisne enote. V prvi vrstici so predstavljeni nazivi popisnih elementov, v nadaljnjih vrsticah pa so zajeti izvirni zapisi s pripadajočimi rezultati, izdelanimi z izvedbenimi aplikativnimi modeli.

Slika 79: Prikaz individualnega zapisa v slovenskem jeziku v izvoženi tekstovni datoteki, ki vključuje vse podatke za izvedbo predogleda popisne enote



"ReferenceCode","Title","AccumulationDate","LevelOfDescription","ExtentAndMedium","NameOfCreator","AdministrativeAndBiographicalHistory","ArchivalHistory","ImmediateSourceOfAcquisitionOrTransfer","Scope","AppraisalDestructionScheduling","Accruals","SystemOfArrangement","ConditionsGoverningAccess","ConditionsGoverningReproduction","Language","PhysicalCharacteristicsAndTechnicalRequirements","FindingAids","ExistenceAndLocationOfOriginals","ExistenceAndLocationOfCopies","RelatedUnitsOfDescription","PublicationNote","Note","ArchivistNote","RulesOrConventions","DateOfDescriptions","Content","Classification"

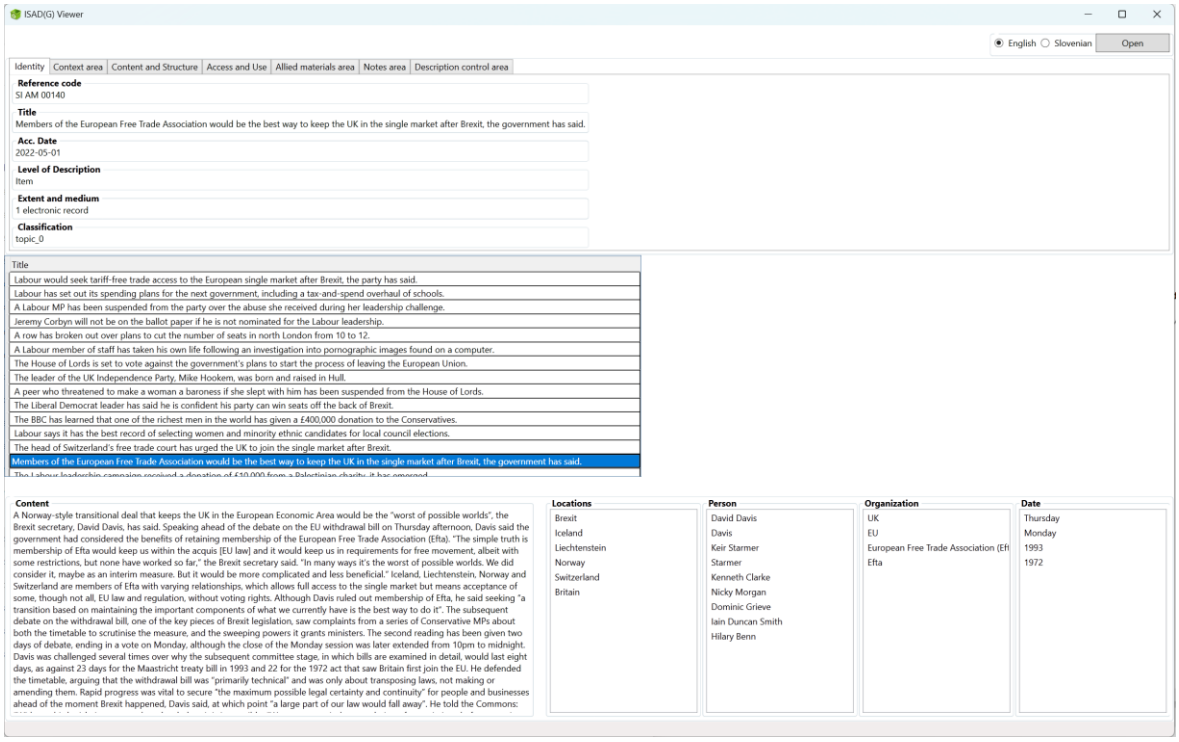
"TVKoper[ORG(SPACY_NE_sl_core_news_lg-3.7.0)];Primorskokroniko[LOC(SPACY_NE_sl_core_news_lg-3.7.0)];Slovence[PER(SPACY_NE_sl_core_news_lg-3.7.0)];Italiji[LOC(SPACY_NE_sl_core_news_lg-3.7.0)];Športelozamejskemšportu[MISC(SPACY_NE_sl_core_news_lg-3.7.0)];Brezmeje[MISC(SPACY_NE_sl_core_news_lg-3.7.0)];Športnamreža[ORG(SPACY_NE_sl_core_news_lg-3.7.0)];BarbaraKampos[PER(SPACY_NE_sl_core_news_lg-3.7.0)];Slovencih[PER(SPACY_NE_sl_core_news_lg-3.7.0)] SI AM 02","Kampos: Televizijske oddaje o Slovencih v Italiji bodo kršene (dopolnjeno)","2022-05-01","Individualni zapis","1 elektronski zapis","Zurnal24","Portal Žurnal24 je eden izmed najbolj obiskanih medijskih portalov v Sloveniji. Z zmagovito kombinacijo ekskluzivnih novic ter prestižnih magazinskih vsebin in vsebin s področja življenjskega sloga Žurnal24 vsak mesec prepriča več kot 700 tisoč slovenskih uporabnikov spleta. ","Izvirni zapisi so del kataloga novic, ki je prosto dostopen na spletu.","Izvirni zapisi so del kataloga novic, ki je prosto dostopen na spletu.","En članek","Arhivsko gradivo","Ni povezanih stroškov","Izvirni digitalni zapis v podatkovnem skladišču","Javni dostop","Avtorske pravice so zadržane","Slovenski","Ni identificiranih lastnosti ali tehničnih zahtev","Celotni vsebinski seznam je na voljo","Zurnal24","Miroslav Milovanovic PHD raziskovalni repozitorij","Ni povezanih opisov","https://www.zurnal24.si/","","Opis je bil pripravljen z modelom strojnega učenja in pristopom samodejne obdelave","Opis je bil pripravljen na podlagi ISAD(G): Splošni mednarodni standardni arhivski standard - druga izdaja","2023-12-20T20:49:19.362","„Vladna prepoved sklepanja avtorskih in svetovalnih pogodb bo močno prizadela slovenski program TV Koper, saj številne oddaje nastajajo s pomočjo sodelavcev, ki imajo sklenjene avtorske pogodbe. Na udaru bodo prav vse oddaje, vključno z dnevno informativno Primorsko kroniko. Predvsem pa ukrep pomeni hud udarec za tiste oddaje, s katerimi izpolnjujemo zakonsko dolžnost oziroma eno temeljnih poslanstev naše televizije – oddaje za Slovence v Italiji. Tako je ogrožena na primer oddaja Športel o zamejskem športu, pa oddaja Brez meje, z ekranov se bo morala začasno umakniti tudi oddaja Športna mreža. Vsebinsko pa bodo okrnjene tudi vse druge oddaje, vključno z dnevno informativno Primorsko kroniko,“ poudarja odgovorna urednica Barbara Kampos.Dodaja še, da bodo z ostrim posegom v programe oddaj o Slovencih v Italiji kršene njihove zakonske pravice do televizijskega programa, ki jih jamči slovenska zakonodaja.Posledice bodo na ekranu vidne že v nedeljo, 1. aprila, ko bi morale v veljavo stop","topic 4"

Vir: Lastni 2024.

3.4.7 Pregledovalnik popisnih enot v angleškem in slovenskem jeziku

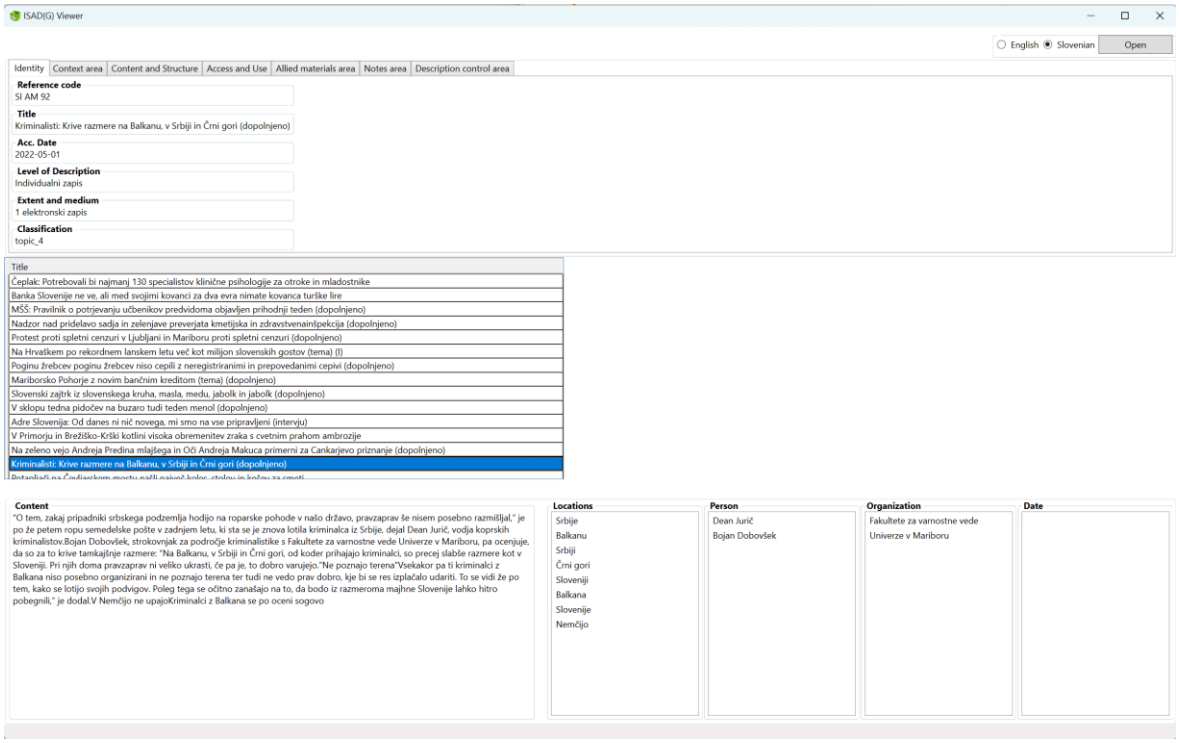
Na slikah 80 in 81 je prikazano aplikativno grafično okolje za pregledavanje zapisov izdelanih popisnih enot. Vhodna vrednost je tekstovna datoteka, ki je bila izvožena iz zadnjega delotoka procesa izdelave popisne enote. Grafični vmesnik je bil razvit kot namizna programska oprema, ki ne potrebuje posebnega namestitvenega postopka.

Slika 80: Prikaz aplikativnega grafičnega okolja za ogled zapisov izdelanih popisnih enot v angleškem jeziku



Vir: Lastni 2024.

Slika 81: Prikaz aplikativnega grafičnega okolja za ogled zapisov izdelanih popisnih enot v slovenskem jeziku



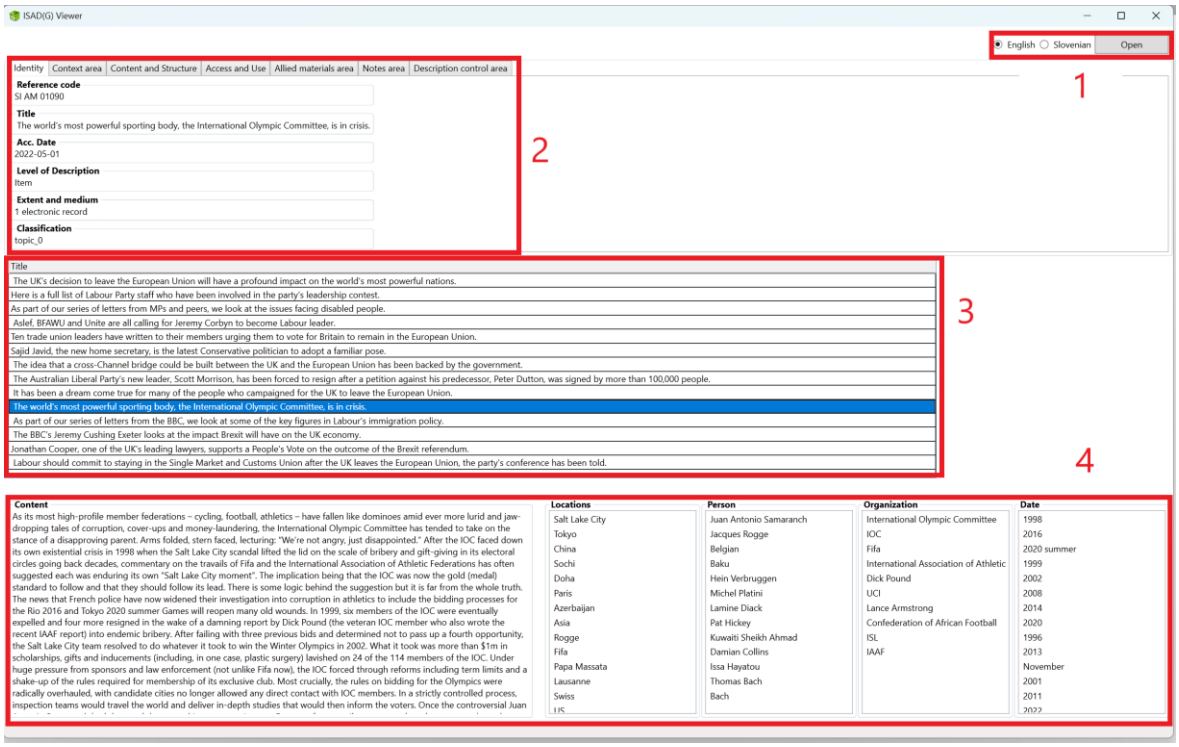
Vir: Lastni 2024.

Pregledovalnik je sestavljen iz štirih sekcij, ki so namenjene izbiri izdelanih popisnih enot in izvedbi predogleda posamezne popisne enote (glej sliko 82):

- sekcija 1 prikazuje možnost izbire jezika izdelanih popisnih enot in funkcijo izbire tekstovne datoteke, kjer se nahajajo zapisi. Pri tem je izbira tekstovne datoteke omejena na obliko, kot je bila ustvarjena v delu izvoza podatkov v zadnjem delotoku, ustvarjenem v programski opremi KNIME (2024);
- sekcija 2 prikazuje zavihke, kjer se nahajajo izdelani popisni elementi, ki so ločeni glede na skupine popisnih podatkov. Na prvem zavihku se neposredno nahaja tudi podatek o klasifikaciji zapisa. Vsi podatki v teh sekcijah (in vseh zavihkih) se prikažejo z izbiro individualnega naslova popisne enote;
- sekcija 3 prikazuje izdelane naslove popisnih enot in je stalno vidna v prikazovalniku. Z izbiro naslova se aktivira prikaz vseh podatkov individualne popisne enote v vseh poljih za predogled individualnega zapisa;
- sekcija 4 prikazuje vsebino individualnega zapisa in prepoznane imenske entitete.

Namen ločevanja sekcij je bil predvsem doseči čim boljši pregled nad pridobljenimi rezultati.

Slika 82: Prikaz sekcij grafičnega okolja za ogled zapisov izdelanih popisnih enot v angleškem jeziku



Vir: Lastni 2024.

3.5 Razprava

Uvodoma je bilo izpostavljeno, da je bil cilj raziskave vzpostaviti delujoč teoretičen in aplikativen izvedbeni model za samodejno masovno urejanje in popisovanje nestrukturiranih besedil ter skušati odgovoriti na naslednja vprašanja:

- Kako različna pojavna oblika posameznih digitalnih zapisov vpliva na možnost vsebinske analize z uporabo strojnega učenja?
- Kako nestrukturiranost podatkov vpliva na možnost vsebinske analize z uporabo strojnega učenja?
- Kakšna je učinkovitost masovnega urejanja in popisovanja nestrukturiranih zapisov z uporabo strojnega učenja?

V povezavi z zgoraj zastavljenimi vprašanji so bila na osnovi ugotovitev in predvidenih odgovorov postavljena dodatna podvprašanja:

- Ali je v primeru nestrukturiranih besedil za namen popisovanja možno izdelati model samodejne klasifikacije vsebin z uporabo strojnega učenja?
- Ali je v primeru nestrukturiranih besedil za namen popisovanja možno izdelati model samodejne identifikacije imenskih entitet z uporabo strojnega učenja?
- Ali je v primeru nestrukturiranih besedil za namen popisovanja možno izdelati model izdelave naslova individualne popisne enote z uporabo strojnega učenja?

Urejanje in popisovanje nestrukturiranih besedil z uporabo strojnega učenja je izvedljivo, pomembna sta predvsem pristop in oblika obdelave predvidenih podatkov. Da bi dosegli dobre rezultate, je bilo treba poizkusiti različne pristope, saj je bilo ravno na podlagi povratnih informacij različnih individualnih aktivnosti potem lažje ugotoviti, kateri pristop in pod kakšnimi pogoji daje boljše rezultate.

V raziskavi smo ugotovili, da je izdelava strukturiranega pristopa, tj. modela urejanja in popisovanja nestrukturiranih besedil z uporabo strojnega učenja, ravno tako v celoti izvedljiva bodisi v obliki izvajanja aktivnosti za dosego le enega od ciljev, npr. identifikacija in dokumentiranost imenskih entitet, ali pa kot celosten pristop z doseganjem končnega cilja t. i. izdelave popisne enote.

Pojavna oblika posameznih digitalnih zapisov lahko vpliva na možnost vsebinske obdelave. Predvsem je omejitev izkazana v delu, da morajo biti podatki strojno berljivi, če je namen, da se obdelava izvede v digitalni obliki z uporabo strojnega učenja. Vhodno vrednost tako predstavljajo zapisi, ki so v digitalni obliki ali pa so bili pretvorjeni v digitalno obliko na način, da je možna kakovostna prepoznava besedila.

Poleg pojavne oblike je bila določena pozornost posvečena tudi obdelavi podatkov v primeru, da gre za različne jezike besedil, pri čemer se skuša zadržati unificiran pristop pri izvedbi vseh aktivnosti, tj. standardizaciji korakov, potrebnih za dosego ciljev. Iz rezultatov je razvidno, da je model urejanja in popisovanja nestrukturiranih besedil z uporabo strojnega

učenja stabilen in izvedljiv tudi v primeru večjezične vsebine. V določenih primerih je sicer treba vzpostaviti specifične vire za obdelavo podatkov, kot so npr. predizdelani modeli, nanašajoči se na specifično jezično področje, vendar pa je sam model urejanja in popisovanja nestrukturiranih besedil koherenten v delu izvajanj aktivnosti ne glede na jezik, ki je uporabljen v besedilu.

Preverjeno je bilo tudi, kako nestrukturiranost podatkov vpliva na možnost vsebinske analize z uporabo strojnega učenja. Ugotovljeno je bilo, da zgolj »nestrukturiranost« vidno ne vpliva na možnost obdelave podatkov oziroma na rezultate. Obdelava podatkov je namreč pokazala, da je bila arhitektura modela in povezanih aktivnosti usmerjena na način, da je bilo besedilo obdelano v svoji izvirni obliki tako na nivoju celote kot na nivoju individualnih enot besedila (besed), pri čemer ni bilo izpostavljenih omejitev za uporabo strojnega učenja.

V zadnjem sklopu prvega dela vprašanj je bilo preverjeno tudi, kakšna je učinkovitost masovnega urejanja in popisovanja nestrukturiranih zapisov z uporabo strojnega učenja. Pri preverjanju učinkovitosti je bilo osnovno izhodišče doseganje čim boljšega rezultata z zasledovanjem dveh možnih pristopov, in sicer z izdelavo modela urejanja in popisovanja nestrukturiranih besedil na način, da so bili v celoti izdelani individualni odločitveni modeli, in s pristopom, kjer je bila vključena izdelava modelov na podlagi predizdelanih modelov.

Doseganje čim boljših rezultatov za določen sklop nestrukturiranih besedil je odvisen predvsem od vzpostavitve pristopa iteracij, pri čemer se izvaja izboljševanje modela na podlagi pridobljenih povratnih informacij predhodnih iteracij in posledično na določanju parametrov končne oblike modela. Kot primer se tako lahko izpostavi npr. določanje števila tematskih modelov za določen sklop besedil, kjer se lahko na podlagi rezultatov iteracij določi parametre za učinkovitejšo obdelavo podatkov (angl. Fine tuning) ali npr. število ciklov in obseg podatkov za učenje modelov v primeru izdelave odločitvenih modelov za izvedbo klasifikacije vsebine itd. Izvedba iteracij je bila pomembna tudi z vidika optimizacij delotokov, saj so povratne informacije ponudile dovolj informacij za oceno nujnosti individualnih korakov in vpliva posameznih korakov na rezultate. Posledično se je zasledoval tudi cilj vzpostavitve t. i. vitkega oziroma optimiziranega modela.

Učinkovitost je bila merjena na način, da se je preverjalo točnost pridobljenih rezultatov glede na predvidene cilje. Pri tem se je zasledovalo razmerje, da je vsaj 75 % pridobljenih rezultatov sprejemljivih, ustreznih oziroma pravilnih. Glede na rezultate je bil prag uspešnosti in učinkovitosti v vseh treh glavnih ciljih dosežen, v določenih primerih pa tudi močno presežen. Ob tem je treba izpostaviti, da je bila učinkovitost lahko kvantitativno merjena le v delu, kjer je bila takšna meritev možna, npr. pri izvedbi klasifikacije vsebine ali pri prepoznavi imenskih entitet. Precej težje je bilo izvesti kvantitativno merjenje v delu, kjer takšno merjenje ni bilo smiselno ali izvedljivo, npr. pri izdelavi naslova popisne enote ali pri določanju obsega izdelanih tematskih modelov. Pri slednjih je kvalitativni pristop merjenja učinkovitosti odvisen predvsem od pridobivanja povratnih informacij v okviru iteracij obdelave podatkov in kriterijev (parametrih), ki jih določi izdelovalec modela.

Posledično se tudi rezultati lahko razlikujejo glede na pogled npr. arhivista, ki je zadolžen za popisovanje gradiva.

Skladno s pridobljenimi rezultati in postavljeno arhitekturo izdelave modela so bili v okviru raziskave doseženi cilji, ki so bili predstavljeni na začetku raziskave:

- vsebinski pregled oblik in možnosti klasifikacije vsebin nestrukturiranih besedil z uporabo strojnega učenja,
- vsebinski pregled oblik in možnosti pristopov k identifikaciji imenskih entitet z uporabo strojnega učenja,
- vsebinski pregled oblik in možnih pristopov k ekstrakciji informacij, potrebnih za izdelavo naslova individualne popisne enote,
- izdelava teoretičnega modela urejanja in popisovanja nestrukturiranih besedil z uporabo strojnega učenja, ki bo vključeval popis procesov, delotokov in posameznih aktivnosti, potrebnih za samostojno urejanje in popisovanje nestrukturiranih besedil,
- temeljit pregled, kateri dejavniki ključno vplivajo na kakovost pridobljenih rezultatov,
- razvoj izvedbenega aplikativnega modela za urejanje in popisovanje nestrukturiranih besedil z uporabo strojnega učenja,
- opredelitev specializiranih orodij ter navodil in priporočil za samostojno izdelavo aplikativnega modela za urejanje in popisovanje nestrukturiranih besedil z uporabo strojnega učenja.

Glede na opredeljeni raziskovalni problem, izpostavljena raziskovalna vprašanja in predvidene pristope k razvoju modela urejanja in popisovanja nestrukturiranih besedil z uporabo strojnega učenja so bile izpostavljene tri hipoteze:

H1: Izdelava modela samodejne klasifikacije vsebin z uporabo strojnega učenja v primeru prepoznavanja nestrukturiranih besedil je izvedljiva in uporabna za potrebe popisovanja individualnih zapisov.

H2: Izdelava modela samodejne identifikacije imenskih entitet z uporabo strojnega učenja v primeru prepoznavanja nestrukturiranih besedil je izvedljiva in uporabna za potrebe popisovanja individualnega zapisa.

H3: Izdelava modela samodejne izdelave naslova individualne popisne enote z uporabo strojnega učenja v primeru prepoznavanja nestrukturiranih besedil je izvedljiva in uporabna za potrebe popisovanja individualnega zapisa.

H1: Izdelava modela samodejne klasifikacije vsebin z uporabo strojnega učenja v primeru prepoznavanja nestrukturiranih besedil je izvedljiva in uporabna za potrebe popisovanja individualnih zapisov?

Raziskava je pokazala, da je izdelava modela in pristopa samodejne klasifikacije vsebin z uporabo strojnega učenja v primeru prepoznavanja nestrukturiranih besedil *izvedljiva in uporabna za potrebe popisovanja individualnih zapisov.*
Pri tem je treba izpostaviti, da sta učinkovitost in posledično uporabnost odvisni predvsem od izbranega pristopa in določanja ciljev s strani osebe, ki določa kriterije za obdelavo podatkov.

Pomemben vidik zagotavljanja učinkovitosti delovanja modela je tudi v tem, da ga je vedno možno dopolnjevati ali spreminjati v delu, ki se nanaša na vsebinsko prepoznavo nestrukturiranih besedil in normativov za izvedbo klasifikacije gradiva. Pri tem je priporočljiv pristop z uporabo iteracij in pridobivanjem povratnih informacij o delovanju modela in pričakovanih rezultatih ter prilagajanjem modela na podlagi zastavljenih ciljev.

Pri izvedbi klasifikacije nestrukturiranih besedil lahko na odločitev, kateri pristop in izvedbeni model predstavlja boljšo možnost, vpliva tudi dinamika potencialnega spreminjanja ali dopolnjevanja izvedbenega modela, kar je odvisno predvsem od vhodnih podatkov, ki so predmet obdelave ali kriterijev posameznega uporabnika, ki želi model uporabiti.

Določen kriterij, ki lahko vpliva na izbiro izvedbenega modela, je tudi, ali so na voljo predizdelani modeli, ki lahko olajšajo izvedbo določenih delov procesa obdelave podatkov, kot je npr. izločitev faze »urejanja podatkov«, če takšna obdelava ni potrebna.

H2: Izdelava modela samodejne identifikacije imenskih entitet z uporabo strojnega učenja v primeru prepoznavanja nestrukturiranih besedil je izvedljiva in uporabna za potrebe popisovanja individualnega zapisa.

Raziskava je pokazala, da je izdelava modela in pristopa identifikacije imenskih entitet z uporabo strojnega učenja v primeru prepoznavanja nestrukturiranih besedil *izvedljiva in uporabna za potrebe popisovanja individualnega zapisa*.

Pri izdelavi pristopa in modela prepoznavanja imenskih entitet je eden od ključnih kriterijev izbira obsega tipov imenskih entitet, ki naj bi bile prepoznane v okviru obdelave nestrukturiranih besedil, iz tega pa izhaja tudi nivo zahtevnosti izdelave modela ali uporabe predizdelanega modela.

V primeru izdelave lastnega odločitvenega modela za prepoznavo imenskih entitet je pomemben predvsem pristop za izdelavo takšnega modela, saj so s tem lahko povezane zahteve po zagotavljanju zadostnih virov za izdelavo modela posebej v delu, ko se potrebujejo ustrezne količine podatkov za izvedbo učenja.

Pomemben vidik doseganja ciljev prepoznavanja imenskih entitet je tudi zagotavljanje kakovosti podatkov, ki lahko vpliva na končen rezultat prepoznane vsebine in prepoznanih imenskih entitet.

H3: Izdelava modela samodejne izdelave naslova individualne popisne enote z uporabo strojnega učenja v primeru prepoznavanja nestrukturiranih besedil je izvedljiva in uporabna za potrebe popisovanja individualnega zapisa.

Raziskava je pokazala, da je izdelava modela in pristopa samodejne izdelave naslova individualne popisne enote z uporabo strojnega učenja v primeru prepoznavanja nestrukturiranih besedil *izvedljiva in uporabna za potrebe popisovanja individualnega zapisa.*

Izdelava pristopa in modela za samodejno izdelavo naslova individualne popisne enote je predstavljala poseben izziv predvsem z vidika merjenja učinkovitosti uporabe takšnega modela in uporabnosti za namen popisovanja gradiva. V tem delu je bilo težko izvesti kvantitativno meritev ustreznosti pridobljenih rezultatov, saj je bil rezultat takšne obdelave tudi v klasičnem smislu, tj. brez uporabe strojnega učenja, predvsem odvisen od individualne interpretacije vsebine s strani posameznega popisovalca (arhivista) in zato individualnega tolmačenja in izdelave naslova popisne enote, ki bi ustrezno odražala vsebino izvirnega zapisa.

Glede na izhodišče, da so predmet obdelave nestrukturiranih besedil obseg in oblika ter količina podatkov v številu, ki za obdelavo s strani posamezne osebe ne bi bilo racionalno izvedljivo, in da izdelani naslovi individualnih popisnih enot predstavljajo ustrezen in pravilen odraz vsebine gradiva, je bil cilj izdelave naslova popisne enote, ki je uporabna za namen popisovanja gradiva, dosežen. Tako kot pri ostalih obdelavah podatkov sta kvalitativni kriterij in doseženi rezultat odvisna predvsem od kriterijev osebe, ki določa parametre za izvedbo obdelave podatkov na podlagi izvedenih iteracij in pridobljenih povratnih informacij.

3.5.1 Gradniki modela za samodejno masovno urejanje in popisovanje nestrukturiranih besedil

V raziskavi so bili natančno predstavljeni vsi koraki, ki so bili uporabljeni za doseganje ciljev in preverjanje hipotez. Izdelani model je predstavljen skozi serijo gradnikov, ki tvorijo funkcionalno celoto za izvedbo vseh aktivnosti, ki so potrebne za izvedbo obdelave podatkov. Med gradnike se tako všteva:

- strojna oprema za obdelavo podatkov,
- programska oprema za obdelavo podatkov,
- definicije delotokov za izvedbo aktivnosti (procesne opredelitve),
- vhodni podatki oziroma izvirno gradivo, ki je predmet obdelave,
- pristopi za obdelavo podatkov,
- arhitekture in predizdelani modeli za strojno učenje,
- izhodni podatki, pripravljeni za nadaljnjo obdelavo, in
- predogled izdelanih popisnih enot.

Strojna oprema za obdelavo podatkov

Strojna oprema je bila v okviru raziskave izbrana in uporabljena na način, da po zmogljivosti predstavlja obliko povprečne delovne postaje srednjega cenovnega razreda. Namen raziskave je bil vzporedno ponuditi enostavno možnost reproduciranja pridobljenih rezultatov na način, ki ne terja poglobljenega znanja s področja informacijskih tehnologij. V okviru specifik strojne opreme je treba izpostaviti potencialno dodatno uporabo grafične kratice in uporabo grafičnih jeder za hitrejše izvajanje določenih aktivnosti pri obdelavi podatkov. Ustrezna strojna oprema tako predstavlja prvi pogoj in gradnik za izdelavo modela za samodejno masovno urejanje in popisovanje nestrukturiranih besedil.

Programska oprema za obdelavo podatkov

Izbira ustrezne programske opreme za izdelavo modela je predstavljala največji izziv pri snovanju celotnega sistema, saj je bilo osnovno izhodišče, da bi morali biti gradniki z izjemo strojne opreme, ki bi bili uporabljeni za izdelavo in uporabo modela, prosto dostopni za uporabo komurkoli, ki bi želel takšen model izdelati ali brez omejitev uporabiti. Specifičen izziv je predstavljala precejšnja razdrobljenost individualnih funkcionalnih enot obdelave podatkov, ki je terjala določen pristop z vidika konsolidacije zaradi lažjega izvajanja opravil in organiziranja izvajanja aktivnosti. Za osrednjo programsko rešitev je bila izbrana odprtokodna programska oprema za analizo podatkov KNIME (2024), ki je bila v prvi vrsti namenjena t. i. orkestraciji oziroma združevanju vseh individualnih programov, ki so omogočali izvedbo celotnega procesa obdelave podatkov. Ob tem je treba izpostaviti, da končna rešitev pri uporabi programske opreme predstavlja neposredno izvedljivo funkcionalnost obdelave podatkov prek uporabe grafičnega vmesnika na način, da sta bili v čim večji meri poenostavljeni uporaba in izvedba vseh procesov obdelave podatkov za končnega uporabnika. Celotna funkcionalnost je predstavljala integrirano funkcionalnost izbrane programske opreme kot tudi integrirano funkcionalnost v obliki individualnih programov ali skript, ki so bile podprto dodane v izvajanje procesa. Končni uporabnik pri uporabi delotokov v orodju KNIME (2024) tako ne potrebuje dodatnih programov ali drugih orodij za izvedbo vseh aktivnosti, povezanih z obdelavo podatkov. Treba je izpostaviti, da je za določeno izvajanje aktivnosti in uporabo namenskih programskih jezikov predhodno treba zagotoviti ustrezno izvajalno okolje za podporo izvajanju določenih programskih jezikov, kot je npr. ustvarjanje ustreznega okolja za izvajanje programske kode, napisane s programskim jezikom »python« (Python.org 2024), v primeru uporabe specializiranih jeder grafične kartice pa tudi programsko opremo in gonilnike, ki omogočajo uporabo takšne strojne opreme.

Med programsko opremo je treba všteti tudi razvoj oziroma izdelavo specifične programske opreme za pregled pripravljenih izhodnih podatkov, tj. popisne enote. Ob tem je bil znova upoštevan pristop, da se zagotovi čim enostavnejšo obliko programske opreme z uporabo grafičnega uporabniškega vmesnika, ki bi končnemu uporabniku omogočal enostavno premikanje med uvoženimi podatki s specifičnim poudarkom na vizualizaciji podatkov, ki

so bili osrednji del raziskave oziroma predogled klasifikacije identificiranih imenskih entitet in izdelanih naslovov popisnih enot.

Uporaba odprtokodne programske opreme za analizo podatkov, vključno z novorazvito programsko opremo za predogled izdelanih popisnih enot, tako predstavlja celovito rešitev za izdelavo in uporabo modela za samodejno masovno urejanje in popisovanje nestrukturiranih besedil ter predogled pridobljenih rezultatov uporabe takšnega modela.

Definicije delotokov za izvedbo aktivnosti (procesne opredelitve)

Zaradi obsežnosti raziskave v delu izdelave modela za samodejno masovno urejanje in popisovanje nestrukturiranih besedil in še bolj natančno v delu vseh aktivnosti, ki so potrebne za doseganje ciljev, ki bi predstavljali zadovoljive rezultate, je bilo treba izbrati pristop, ki bi dovolj pregledno in jasno prikazoval vse korake, ki so bili uporabljeni. Pri tem je bil izbran pristop z izdelavo delotokov, v okviru katerih so bili natančno dokumentirani vsi koraki, ki so bili potrebni za izdelavo in uporabo modela za samodejno masovno urejanje in popisovanje nestrukturiranih besedil. Zaradi precejšnje razvejanosti aktivnosti so bili delotoki razdeljeni v tri krovne delotoke, ki so zasledovali tri ločene, vendar povezljive cilje:

- izdelavo procesa za samostojno prepoznavo imenskih entitet,
- izdelavo procesa za samostojno klasifikacijo vsebin in
- izdelavo procesa za samostojno izdelavo naslova popisne enote.

Delotoki in povezane aktivnosti so bili izdelani na način, da so skozi zaporedje korakov zagotovili takšno obdelavo podatkov, ki zagotavlja preglednost vseh vhodnih oblik oziroma potrebnih parametrov ter potencialnih predizdelanih modelov in vseh izhodnih oblik, vključno z jasno opredelitvijo celotne aktivnosti in pričakovanih rezultatov. Dokumentiranost delotokov tako predstavlja najboljši pregled na vsemi koraki, ki so potrebni za reprodukcijo rezultatov in doseganje ciljev. V določenih krovnih delotokih so bili vključeni tudi t. i. podprocesi ali manjši delotoki, ki so predstavljali zaključeno celoto določenega obsega aktivnosti znotraj obstoječega procesa, predvsem z namenom ponuditi možnost opcijske uporabe takšnih podprocesov. Delotoki ne vključujejo komponent, ki so nespremenljive ali nedefinirane, vsak korak delotoka pa je jasno dokumentiran in vizualno predstavljen na način, da je vedno vidna umeščenost tega koraka in povezljivost s prejšnjimi ali naknadnimi koraki.

Delotoki so bili ravno tako pripravljeni na način, da se lahko kot celote s podatki ali brez njih izvozijo iz uporabljene odprtokodne programske opreme in uporabijo v enaki obliki v enaki programski opremi na drugih delovnih postajah ali informacijskih okoljih. Skozi raziskavo je bila izvedena jasna dokumentiranost posameznih delotokov in povezanih aktivnosti z namenom izdelave modela kot specifikacije korakov za izvedbo vseh potrebnih aktivnosti za doseganje zadanih ciljev, zato lahko končni uporabnik uporabi tudi drugo programsko opremo in na podlagi izdelanega modela izgradi lasten model za samodejno masovno urejanje in popisovanje nestrukturiranih besedil oziroma razvije svojo lastno rešitev, ki bo vključevala dokumentirane korake izdelanega modela v okviru te raziskave.

Posamezni koraki so bili dokumentirani na način, da so vključevali definicijo potrebnih vhodnih podatkov in informacij kot tudi rezultate izvedene obdelave podatkov.

Delotoki so bili izdelani v treh ločenih segmentih z zasledovanjem treh različnih ciljev tudi zaradi možnosti zasledovanja zgolj enega od ciljev, zaradi česar tudi izhodne vrednosti individualnega delotoka ne predstavljajo omejitev za izvedbo drugega delotoka, končni uporabnik pa je tisti, ki sprejema odločitev o obsegu uporabe izdelanega modela. Če je npr. namen izvesti le prepoznavo imenskih entitet, se lahko izvede le ta delotok brez povezane obveznosti izvajanja ostalih dveh delotokov.

Delotoki, ki so del izdelanega modela, omogočajo posodobitve in nadgradnje ter modifikacije tako, da lahko končni uporabnik modela in delotokov izvede poljubne prilagoditve oziroma tudi spremeni arhitekturo in delovanje modela glede na lastne potrebe. S tem doseže, da se lahko izdelani model uporabi kot končno rešitev ali pa predstavlja izhodišče za nadaljnje raziskave ali razvoj novih modelov.

Pri izdelavi delotokov in uporabljenih programov je bil zasledovan cilj, da se uporabi le programska oprema, ki je odprtokodna ali prosto dostopna za končnega uporabnika in nima povezanih komercialnih licenc ali drugačnih omejitev pogojev uporabe. Kjer je bilo možno in izvedljivo oziroma ni predstavljalo predhodno že objavljene izvorne kode, je bila predstavljena in dokumentirana tudi izvorna koda določenih aktivnosti ali individualnih programov, kot so npr. aktivnosti ali deli aktivnosti, ki so predstavljali skripte ali individualno razvito programsko opremo, ko so npr. dodane skripte »python« (Python 2024).

Vhodni podatki oziroma izvirno gradivo, ki je predmet obdelave

Naslednji gradnik, ki je bil kritičen za izdelavo modela, so bili podatki oziroma natančneje nestrukturirano besedilo, ki je predstavljalo izvirne zapise za obdelavo. Vhodni podatki so bili za namen raziskave kot vhodna vrednost pripravljeni v tekstovni obliki, o kateri ni bilo vnaprej znanih nobenih lastnosti, ki bi na kakršenkoli način omogočale lažje ali enostavnejše doseganje zastavljenih ciljev. Ob tem je treba izpostaviti, da namen raziskave ni bila obdelava katerihkoli drugih oblik zapisov v nestrukturiranih podatkih z izjemo besedil.

Izdelani model eksplicitno ne omejuje vhodne oblike in se lahko ta v prvem koraku prilagodi, da je možen uvoz podatkov (besedil) iz različnih formatov, dokler je osnova za obdelavo podatkov tabela, ki vključuje individualen nestrukturiran zapis (besedilo) v eni vrstici tabele. Posledično lahko končni uporabnik prilagodi uvoz podatkov na način, da izvede uvoz podatkov iz različnih formatov ali s predhodno obdelavo, npr. uvozi formate elektronske pošte, s prepoznavo besedil iz slikovnih formatov itd.

Vhodni podatki so bili izbrani kot vsebina člankov novinarskih hiš, ki so omogočile dostop do omenjenih podatkov za namene raziskav. Namen izbire člankov novinarskih hiš je bil predvsem v tem, da se uporabi čim bolj razpršeno vsebino, ki vnaprej ni znana in se ne nanaša na specifično področje delovanja neke specifične organizacije. Pri tem je bil

zasledovan cilj, da uporabljeni podatki ne vsebujejo dodatne vsebine, ki bi omogočala ali olajšala lažjo izvedbo izdelave modela, kot so npr. metapodatki o posameznem izvirnem zapisu. Če je kakšen izvirni zapis vključeval takšne metapodatke, so bili pred nadaljnjo obdelavo izločeni.

Pri izdelavi modela za samodejno masovno urejanje in popisovanje nestrukturiranih besedil je bila posebna pozornost namenjena preverjanju uporabnosti izdelanega modela v primeru novega jezika besedila. Izkazalo se je, da je bil model ravno tako učinkovit, tudi če je bil uporabljen drug jezik besedila. Ob tem je treba izpostaviti, da se boljši rezultati lahko dosežejo z izbiro parametrov individualnih aktivnosti v delotoku, ki se nanašajo na specifičen jezik. Izdelava in uporaba modela za samodejno masovno urejanje in popisovanje nestrukturiranih besedil tako predstavlja univerzalno večjezično rešitev za različne uporabnike.

Pristopi za obdelavo podatkov

Metode in pristopi za obdelavo podatkov so razvidni predvsem iz individualnih aktivnosti, ki predstavljajo zaključene funkcije za obdelavo podatkov. Vsak individualen korak ali aktivnost v delotoku ima določeno funkcijo, bodisi kot del neposredne obdelave podatkov ali pa kot del merjenja učinkovitosti oziroma uspešnosti izvajanja delotoka in povezanih aktivnosti. Aktivnosti so lahko tudi povezane v določene sklope obdelav podatkov, kot je npr. »urejanje podatkov« kot opcijska izvedba, odvisno od izbranega pristopa in potrebe po uporabi takšnih sklopov obdelav podatkov. Med njih lahko prištevamo predvsem naslednje:

- urejanje besedil (formatiranje, krnjenje, izločevanje nedovoljenih simbolov, pretvorba v male/velike črtke itd.),
- ocenjevanje ustreznosti besedil (izločevanje besedil, izdelovanje lem, izračun frekvenc izrazov itd.),
- pretvorba besedil (programski objekti, vektorska oblika, vizualni elementi, izdelava skupin besed itd.),
- izdelava in uporaba odločitvenih modelov.

Večina aktivnosti v posameznem delotoku je del integrirane funkcionalnosti oziroma razpoložljivih možnosti znotraj krovne odprtokodne programske opreme, ki je bila uporabljena za obdelavo podatkov (KNIME 2024). V določenih primerih, aktivnostih ali nizu aktivnosti, ki jih omenjena programska oprema ni mogla zagotoviti, pa je bil zagotovljen dodaten razvoj funkcionalnosti prek vmesnikov za dodajanje novih programov kot del omogočanja dodatnih funkcionalnosti znotraj uporabe KNIME (2024), tj. pisanje dodatnih skriptov, ki pa so v zaključni fazi integrirani kot končna rešitev in del delotoka za uporabo izdelanega modela.

Pristop k obdelavi podatkov je zasledoval tudi vključitev izvirnih zapisov skozi izvedbo vseh korakov delotokov posebej z namenom povezovanja pridobljenih rezultatov z izvirnimi zapisi kot tudi zaradi primerjanja vmesnih rezultatov (učinkovitosti) za namen posodabljanja parametrov individualni aktivnosti. Namen povezovanja izvirnih zapisov s pridobljenimi

rezultati je bil tudi, da je bil skozi vse korake zagotovljen nadzor nad konsistentnostjo podatkov po vsaki uporabljeni obdelavi oziroma sklopu obdelave podatkov. Osnovni cilj je bil, da je izhodna vrednost delotokov enaka vhodni vrednosti, vključno s pridobljenimi rezultati.

Del izvedbe delotokov in uporabe različnih pristopov ter metod je bilo tudi merjenje časovne učinkovitosti izvajanja celotnih delotokov. Pri vsakem delotoku je bilo izvedeno merjenje, koliko časa je potrebnega, da se delotok izvede od uvoza podatkov v prvo aktivnost do izvoza podatkov za naslednji delotok. Koliko časa je potrebnega za izvedbo celotnega delotoka, tj. kako dolgo traja izvajanje vseh aktivnosti delotoka, je predvsem odvisno od količine podatkov in zahtevnosti operacij, ki se izvajajo v okviru individualne aktivnosti. Osnovne operacije, kot so npr. aktivnosti, ki se nanašajo na urejanje besedil, niso časovno potratne, na drugi strani pa so določene aktivnosti, ki so računsko obremenjujoče, v precejšnji meri tudi časovno potratne, kot je npr. izdelava odločitvenega modela.

Na hitrost obdelave podatkov predvsem v delu izdelave odločitvenih modelov lahko močno vpliva izbira strojne opreme, saj omenjene aktivnosti v prvi vrsti predstavljajo najbolj obremenjujoč del delotoka, posebej zaradi potrebe po intenzivni računski obdelavi t. i. »procesiranja podatkov«. Večina operacij se lahko pospeši z uporabo namenskih grafičnih kartic, ki omogočajo uporabo grafičnih jeder za hitrejše preračunavanje. Izhodišče raziskave je bilo namreč preveriti, ali se lahko v določenem časovnem okviru (96 ur) izvede vsak od treh delotokov v delu ene od izbranih možnosti zgolj z uporabo procesorja delovne postaje in brez uporabe namenske grafične kartice. Vsi delotoki so na podlagi uporabljene strojne opreme z izjemo ene od izbranih možnosti izvedljivi zgolj z uporabo procesorja delovne postaje. Edina možnost, ki ni bila izvedljiva v predvidnem časovnem okviru, je bila možnost izbire izdelave odločitvenega modela za klasifikacijo vsebine z uporabo arhitekture in predizdelanega modela BERT (2023).

Omenjeno vsekakor ne pomeni, da izdelava odločitvenega modela ni časovno izvedljiva brez uporabe jeder grafične kartice, le da je časovno neučinkovita. Pri uporabi arhitekture in predizdelanega modela BERT (2023) so bila za izdelavo odločitvenega modela zato uporabljena jedra grafične kartice, ki so močno pospešila izvedbo aktivnosti, tudi z uporabo jeder grafične kartice pa je bila omenjena aktivnost ena od aktivnosti, ki je potrebovala največ časa za zaključek vseh operacij.

Časovni okvir izvajanja aktivnosti je pomemben predvsem tudi zaradi pridobivanja povratnih informacij zaradi izvajanja natančnejših nastavitev (angl. Fine tuning), ki terja izvajanje večkratnih iteracij obdelave enakih podatkov. Če bi vsaka iteracija vzela preveč časa, bi to lahko ravno tako predstavljalo časovno neučinkovit pristop k izvedbi vseh potrebnih aktivnosti.

Določena izbira računske in posledično časovne obremenitve se za končno uporabo lahko opredeljuje tudi skozi individualizacijo potrebnega izvajanja posamezne aktivnosti. Npr. izdelava odločitvenega modela je računsko obremenjujoč del delotoka le v delu, ko je takšno

aktivnost treba izvesti, nadalje pri uporabi izdelanega odločitvenega modela takšne aktivnosti ni treba vnovič izvajati, saj je model že izdelan in se ga zgolj uporabi. V takšnem primeru se lahko končni uporabnik odloči, da več časa »investira« v izdelavo npr. odločitvenega modela, če se ta v praksi pogosto ne bo spreminjal, in zato ne bo treba znova izvajati procesno in časovno obremenjujočega dela delotoka.

Boljše rezultate z vidika časovne učinkovitosti je tako možno doseči z bolj zmogljivo strojno opremo, odločitev pa je odvisna predvsem od dinamike obdelav podatkov in potrebe po vnovični izdelavi odločitvenih modelov.

Arhitekture in predizdelani modeli za strojno učenje

Del raziskave je bil poleg zasledovanju temeljnih ciljev namenjen vzpostavitvi pristopa, v okviru katerega je bil namen izdelati model za samodejno masovno urejanje in popisovanje nestrukturiranih besedil z izdelavo odločitvenih modelov z uporabo predizdelanih modelov ali brez njih. Ob tem je bilo ključno preveriti, katere arhitekture in kateri predizdelani modeli predstavljajo optimalno izbiro. Izbrani pristopi za izdelavo končnega modela so tako zasledovali metode in pristope, ki so zagotavljali najboljše rezultate pri obdelavi podatkov, v določenih primerih tudi kombinacijo različnih pristopov.

V vseh treh delotokih je bilo iz pridobljenih rezultatov razvidno, da so bili boljši rezultati zagotovljeni z izdelavo odločitvenih modelov na podlagi predizdelanih modelov, kar je bilo pričakovano, saj so bili predizdelani modeli učeni/trenirani na izjemno velikih količinah podatkov. Ne glede na to, da so odločitveni modeli, ki so bili izdelani na podlagi predizdelanih modelov, učinkovitejši, pa je izbira predizdelanega modela ravno tako zelo pomembna za doseganje dobrih rezultatov, saj je pomembno, na kakšnih podatkih se je predizdelani model učil in ali ti zagotavljajo trajnejšo obliko za stabilno obdelavo podatkov, predvsem z vidika vzdrževanja predizdelanih modelov.

Izpostavljeno je predvsem pomembno tudi z vidika operacij ali ciljev, ki jih želimo doseči z uporabo takšnega predizdelanega modela. V primeru, da želimo uporabiti predizdelani model za npr. prepoznavo imenskih entitet ali za izvedbo klasifikacije vsebine, je treba izbrati predizdelani model, ki podpira omenjene pristope oziroma je bil učen z namenom, da podpre omenjene pristope, zaradi česar so tudi rezultati pri uporabi takšnega modela boljši.

Iz rezultatov raziskave pri npr. uporabi klasifikacije vsebine z izdelavo odločitvenega modela na podlagi predizdelanega modela je tako razvidno, da kakovost podatkov ne vpliva na učinkovitost in pravilnost klasifikacije, ravno obratno, rezultati so bili slabši, če se je predhodno izvedel podproces urejanja besedil. Omenjeno je predvsem posledica tega, da je bil uporabljeni predizdelani model učen/treniran na celotnih besedilih z vsemi t. i. »nepotrebnimi« informacijami, zato je bil izdelani odločitveni model natančnejši in učinkovitejši brez predhodnega urejanja besedil. Izdelava odločitvenega modela na podlagi takšnega predizdelanega modela predstavlja tudi določeno prednost, saj pri takšnih

obdelavah podatkov ni treba »skrbeti« za kakovost podatkov, kar lahko precej poenostavi delotok za klasifikacijo vsebine.

Drugi vidik, ki je pomemben za doseganje čim boljših rezultatov, je potencialna uporaba večjezičnega predizdelanega modela, predvsem v delu obdelave podatkov, ki predstavljajo večjezično vsebino.

Ne glede na to, da predizdelani modeli predstavljajo odlično podlago za izdelavo lastnih odločitvenih modelov, pa je z uporabo ustreznih metod možna tudi izdelava učinkovitih odločitvenih modelov na osnovi omejenega obsega podatkov, namenjenega učenju modela. Iz rezultatov raziskave je tako razvidno, da je npr. klasifikacija vsebine izdelave odločitvenega modela na podlagi metode SVM (Cervantes idr. 2020) dala izredno dobre rezultate brez uporabe predizdelanega modela. Na drugi strani pa je lahko izdelava odločitvenih modelov tudi močno odvisna od obsega in kakovosti podatkov, kar je npr. razvidno iz rezultatov izdelave odločitvenega modela za prepoznavo imenskih entitet z uporabo slovarjev, kjer bi za izdelavo splošno učinkovitega odločitvenega modela potrebovali izjemno obsežne slovarje različnih tipov imenskih entitet. Postopek bi bil zaradi tega zelo procesorsko in časovno obremenjujoč. Uporaba predizdelanega modela lahko sicer precej skrajša postopek izdelave odločitvenega modela, vendar je celoten čas izdelave izvedbenega modela še vedno precej odvisen od načina uporabe predizdelanega modela. zlasti pri določanju parametrov učenja na podlagi predizdelanega modela (angl. Fine tuning).

Izhodni podatki, pripravljeni za nadaljnjo obdelavo

Da bi lahko združevali vse izhodne podatke različnih delotokov, je bilo treba vzpostaviti pristop povezovanja vseh pridobljenih rezultatov. Za ta namen je bil izbran najenostavnejši pristop, in sicer priprava in prenos izhodnih podatkov posameznega delotoka v obliko, ki je predstavljala vhodno obliko za naslednji delotok. Na takšen način se je zadržala osnovna struktura izvirnih zapisov, vključno s pridobljenimi rezultati vseh delotokov. Ob tem v okviru delotokov ni omejitev glede obsega izvoza podatkov. Zaradi praktičnih razlogov so bili podatki, namenjeni za izvoz, omejeni le na tisti del, ki je bil nujno potreben za nadaljnjo obdelavo podatkov. Končni uporabnik lahko v okviru delotoka poljubno razširi obseg podatkov, ki jih želi izvoziti kot rezultat npr. poleg klasifikacijske oznake, ki ga v praksi predstavlja. Določen tematski model lahko izvozi tudi podatke o ključnih besedah, ki tvorijo posamezen tematski model oziroma določeno klasifikacijo. Takšen pristop je lahko koristen zaradi lažjega razumevanja določanja klasifikacijskih oznak ali pa zaradi ocenjevanja ustreznosti izvedenih odločitvenih modelov.

Izhodne vrednosti podatkov so bile izvožene v tekstovni obliki zaradi lažje potencialne naknadne obdelave. Namen uporabe standardiziranega tekstovnega formata je bil tudi v tem, da je z uporabo enostavnih programskih orodij za predogled tekstovne oblike vedno možno preveriti vmesne pridobljene rezultate in konsistentnost gradiva, pa tudi obliko, ki omogoča enostavno ekstrakcijo podatkov za namen optimizacije modela. Pri tem je bila posebna pozornost namenjena ustrezni dokumentiranosti podatkov, tj. preglednosti v obliki izdelanih

povezanih metapodatkov. Cilj raziskave je bila izdelava modela za samodejno masovno urejanje in popisovanje nestrukturiranih besedil, pri čemer se je zasledovalo tri glavne cilje, ki so bili tudi glavni del izhodne vrednosti podatkov: prepoznane imenske entitete, določene klasifikacije ter izdelani naslovi individualnih zapisov. Poleg omenjenih informacij so kot del izhodnih podatkov dodani še ostali fiksni podatki zaradi lažjega dokumentiranja (metapodatki) individualne popisne enote na podlagi individualnega izvirnega zapisa. Celotna obdelava podatkov je bila izvedena v kodnem načinu, ki je zadržal ustrezno strukturo podatkov skozi vse korake obdelave, od uvoza podatkov skozi morebitno urejanje podatkov do izvoza podatkov.

Izdelani model tako kot v primeru vhodnih podatkov eksplicitno ne omejuje izhodne oblike in se lahko v prvem koraku prilagodi, da je možen izvoz podatkov v različne strukture ali formate, odvisno od potreb končnega uporabnika. Ravno tako večjezičnost ne vpliva na izhodno obliko podatkov, saj je izvoz primarnih podatkov namenjen strukturiranju podatkov za potencialno nadaljnjo obdelavo in ne dodatni vsebinski obdelavi.

Predogled izdelanih popisnih enot

Eden od namenov raziskave in izdelave modela za samodejno masovno urejanje in popisovanje nestrukturiranih besedil je bil preveriti uporabnost izdelanega modela in pridobljenih rezultatov. Za ta namen je bil za predogled pridobljenih rezultatov razvit grafični vmesnik, ki omogoča jasen in pregleden način pregledovanja izvirnih zapisov in povezanih pridobljenih rezultatov.

Predogled rezultatov je namenjen tako končnim uporabnikom, ki lahko z vsebinskega vidika ocenjujejo ustreznost pridobljenih rezultatov, kot tudi snovalcem oziroma osebam, ki optimizirajo model, da lahko ocenjujejo delovanje oziroma uporabo modela. Namen uporabe modela za samodejno masovno urejanje in popisovanje nestrukturiranih besedil ni omejitev uporabe le na eno kategorijo uporabnikov, npr. uporabo zgolj s strani informatikov ali razvijalcev sistemov za strojno učenje, ampak je namenjen predvsem uporabi s strani arhivistov kot končnih uporabnikov, zato je pomembno, da je tudi končna rešitev čim bolj prijazna do končnega uporabnika.

Glede na to, da so izhodni podatki v tekstovni obliki oziroma da ni praktičnih omejitev v okviru delotokov, v kakšni obliki se podatki lahko izvozijo, lahko končni uporabnik modela prilagodi izvoz rezultatov v obliko, ki mu najbolj ustreza za vizualizacijo ali uporabo v drugih programskih orodjih.

Grafični vmesnik, ki je bil razvit za potrebe predogleda pridobljenih rezultatov, je omejen na predogled pridobljenih rezultatov izključno v izhodni obliki, kot je bila vzpostavljena v zadnjem delotoku izdelave popisne enote. V primeru spreminjanja strukture izvoženih podatkov je potreben dodaten razvoj, da se zagotovi podpora obdelavi podatkov za izvedbo pravilnega predogleda izdelanih popisnih enot ali pa izvoz podatkov v eno od standardnih metapodatkovnih struktur, ki se uporabljajo na področju popisovanja arhivskega gradiva.

4 ZAKLJUČEK

Uvodoma je bil opredeljeno, da je temeljni problem pri velikih količinah gradiva v digitalni obliki, ki ima neznano vsebino in nestrukturirano obliko v tem, da je lahko pri izjemno velikih količinah takšnega gradiva težko zagotoviti hitro in učinkovito ter ustrezno vsebinsko obravnavo za namen klasificiranja gradiva oziroma za namen urejanja in popisovanja gradiva v digitalni obliki.

Izdelani model za samodejno masovno urejanje in popisovanje nestrukturiranih besedil je eden od možnih odgovorov, kako se lahko pristopi k reševanju omenjenega problema. Raziskava je pokazala, da je možno izdelati rešitve, ki omogočajo učinkovito obdelavo večjih količin gradiva na način, da sta zagotovljeni ustrezna kakovost in varnost obdelanega gradiva. Pri tem ključno vlogo igra izbira ustrezne tehnologije in metod ter pristopov, s katerimi so bili cilji lažje doseženi.

Velike količine besedil z nestrukturirano vsebino so realnost, s katero se arhivisti srečujejo dnevno, količine takšnega gradiva pa so ravno tako iz dneva v dan večje. Tradicionalni pristopi za urejanje in popisovanje gradiva v večjih količinah brez uporabe modernih tehnologij so že danes neizvedljivi ali neuporabni. Pristop, kjer bi se izvedlo urejanje in popisovanje gradiv na t. i. ročni način s strani posameznega arhivista za vsak individualen zapis, časovno ni vzdržen in niti izvedljiv glede na prirast takšnega gradiva.

Raziskava je bila razdeljena na zasledovanje treh ciljev:

- ali je možna vsebinska opredelitev (klasifikacija) besedil z nestrukturirano vsebino,
- ali je možna prepoznava imenskih entitet v besedilih z nestrukturirano vsebino in
- ali je možna izdelava naslova za posamezno popisno enoto gradiva, t. i. »izvirni zapis« z nestrukturirano vsebino.

Identifikacija omenjenih kategorij informacij bi lahko močno olajšala delo arhivistov oziroma dokumentalistov v primeru urejanja in popisovanja gradiva, pri tem pa je bilo treba realno oceniti tudi morebitne omejitve ali probleme, ki niso bili rešeni v izvedeni raziskavi in predstavljajo možnosti za dodatne raziskave in razvoj izboljšav, ki bi omogočale učinkovitejšo izdelavo modela za samodejno masovno urejanje in popisovanje nestrukturiranih besedil.

4.1 Vsebinska klasifikacija gradiva

Vsebinska klasifikacija je eden od možnih načinov, s pomočjo katerega se lahko izvede določena oblika vrednotenja gradiva predvsem v smislu ocenjevanja potreb po izvedbi urejanja in popisovanja gradiva. Vse gradivo, ki nastaja, še ne predstavlja tudi gradiva, ki ga je treba hraniti, zato je pomembno, da se z vidika optimizacije del gradiva loči na tisti del, ki ga je treba urejati in popisovati, od tistega gradiva, ki se ga lahko izloči oziroma za katerega urejanje in popisovanje ni potrebno.

Rezultati raziskave so pokazali, da vsebinska klasifikacija z uporabo strojnega učenja predstavlja učinkovito rešitev za reševanje temeljnega problema obdelave nestrukturiranih besedil v velikih količinah. Obstoječa tehnologija, metode in pristopi so zreli in stabilni za izdelavo odločitvenih modelov za izvedbo klasifikacije, ravno tako se postopoma poenostavlja uporaba takšnih tehnologij v obliki naprednih programskih orodij, standardizacij, ugotovitev obstoječih raziskav in dokumentiranosti pristopov. Rezultati vsebinske klasifikacije tako lahko predstavljajo podlago za končno obliko popisanega gradiva, saj lahko omogočajo nadaljnjo obdelavo v kontekstu združevanja podatkov, iskanja povezav med različnimi zapisi ali pa kot zelo pomemben del vrednotenja gradiva.

Ob tem se pri izbranem pristopu v okviru raziskave pojavlja dodatno nerešeno vprašanje oziroma problem potencialnih subjektivnih kriterijev pri izbiri klasifikacijskega okvira za izvedbo klasifikacije, tj. obseg izdelanih tematskih modelov. Učinkovitost in smiselnost vsebinske klasifikacije se tako še vedno nanaša na opredelitev obsega in strukture izdelave tematskih modelov s strani snovalca sistema (delotoka) za izdelavo tematskih modelov. V primeru, da se kriterij obsega tematskih modelov za izvedbo klasifikacije zastavi preširoko ali preozko, se lahko takšni kriteriji prenesejo na končne rezultate vsebinske klasifikacije in ne predstavljajo najbolj optimalne oblike za popisovanje gradiva ali npr. za vrednotenje gradiva. Lahko predstavljajo tudi neustrezno klasifikacijo gradiva. V tem delu bi bila vsekakor potrebna dodatna raziskava kako pristopiti k zmanjševanju subjektivnih kriterijev snovalcev sistema kot dejavnik vpliva na končne rezultate obdelave podatkov. Ena od možnih rešitev bi lahko bil sistem z več kontrolami, ki bi bile namenjene večplastnemu preverjanju ustreznosti sprejetih odločitev, ali pa pristop, ki bi vključeval več strokovnih deležnikov za potrjevanje sprejetih odločitev.

Dobra praksa pri uporabi modela za samodejno masovno urejanje in popisovanje nestrukturiranih besedil za namen vsebinske klasifikacije je tudi izbira ustreznega pristopa z uporabo predizdelanih modelov ali brez nje. Omenjeno je predvsem odvisno od vhodne oblike podatkov. V tem delu je zlasti mišljeno, v kakšnem jeziku se nahaja gradivo in ali je potrebna kakršnakoli predhodna obdelava besedila, kot je npr. »urejanje podatkov« z namenom zagotavljanja višje kakovosti podatkov. Za ta namen je dobra praksa, da se predhodno vsaj osnovno pregleda in preveri vhodno obliko podatkov in ustrezno prilagodi parametre za izvedbo delotokov, da bo uporaba modela za samodejno masovno urejanje in popisovanje nestrukturiranih besedil čim bolj učinkovita.

4.2 Prepoznava imenskih entitet

Prepoznava imenskih entitet z uporabo modela za samodejno masovno urejanje in popisovanje nestrukturiranih besedil omogoča identifikacijo osnovnih informacij, na podlagi katerih je naknadno možno lažje uporabljati popisano gradivo. Prepoznane imenske entitete lahko tako poleg prepoznanih entitet omogočajo tudi potencialno lažjo predvsem pa hitrejšo kontekstualizacijo vsebine v navezi z vsebinsko klasifikacijo gradiva.

Pomemben del procesa prepoznavanja imenskih entitet je odločitev glede obsega tipov imenskih entitet, ki bi jih radi prepoznali, zato je tudi pristop k izdelavi modela ali uporabi predizdelanih modelov lahko drugačen. Enako velja za vsebinsko klasifikacijo, za katero je v primeru prepoznave imenskih entitet ključno predhodno jasno določanje ciljev in pričakovanih rezultatov, saj je od tega, tj. določanja parametrov izvedbe, potem odvisna oblika uporabe izdelanega modela.

Kot v primeru vsebinske klasifikacije so tudi na področju prepoznavanja imenskih entitet obstoječe tehnologije, metode in pristopi zreli in stabilni za izdelavo odločitvenih modelov oziroma prepoznavanje imenskih entitet. Z ustrezno nastavitvijo parametrov pri uporabi predizdelanih modelov ali z dovolj velikim obsegom podatkov za učenje omogočajo zelo učinkovito prepoznavanje.

V raziskavi smo se pri prepoznavi imenskih entitet soočili s specifičnim problem, ki sicer ni v celoti vplival na izvedbo delotokov, je pa lahko v izjemno majhnem in izoliranem obsegu vplival na potencialno kakovost pridobljenih rezultatov. Večpomenskost imenskih entitet je v določenem delu lahko vplivala na pravilnost določitve tipa imenske entitete. Npr. beseda »petek« je lahko identificirana kot tip imenske entitete »oseba« kot tudi tip imenske entitete »čas«. Zaradi pomanjkanja ustreznih mehanizmov za ločevanje ugotavljanja ustreznosti pripadnosti pravemu tipu imenske entitete je model določil imensko entiteto večkratno za vsak identificiran tip entitete. Omenjeni problem zato lahko povzroči neustrezno določanje tipa imenskih entitet in uvrščanje neustrezno prepoznane imenske entitete v končni rezultat obdelave besedila.

Drug pomemben vidik oziroma dobra praksa pri izbiri pristopa za prepoznavo imenskih entitet in uporabo potencialnega predizdelanega modela je predhodno preverjanje, v katerem jeziku se nahaja besedilo. V primeru uporabe predizdelanih modelov sledi izbira ustreznega predizdelanega modela. Prepoznava določenih vsebin oziroma entitet in povezana terminologija sta lahko precej zahtevni, če odločitveni modeli niso učeni/trenirani na vsebini, ki je primerljiva oziroma v ustreznem enakem jeziku, kot je jezik, v katerem izvajamo prepoznavo.

4.3 Izdelava naslova popisne enote

Izmed vseh zastavljenih ciljev se je izdelava naslova z uporabo strojnega učenja izkazala za najzahtevnejši cilj, ker je ocenjevanje ustreznosti doseženih rezultatov precej odvisno od kriterijev sprejemljivosti, ki so odvisni od posameznega uporabnika modela. Ena od temeljnih nalog popisovanja gradiva je zagotoviti pregleden in reprezentativen naslov popisne enote glede na vsebino gradiva. Različni popisovalci bi tako v praksi isto enoto gradiva praviloma naslovili drugače glede na lastno interpretacijo in razumevanje vsebine, zato bi se tudi v realnosti z omenjenim tradicionalnim načinom lahko za popisno enoto dokumentiralo različne naslove popisne enote.

Pri izdelavi naslova popisne enote tako ni bila v ospredju t. i. točnost ali učinkovitost, saj bi jo bilo, kot je bilo omenjeno v prejšnjem odstavku, težko meriti ali ocenjevati. V tem delu je bila edina meritev, ki je bila izvedena, subjektivna ocena, ali je izdelani naslov pregleden in ali predstavlja ustrezen vsebinski opis vsebine gradiva.

Za izdelavo naslova je bila uporabljena metoda abstraktnega povzemanja, ki se za razliko od ekstrakcijskega povzemanja skuša izogniti zgolj povzemanju vsebine na najbolj reprezentativen stavek ali skupek besed iz besedila, temveč je bil namen, da naslov skozi abstraktno povzemanje predstavlja svojo novo obliko opisa vsebine v skladu z razumevanjem vsebine.

Določen problem, ki se pri izbrani metodi povzemanja lahko pojavlja, je ponavljanje oziroma podobna sintaksa pri izbiri besed in tvorjenju reprezentativne vsebine v obliki stavkov. Da bi se lahko izognili potencialnemu ponavljanju bodisi oblik stavkov ali izbire besed, bi bila v tem delu smiselna dodatna raziskava, ki bi se lahko osredotočila na raziskovanje in izdelavo bolj dinamične oblike naslova oziroma diferenciacijo sintakse za izdelavo naslova. Ena od možnih rešitev bi lahko bilo omejevanje že uporabljenih struktur ali besed v določenem vzorcu obdelanega besedila.

4.4 Model za samodejno masovno urejanje in popisovanje nestrukturiranih besedil

Uvodoma je bilo opredeljeno, da količina nestrukturiranih podatkov nezadržno raste in že danes predstavlja glavnino vseh podatkov, do leta 2025 pa se bo predvidoma ta delež še dvignil do predvidenih 80 % vseh podatkov (Burgener 2021). Glede na to, da danes še ne razpolagamo z celovitimi enostavnimi produkcijskimi sistemi, ki bi nam omogočili enostavno obdelavo večjih količin nestrukturiranih podatkov z uporabo strojnega učenja, je treba k reševanju temeljnega problema in izdelavi takšnih modelov ter povezanih rešitev pristopiti čim prej. V raziskavi je bil izdelan model za samodejno masovno urejanje in popisovanje nestrukturiranih besedil, ki predstavlja možno celovito rešitev za reševanje omenjene problematike tako z vidika izdelave načrtnega referenčnega okvira kot z vidika zagotavljanja informacijskega sistema kot praktične rešitve za končnega uporabnika.

Izdelani model za samodejno masovno urejanje in popisovanje nestrukturiranih besedil je pripravljen za takojšnjo uporabo v praksi z ustrezno izvedbo nastavitev s strani končnega uporabnika tako v aplikativni kot dokumentacijski obliki na podlagi jasno dokumentiranih aktivnosti, ki jih je zagotovila raziskava, in s preglednostjo nad vključenimi koraki in pričakovanimi rezultati.

Ob vseh rezultatih in funkcionalni vzpostavitvi modela za samodejno masovno urejanje in popisovanje nestrukturiranih besedil pa je treba izpostaviti trenutno največji problem oziroma rezultat obdelave podatkov, ki ga ta raziskava ni zajela, in sicer, kako ravnati z nepravilno in neustrezno obdelanimi podatki oziroma s t. i. »nepravilnimi rezultati«. Npr. v primeru klasifikacije vsebine je znašal najboljši rezultat, ki je bil izmerjen, 94 % točnosti izvedene klasifikacije glede na predhodno znano klasifikacijo posameznega izvirnega zapisa

iz delotoka določanja tematskih modelov. Pri obsegu 50.000 izvirnih zapisov, kolikor je predstavljal obseg gradiva, 6 % napačno klasificiranih izvirnih zapisov predstavlja 3000 enot popisnega gradiva, ki niso bile pravilno klasificirane.

V primeru prepoznave imenskih entitet je merjenje učinkovitosti izvedene prepoznave izjemno težko, saj neznano besedilo v nestrukturirani obliki predhodno ne vključuje prepoznanih entitet, da bi lahko takšno primerjavo izvedli, zato je bil edini način, da se preveri učinkovitost prepoznavanja, da se izvede t. i. »ročno« prepoznavanje in primerja rezultate s tistimi, ki so bili pridobljeni z uporabo modela. Slednji pristop nam ravno tako ne ponudi odgovora, kakšna je učinkovitost na celotnem obsegu vseh izvirnih zapisov, ki so bili predmet obdelave.

Kot je že bilo omenjeno, se podobna težava pojavlja tudi v primeru izdelave naslova popisne enote, kjer je ravno tako izjemno težko meriti učinkovitost in pravilnost končnih rezultatov, saj so pogojeni s subjektivnimi kriteriji glede ustreznosti opisa vsebine.

V osnovi se kot največji problem, ki ga raziskava ni zajela, kaže v tem, kako ravnati in obdelati nepravilne ali neustrezne rezultate, da bi dosegli čim boljši učinek z vidika celovitosti obdelave. Dodatna raziskava za obravnavo potencialnih rešitev za reševanje omenjenega problema bi vsekakor ponudila velik korak k zagotavljanju celovite rešitve za izdelavo modela za samodejno masovno urejanje in popisovanje nestrukturiranih besedil, ki bi vključeval tudi dodatno obdelavo t. i. »nepravilnih rezultatov«.

Na področju klasifikacije vsebine je omenjeni izziv mogoče malce enostavnejši, saj je z izdelavo tematskih modelov zagotovljena referenčna predklasifikacija, ki je lahko rešitev za analizo odstopanj in prilagajanje izdelave tematskih modelov v bolj optimalno obliko. Možna bi bila tudi npr. uporaba uteži in presojanja, ali je izvirni zapis možno pod določeno verjetnostjo uvrstiti med obstoječe klasifikacijske razrede (tematske modele). Sledilo bi potencialno izločevanje in preusmeritev v posebno obravnavo. Spet ena od možnih rešitev bi bila večdimenzijska uporaba modelov, kjer bi se lahko uporabljale gruče različnih izdelanih tematskih modelov iz različnih vsebin itd.

Večja težava se kaže v primeru prepoznave imenskih entitet in izdelave naslova popisne enote, kjer sta obseg in oblika rešitve odvisna predvsem od uporabnika, npr. obseg tipov imenskih entitet ali maksimalna dolžina besed, ki lahko tvorijo naslov popisne enote. Identifikacija, ali je bilo nekaj nepravilno, pomanjkljivo ali neustrezno narejeno, se lahko opravi le z osebnim oziroma »ročnim pregledom«, kar pa v primeru velikih količin podatkov ni racionalno. Zato to ostaja problem, ki ga je treba raziskati in zanj poskušati najti optimalne rešitve, ki bi dvignile kakovost končnih rezultatov. Hkrati je treba zagotoviti merljive normative, ki bi lahko bili uporabljeni za preverjanje kakovosti tako pridobljenih rezultatov.

Raziskava je pokazala, da je izdelava modela za samodejno masovno urejanje in popisovanje nestrukturiranih besedil izvedljiva, uporaba modela pa učinkovita rešitev za reševanje temeljnega problema kako pristopiti k urejanju in popisovanju nestrukturiranih besedil.

5 LITERATURA

1. Abacha Asma, Ben, Faisal Mahbub Chowdhury, Aikaterini Karanasiou, Yassine Mrabet, Alberto Lavelli in Pierre Zweigenbaum. 2015. »Text mining for pharmacovigilance: Using machine learning for drug name recognition and drug-drug interaction extraction and classification«. *Journal of Biomedical Informatics* 58: 122–132. Dostop 5. januar 2024. https://www.sciencedirect.com/science/article/pii/S1532046415002099.
2. Abdel-Salam, Shehab in Ahmed Rafea. 2022. Performance Study on Extractive Text Summarization Using BERT Models. *Information* 2022(13): 67. Dostop 5. januar 2024. https://www.researchgate.net/publication/358207315_Performance_Study_on_Extractive_Text_Summarization_Using_BERT_Models.
3. Adnan, Kiran. Rehan Akbar in Khor Siak Wang. 2021. »Development of Usability Enhancement Model for Unstructured Big Data Using SLR«. *IEEE Access* 9: 87391–87409. Dostop 5. januar 2024. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9454431.
4. Afantenos, Stergos, Vangelis Karkaletsis in Panagiotis Stamatopoulos. 2005. »Summarization from medical documents: a survey«. *Artificial Intelligence in Medicine* 33(2): 157–177. Dostop 5. januar 2024. https://www.researchgate.net/publication/220103096_Summarization_from_Medical_Documents_A_Survey.
5. Agrawal, Amritanshu, Wei Fu in Tim Menzies. 2018. »What is wrong with topic modeling? And how to fix it using search-based software engineering«. *Information and Software Technology* 98: 74–88. Dostop 5. januar 2024. https://arxiv.org/pdf/1608.08176.
6. Akerkar, Rajendra. 2019. *Artifcial Intelligence for Business*. Switzerland: Springer International Publishing AG.
7. Allahyari, Mehdi, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez in Krys Kochut. 2017. »Text Summarization Techniques: A Brief Survey«. *International Journal of Advanced Computer Science and Applications* 8(10). Dostop 5. januar 2024. https://thesai.org/Downloads/Volume8No10/Paper_52-Text_Summarization_Techniques_a_Brief_Survey.pdf.
8. Andhale, Narendra in Laxmi Bewoor. 2016. »An overview of Text Summarization techniques«. V *International conference on computing communication control and automation*. Dostop 15. januar 2024. https://www.researchgate.net/publication/313955830_An_overview_of_Text_Summarization_techniques.
9. Anthony, Patricia, Rayner Alfred, Leow Ching Leong, Chin Kim On, Tan Soo Fun, Mohd Norhisham Bin Razali in Mohd Hanafi Ahmad Hijazi. 2013. »A Rule-Based Named-Entity Recognition for Malay Articles«. *Lecture Notes in Computer Science* 8346: 288–299. Dostop 5. januar 2024. https://www.researchgate.net/publication/275644054_A_Rule-Based_Named-Entity_Recognition_for_Malay_Articles.
10. Aparna, M. in Srinath Srinivasa. 2023. »Active learning for Named Entity Recognition in Kannada«. *TechRxiv*. Dostop 5. januar 2024.

https://www.researchgate.net/publication/375976751_Active_learning_for_Named_Entity_Recognition_in_Kannada.
11. *Arhivski slovarček*. 2024. Dostop 1. januar 2024. https://www.zal-lj.si/project/arhivski-slovarcek/#:~:text=Arhiv%3A%20ustanova%2C%20ki%20opravlja%20javno,nastalo%20v%20%C4%8Dasu%20njegovega%20delovanja.
12. Arora, Sanjeev, Rong Ge, Yoni Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu in Michael Zhu. »A Practical Algorithm for Topic Modeling with Provable Guarantees«. Dostop 27. februar 2020. http://proceedings.mlr.press/v28/arora13.pdf.
13. Arras, Leila, Franziska Horn, Gregoire Montavon, Klaus-Robert Muller in Wojciech Samek. 2017. »What is relevant in a text document?: An interpretable machine learning approach«. *Plos One* 12(8): 23. Dostop 1. januar 2024. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0181142.
14. Asmussen, Claus Boye in Charles Møller. 2019. »Smart literature review: a practical topic modelling approach to exploratory literature review«. *Journal of Big Data* 6: 93. Dostop 1. januar 2024. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0255-7.
15. Augenstein, Isabelle, Leon Derczynski in Kalina Bontcheva. 2017. »Generalisation in named entity recognition: A quantitative analysis«. *Computer Speech and Language* 44: 61–83. Dostop 1. januar 2024. https://www.sciencedirect.com/science/article/pii/S088523081630002X.
16. Baig, Junaid. 2023. *Unstructured Data Challenges for 2023 and their Solutions*. Dostop 12. januar 2024. https://www.astera.com/type/blog/unstructured-data-challenges/.
17. Balič, Jože. 2004. *Inteligentni obdelovalni sistemi*. Maribor: Fakulteta za strojništvo, Univerza v Mariboru.
18. Barredo, Alejandro Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila in Francisco Herrera. 2020. »Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI«. *Information Fusion* 58: 82–115. Dostop 1. januar 2024. https://arxiv.org/pdf/1910.10045v1.
19. Benedetti del Rio, Valeria. 2023. »Ethical AI: The European Approach to Achieving the SDGs Through AI«. V *The Ethics of Artificial Intelligence for the Sustainable Development Goals,* ur. Francesca Mazzi in Luciano Floridi, 167-181. Switzerland: Cham. Springer Nature AG.
20. Bennasar, Mohamed, Yulia Hicks in Rossitza Setchi. 2015. »Feature selection using Joint Mutual Information Maximisation«. *Expert Systems with Applications* 42(22): 8520–8532. Dostop 1. januar 2024. https://www.sciencedirect.com/science/article/pii/S0957417415004674.
21. *BERT*. 2023. Dostop 1. januar 2023. https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270.
22. *Bert-base-multilingual-cased*. 2023. Dostop 19. december 2023. https://huggingface.co/google-bert/bert-base-multilingual-cased.
23. *BERT-large-NER*. 2023. Dostop 19. december 2023. https://huggingface.co/dslim/bert-large-NER.

24. Bhbosale, Sachin, Vinayak Pujari in Zameer Multani. 2020. »Advantages And Disadvantages Of Artificial Intellegence«. *Aayushi International Interdisciplinary Research Journal 77*. Dostop 1. januar 2024. https://www.researchgate.net/publication/344584269_Advantages_And_Disadvantages_Of_Artificial_Intellegence.
25. Bidoki, Mohammad, Mohammad R. Moosavi in Mostafa Fakhrahmad. 2020. »A semantic approach to extractive multi-document summarization: Applying sentence expansion for tuning of conceptual densities«. *Information Processing & Management* 57(6).
26. Blei M. David, Andrew Y. Ng, Michael I. Jordan. 2003. »Latent Dirichlet Allocation«. *Journal of Machine Learning Research* 3. Dostop 1. januar 2024. http://www.jmlr.org/papers/volume3/ blei03a/blei03a.pdf.
27. Boddington, Paula. 2023. *AI Ethics*. Singapore: Springer Nature Singapore Pte Ltd.
28. Bottou, Leon, Frank E. Curtis in Jorge Nocedal. 2018. »Optimization methods for large-scale machine learning«. *SIAM Review* 60(2): 223–311. Dostop 1. januar 2024. https://epubs.siam.org/doi/10.1137/16M1080173.
29. Burgener, Eric. 2021. »Meeting the New Unstructured Storage Requirements for Digitally Transforming Enterprises«. *IDC White paper*. Dostop 10. januar 2024. https://www.delltechnologies.com/asset/en-th/ products/storage/industry-market/ h18811-wp-idc-new-unstructured-storage-requirements-for-enterprises.pdf.
30. Caliskan, Aylin, Joanna J. Bryson in Arvind Narayanan. 2017. »Semantics derived automatically from language corpora contain human-like biases«. *Science* 356(6334): 4. Dostop 10. januar 2024. https://pubmed.ncbi.nlm.nih.gov/28408601/.
31. Canedo, Edna Dias in Bruno Cordeiro Mendes. 2020. »Software Requirements Classification Using Machine Learning Algorithms«. *Entropy* 22(9): 20. Dostop 10. januar 2024. https://www.mdpi.com/1099-4300/22/9/1057.
32. Cath, Corinne. 2018. »Governing artificial intelligence: Ethical, legal and technical opportunities and challenges«. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376(2133). Dostop 10. januar 2024. https://royalsocietypublishing.org/doi/10.1098/rsta.2018.0080.
33. Cervantes, Jair, Farid Garcia-Lamont, Lisbeth Rodríguez-Mazahua in Asdrubal Lopez. 2020. »A comprehensive survey on support vector machine classification: Applications, challenges and trends«. *Neurocomputing*, 408(2020): 189–215.
34. *CLARIN.SI*. 2022. Dostop 1. januar 2022. http://www.clarin.si/info/about/.
35. Chan, Leong, Liliya Hogaboam in Renzhi Cao. 2022. *Applied Artificial Intelligence in Business-Concepts and Cases*. Switzerland: Springer Nature.
36. Chen, Haihua, Junhua Ding, Wei Lu in Jay Bhuyan. 2022. »"Data Quality for Big Data and Machine Learning" Frontiers in Big Data«. Dostop 5. januar 2024. https://www.researchgate.net/publication/359245326_Data_Quality_for_Big_Data_and_Machine_Learning_Frontiers_in_Big_Data.
37. *COBISS*. 2024. Dostop 1. januar 2024. https://www.cobiss.si/.
38. Collins, Christopher, Denis Dennehy, Kieran Conboy in Patrick Mikalef. 2021. »Artificial intelligence in information systems research: A systematic literature review and research agenda«. *International Journal of Information Management* 60. Dostop

10. januar 2024. https://www.sciencedirect.com/ science/article/ pii/S0268401221000761.
39. Conneau, Alexis, Douwe Kiela, Holger Schwenk, Loıc Barrault in Antoine Bordes. 2017. »Supervised learning of universal sentence representations from natural language inference data«. Dostop 01. januar 2022. https://arxiv.org/pdf/1705.02364.pdf.
40. Cordell, Ryan. 2020. Machine Learning + Libraries: A Report on teh State of the Field. *Library of Congress*. Dostop 10. januar 2022. https://labs.loc.gov/static/ labs/work/reports/Cordell-LOC-ML-report.pdf.
41. Couture, Carol. 2005. »Archival Appraisal: A Status Report«. *Archivaria* 59: 83–107. Dostop 10. januar 2024. https://archivaria.ca/index.php/ archivaria/article/view/ 12502/13624.
42. CSV. 2023. *Comma Separated Values*. Dostop 15. september 2023. https://www.loc.gov/preservation/digital/formats/fdd/fdd000323.shtml.
43. Čargo, Danijela Juričič. 2018. »Popisovanje v slovenskih arhivih – standardi in njihova uporaba«. Moderna arhivistika 2018(2). Dostop 5. januar 2024. http://www.pokarh-mb.si/uploaded/datoteke/2_2018_365-384_juricic.pdf.
44. Dada, Gbenga Emmanuel, Joseph Stephen Bassi, Haruna Chiroma, Shafi'i Muhammad Abdulhamid, Adebayo Olusola Adetunmbi, Opeyemi Emmanuel Ajibuwa. 2019. »Machine learning for email spam filtering: review, approaches and open research problems«. *Heliyon* 5 (6). Dostop 5. januar 2024. https://www.cell.com/action/showPdf?pii=S2405-8440%2818%2935340-4.
45. Dey, Ritwik. 2021. »What Is Artificial Intelligence? Everything You Need to Know in 2021«. Dostop 10. januar 2024. https://www.sganalytics.com/blog/what-is-artificial-intelligence-everything-you-need-to-know-in-2021/.
46. Davenport, Thomas H. in Rajeev Ronanki. 2018. Artificial Intelligence fort je Real World. *Harvard Business Review January–February 2018*. Dostop 10. januar 2024. https://blockqai.com/wp-content/uploads/2021/01/analytics-hbr-ai-for-the-real-world.pdf.
47. Derczynski, Leon, Diana Maynarda, Giuseppe Rizzo, Marieke van Erp, Genevieve Gorrell, Raphael Troncy, Johann Petraka in Kalina Bontcheva. 2015. »Analysis of named entity recognition and linking for tweets«. *Information Processing & Management* 51(2): 32–49. Dostop 10. januar 2024. https://arxiv.org/abs/1410.7182.
48. Devlin, Jacob, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2018. »BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding«. Dostop 18. december 2023. https://arxiv.org/abs/1810.04805.
49. Dernoncourt, Franck, Ji Young Lee, Ozlem Uzuner in Peter Szolovits. 2017. »De-identification of patient notes with recurrent neural networks«. *Journal of the American Medical Informatics Association* 24(3): 596–606. Dostop 10. januar 2024. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7787254/.
50. Dierickx, Laurence, Carl-Gustav Linden, Andreas Opdahl, Sohail Ahmed Khan in Diana Rojas. 2023. AI in the newsroom: A data quality assessment framework for employing machine learning in journalistic workflows. V *5th International Conference on Advanced Research Methods and Analytics (CARMA 2023)*. Dostop 19. december 2023. https://www.researchgate.net/ publication/374233511_AI_in_the_newsroom_

A_data_quality_assessment_framework_for_employing_machine_learning_in_journalistic_workflows.
51. Dilmegani, Cem. 2024. »100+ AI Use Cases & Applications: In-Depth Guide for 2024«. Dostop 10. januar 2024. https://research.aimultiple.com/ai-usecases/.
52. *Distilbart-xsum-6-6*. 2023. Dostop 19. december 2023. https://huggingface.co/sshleifer/distilbart-xsum-6-6.
53. Duan, Yanqing, John S. Edwards in Yogesh K Dwivedi. 2019. »Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda«. *International Journal of Information Management* 48: 63–71. Dostop 10. januar 2024. https://cronfa.swan.ac.uk/Record/cronfa48788.
54. EAD. 2019. *Encoded Archival Description*. Chicago: Society of American Archivists.
55. Eftimov, Tome, Barbara Koroušić Seljak in Peter Korošec. 2017. »A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations«. *PLoS ONE* 12(6). Dostop 10. januar 2024. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0179488.
56. Ehrlinger, Lisa in Wolfram Wöß. 2022. »A Survey of Data Quality Measurement and Monitoring Tools«. *Frontiers in Big Data* 5. Dostop 10. januar 2024. https://www.frontiersin.org/articles/10.3389/fdata.2022.850611/full.
57. Eisenstein, Jacob. 2018. »Natural language processing«. *MIT press*. Dostop 18 december 2023. https://cseweb.ucsd.edu/~nnakashole/teaching/eisenstein-nov18.pdf.
58. El-Kassas, Wafaa S., Cherif R. Salama, Ahmed A. Rafea in Hoda K. Mohamed. 2020. »Automatic text summarization: A comprehensive survey«. *Expert Systems with Applications* 165. Dostop 10. januar 2024. https://www.researchgate.net/publication/342878502_Automatic_Text_Summarization_A_Comprehensive_Survey.
59. Elouataoui, Widad, Imane El Alaoui, Saida El Mendili, in Youssef Gahi. 2022. »An Advanced Big Data Quality Framework Based on Weighted Metrics«. *Big Data and Cognitive Computing* 6(4): 153. Dostop 10. januar 2024. https://www.mdpi.com/2504-2289/6/4/153.
60. Eltyeb, Safaa in Naomie Salim. 2014. »Chemical named entities recognition: a review on approaches and applications«. *Journal of Cheminformatics* 2014: 6–17. Dostop 10. januar 2024. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4022577/.
61. *EN-core-web-lg*. 2023. Dostop 21. december 2023. https://spacy.io/models/en#en_core_web_lg.
62. *EN-core-web-sm*. 2023. Dostop 21. december 2023. https://spacy.io/models/en#en_core_web_sm.
63. European Commission. 2014. »Open Data Support«. Dostop 10. januar 2024. https://www.europeandataportal.eu/sites/default/files/d2.1.2_training_module_2.2_open_data_quality_sl_edp.pdf.
64. Ficek, Aleksander, Fangyu Liu, Nigel Collier. 2022. »How to tackle an emerging topic? Combining strong and weak labels for Covid news NER«. Dostop 10. januar 2024. https://www.researchgate.net/publication/364110075_How_to_tackle_an_emerging_topic_Combining_strong_and_weak_labels_for_Covid_news_NER.
65. Finkel, Jenny Rose, Trond Grenager in Christopher Manning. 2005. »Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling«. V

Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005), 363–370. Dostop 10. januar 2024. https://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf.
66. Foley. 2024. »What to Expect in Evolving U.S. Regulation of Artificial Intelligence in 2024«. Dostop 10. januar 2024. https://www.foley.com/insights/publications/2023/12/us-regulation-artificial-intelligence-2024/.
67. Gao, Yang, Yue Xu in Yuefeng Li. 2015. »Pattern-based Topics for Document Modelling in Information Filtering«. *Ieee Transactions on Knowledge and Data Engineering* 27(6): 1629-1642. Dostop 10. januar 2024. https://eprints.qut.edu.au/66686/15/ Pattern-based_Topic_Models_for_Information_Filtering_.pdf.
68. Garcia-Pablos, Aitor, Montse Cuadros in German Rigau. 2018. »W2VLDA: Almost unsupervised system for Aspect Based Sentiment Analysis«. *Expert Systems with Applications* 91: 127–137. Dostop 10. januar 2024. https://arxiv.org/abs/1705.07687.
69. Genuine Impact. 2023. »A brief history of … Artificial Intelligence«. Dostop 10. januar 2024. https://substack.com/profile/42660574-genuine-impact/note/c-14871872.
70. Giménez-Chornet, Vicent. 2017. »Ethics and Social Responsibility in Archival Institutions: Elements to Consider«. *El Profesional De La Información* 26(4): 765–770. Dostop 10. januar 2024. https://www.researchgate.net/publication/318879007_Ethics_and_social_responsibility_in_archival_institutions_Elements_to_consider.
71. González-Carvajal, Santiago in Eduardo C. Garrido-Merchán. 2023. »Comparing BERT against traditional machine learning text classification«. *Journal of Computational and Cognitive Engineering* 2: 352-356. Dostop 10. januar 2024. https://arxiv.org/abs/2005.13012.
72. Gui, Min, Zhengkun Zhang, Zhenlu Yang, Yanhui Gu in Guandong Xu. 2018. »An Effective Joint Framework for Document Summarization«. V *27th World Wide Web (WWW) Conference, Apr 23–27*: 121-122. Dostop 10. januar 2024. https://dl.acm.org/doi/abs/10.1145/3184558.3186959.
73. Guid, Nikola in Damjan Strnad. 2007. *Umetna inteligenca*. Maribor: Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru.
74. Gupta, Anshul, Sunil Kr. Singh in Muskaan Chopra. 2024. »Impact of artificial intelligence and the internet of things in modern times and hereafter: an investigative analysis«. V *Advanced computer science applications – Recent Trends in AI, Machine Learning, and Network Security,* ur. Karan Singh, Latha Banda in Masniha Manjul. Burlington: Apple Academic Press, Inc.
75. Hagendorff, Thilo. 2021. »Linking Human And Machine Behavior: A New Approach to Evaluate Training Data Quality for Beneficial Machine Learning«. *Minds and Machines* 31(4): 563–593. Dostop 10. januar 2024. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8475847/.
76. Hajtnik, Tatjana in Aida Škoro Babić. 2018. »Ali nam pri vrednotenju in odbiranju elektronskega gradiva pomaga tehnologija«. *Moderna arhivistika* 2018(1). Dostop 10. januar 2024. https http://www.pokarh-mb.si/uploaded/ datoteke/radenci_20181/1_2018_169-196_%C5%A0koro.pdf.

77. Hartmann, Jochen, Juliana Huppertz, Christina Schamp in Mark Heitmann. 2019. »Comparing automated text classification methods«. *International Journal of Research in Marketing* 36(1): 20–38. Dostop 10. januar 2024. https://www.sciencedirect.com/science/article/pii/S0167811618300545.
78. Hashimoto, Kazuma, Georgios Kontonatsios, Makoto Miwa in Sophia Ananiadou. 2016. »Topic detection using paragraph vectors to support active learning in systematic reviews«. *Journal of Biomedical Informatics* 62: 59–65. Dostop 10. januar 2024. https://www.sciencedirect.com/science/article/pii/S1532046416300442.
79. Hlede, Irena. 2009. »Umetna inteligenca«. *Klik* 113: 12−14.
80. Holterhoff, Kate. 2017. »From Disclaimer to Critique: Race and the Digital Image Archivist«. *Digital Humanities Quarterly* 11(3). Dostop 10. januar 2024. https://www.digitalhumanities.org/dhq/vol/11/3/000324/000324.html.
81. Howard, Jeremy in Sebastian Ruder. 2018. »Universal language model fine-tuning for text classification«. V *Proceedings of the 56th Annual Meeting of the Assossiation for Computational Linguistics: Volume 1, Long Papers*, 328–339. Melbourne: Assossiation for Computational Linguistics. Dostop 10. januar 2024. https://aclanthology.org/P18-1031/.
82. *HuggingFace*. 2023. Dostop 1. januar 2023. https://huggingface.co/.
83. ISAD(G). 2000. *General International Standard Archival Description. Second edition*. Ottawa: International council on archives.
84. ISAAR (CPF). 2004. *International Standard Archival Authority Record for Corporate Bodies, Persons and Families*. Ottawa: International council on archives.
85. ISDF. 2007. *International Standard for Describing Functions*. Ottawa: International council on archives.
86. ISDIAH. 2008. *International Standard for Describing Institutions with Archival Holdings*. Ottawa: International council on archives.
87. Jaillant, Lise. 2022. »How can we make born-digital and digitised archives more accessible«. *Arch Sci* 22, 417–436. Dostop 12. januar 2024. https://link.springer.com/article/10.1007/s10502-022-09390-7.
88. Jelodar, Hamed, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li in Liang Zhao. 2019. »Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey«. *Multimedia Tools and Applications* 78(11): 15169–15211. Dostop 10. januar 2024. https://arxiv.org/abs/1711.04305.
89. Jolliffe, Ian T. in Jorge Cadima. 2016. »Principal component analysis: a review andrecent developments«. *Phil. Trans. R. Soc. A.* 374: 20150202. Dostop 4. januar 2024. http://doi.org/10.1098/rsta.2015.0202.
90. Jordan, M. I in T. M. Mitchell. 2015. »Machine learning: Trends, perspectives, and prospects«. *Science* 349(6245). Dostop 10. januar 2024. https://www.cs.cmu.edu/~tom/pubs/Science-ML-2015.pdf.
91. Joulin, Armand, Edouard Grave, Piotr Bojanowski in Tomas Mikolov. 2017. »Bag of tricks for efficient text classification«. V *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 427–431. Valencia: Assossiation for Computational Linguistics. Dostop 10. januar 2024. https://aclanthology.org/E17-2068/.

92. *KAGGLE*. 2024. Dostop 1. januar 2024. https://www.kaggle.com/.
93. Kanade, Vijay. 2022. »What Is a Support Vector Machine? Working, Types, and Examples«. Dostop 1. januar 2024. https://www.spiceworks.com/tech/big-data/articles/what-is-support-vector-machine/.
94. Kaneko, Masahiro in Danushka Bollegala. 2021. »Unmasking the Mask – Evaluating Social Biases in Masked Language Models«. Dostop 2. januar 2024. https://arxiv.org/abs/2104.07496.
95. Karabadji, Nour El Islem, Hassina Seridi, Fouad Bousetouane, Wajdi Dhifli in Sabeur Aridhi. 2017. »An evolutionary scheme for decision tree construction«. *Knowledge-Based Systems* 119: 166–177. Dostop 10. januar 2024. https://inria.hal.science/hal-01574079/document.
96. Karjule, Vivek, Jayesh Dange, Sneha Thange, Janhavi Sase in Naina Kokate. 2023. »A Survey on Text Summarization Techniques«. *Journal of Natural Language Processing*. Dostop 5. januar 2024. https://www.researchgate.net/publication/375120640_A_Survey_on_Text_Summarization_Techniques.
97. Khanzode, Chhaya A. in Ravindra D. Sarode. 2020. »Advantages and disadvantages of artificial intelligence and machine learning: a literature review«. *International Journal of Library & Information Science (IJLIS)* 9(1). Dostop 10. januar 2024. https://iaeme.com/MasterAdmin/Journal_uploads/IJLIS/VOLUME_9_ISSUE_1/IJLIS_09_01_004.pdf.
98. Khurana, Diksha, Aditya Koli, Kiran Khatter in Sukhdev Singh. 2022. »Natural Language Processing: State of The Art, Current Trends and Challenges«. Multimedia Tools and Applications 82: 3713–3744. Dostop 10. januar 2024. https://www.researchgate.net/publication/319164243_Natural_Language_Processing_State_of_The_Art_Current_Trends_and_Challenges.
99. *KNIME*. 2024. Dostop 14. januar 2024. https://www.knime.com/.
100. Kowsari, Kamran, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes in Donald Brown. 2019. »Text classification algorithms: A survey«. *Information (Switzerland)* 10(4). Dostop 10. januar 2024. https://www.mdpi.com/2078-2489/10/4/150.
101. Kreimeyer, Kory, Matthew Foster, Abhishek Pandey, Nina Arya, Gwendolyn Halford, Sandra F Jones, Richard Forshee, Mark Walderhaug in Taxiarchis Botsis. 2017. »Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review«. *Journal of Biomedical Informatics* 73: 14–29. Dostop 10. januar 2024. https://www.sciencedirect.com/ science/article/pii/S1532046417301685.
102. Kreutzer, Ralf T. in Marie Sirrenberg. 2020. *Understanding Artificial Intelligence Fundamentals, Use Cases and Methods for a Corporate AI Journey*. Switzerland: Springer Nature AG.
103. Kumar, Puneet, Vinod Kumar Jain in Dharminder Kumar. 2021. *Artifcial Intelligence and Global Society.* Boca Raton: *Taylor & Francis Group, LLC.*
104. Kurian, K. Sheena in Sheena Mathew. 2020. »Survey of scientific document summarization methods«. *Computer Science* 21(2): 139–177. Dostop 10. januar 2024. https://journals.agh.edu.pl/csci/article/view/3356.

105. Lake, Brenden M., Tomer D. Ullman, Joshua B. Tenenbaum in Samuel J. Gershman. 2017. »Building machines that learn and think like people«. *Behavioral and Brain Sciences* 40: 25. Dostop 10. januar 2024. https://doi.org/10.1017/S0140525X16001837.
106. Leaman, Robert in Zhiyong Lu. 2016. »TaggerOne: joint named entity recognition and normalization with semi-Markov Models«. *Bioinformatics* 32(18): 2839–2846. Dostop 10. januar 2024. https://academic.oup.com/bioinformatics/article/32/18/2839/1744190.
107. Li, Jing, Aixin Sun, Jianglei Han in Chenliang Li. 2022. »A Survey on Deep Learning for Named Entity Recognition«. *Ieee Transactions on Knowledge and Data Engineering* 34(1): 50–70. Dostop 20. januar 2024. http://www.li-jing.com/papers/22nersurvey.pdf.
108. Liu, Bo, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi in Zihuai Lin. 2022. »When Machine Learning Meets Privacy: A Survey and Outlook«. *ACM Computing Surveys* 54(2): Article 31. Dostop 12. januar 2024. https://www.semanticscholar.org/reader/b293e4659e20815bcf0b6d31ce46b8bd9437c1fa.
109. Liu, Yang, Jian-Wu Bi in Zhi-Ping Fan. 2017. »Multi-class sentiment classification: The experimental comparisons of feature selection and machine learning algorithms«. *Expert Systems with Applications* 80: 323–339. Dostop 05. januar 2024. https://www.sciencedirect.com/science/article/pii/S0957417417301951.
110. Liu, Zengjian, Ming Yang, Xiaolong Wang, Qingcai Chen, Buzhou Tang, Zhe Wang in Hua Xu. 2017. »Entity recognition from clinical texts via recurrent neural network«. *BMC Medical Informatics and Decision Making* 17. Dostop 06. januar 2024. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5506598/pdf/12911_2017_Article_468.pdf.
111. Ljubešić, Nikola, Marija Stupar, Tereza Jurić. 2012. »Building Named Entity Recognition Models for Croatian and Slovene«. *Jezikovne tehnologije*. Dostop 20. december 2021. https://nl.ijs.si/isjt12/proceedings/isjt2012_24.pdf.
112. Luo, Wei, Dinh Phung, Truyen Tran, Sunil Gupta, Santu Rana, Chandan Karmakar, Alistair Shilton, John Yearwood, Nevenka Dimitrova, Tu Bao Ho, Svetha Venkatesh in Michael Berk. 2016. »Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View«. *Journal of Medical Internet Research* 18(12): 10. Dostop 05. januar 2024. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5238707/.
113. Luo, Xiaoyu. 2021. »Efficient English text classification using selected Machine Learning Techniques«. *Alexandria Engineering Journal* 60(3): 3401–3409. Dostop 05. januar 2024. https://www.sciencedirect.com/science/article/pii/S1110016821000806.
114. Maaten, Laurens van der in Geoffrey Hinton. 2008. »Visualizing Data using t-SNE«. *Journal of Machine Learning Research* 9: 2579–2605. Dostop 05. januar 2024. https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf.
115. Mayhew, Stephen, Chen-Tse Tsai in Dan Roth. 2017. »Cheap translation for cross-lingual named entity recognition«. V *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2536–2545. Copenhagen: Assossiation for

Computational Linguistics. Dostop 07. januar 2024. https://aclanthology.org/D17-1269/.
116. Mazzi, Francesca, Mariarosaria Taddeo in Luciano Floridi. 2023. AI in Support of the SDGs: Six Recurring Challenges and Related Opportunities Identifed Through Use Cases. V *The Ethics of Artificial Intelligence for the Sustainable Development Goals,* ur. Francesca Mazzi in Luciano Floridi. Switzerland: Springer Nature AG.
117. McMullen, Matthew. 2023. »11 NLP Use Cases: Putting the Language Comprehension Tech to Work«. Dostop 10. januar 2024. https://readwrite.com/11-nlp-use-cases-putting-the-language-comprehension-tech-to-work/.
118. Medvedeva, Masha, Michel Vols in Martijn Wieling. 2020. »Using machine learning to predict decisions of the European Court of Human Rights«. *Artificial Intelligence and Law* 28(2): 237–266. Dostop 07. januar 2024. https://link.springer.com/ article/10.1007/s10506-019-09255-y.
119. Miller, Tim. 2019. »Explanation in artificial intelligence: Insights from the social sciences«. *Artificial Intelligence* 267: 1–38. Dostop 07. januar 2024. https://www.sciencedirect.com/science/article/pii/S0004370218305988.
120. Milovanović, Miroslav. 2020. *Klasifikacija elektronske pošte z uporabo umetne inteligence*: *magistrsko delo študijskega programa druge bolonjske stopnje Arhivistika in dokumentologija*. Maribor: Alma Mater Europaea – ECM.
121. Morley, Jessica, Luciano Floridi, Libby Kinsey in Anat Elhalal. 2020. »From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices«. *Science and Engineering Ethics* 26(4): 2141–2168. Dostop 07. januar 2024. https://link.springer.com/article/10.1007/s11948-019-00165-5.
122. Murty, M. N. in M. Avinash. 2023. *Representation in Machine Learning*. Singapore: Springer Nature Singapore Pte Ltd.
123. *Nestrukturirani podatki*. 2024. Dostop 10. januar 2024. https://www.opentext.com/what-is/unstructured-data.
124. Ni, Jian, Georgiana Dinu in Radu Florian. 2017. »Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection«. V *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1470–1480. Vancouver: Assossiation for Computational Linguistics. Dostop 07. januar 2024. https://aclanthology.org/P17-1135/.
125. Nichols, James A, Herbert Chan HW, Baker Mattehew AB. 2019. »Machine learning: applications of artificial intelligence to imaging and diagnosis«. *Biophys Rev* 11(1): 111–118. Dostop 07. januar 2024. https://www.ncbi.nlm.nih.gov/ pmc/articles/ PMC6381354/.
126. Novak, Miroslav. 2015. »Vrednotenje arhivskega gradiva in ponovna uporaba informacij v javnem sektorju«. *Atlanti* 25(1): 53–59. Dostop 07. januar 2024. https://journal.almamater.si/index.php/Atlanti/article/download/108/97.
127. Nozza, Debora, Pikakshi Manchanda, Elisabetta Fersini, Matteo Palmonari in Enza Messina. 2021. »LearningToAdapt with word embeddings: Domain adaptation of Named Entity Recognition systems«. *Information Processing & Management* 58(3): 21.

Dostop 07. januar 2024. https://www.deboranozza.com/ publication/ 2021_learningtoadapt/ 2021_learningtoadapt.pdf.
128. *NVIDIA*. 2023. Dostop 19. december 2023. https://www.nvidia.com/en-gb/geforce/technologies/cuda/technology/.
129. Onan, Aytug. 2018. »Biomedical Text Categorization Based on Ensemble Pruning and Optimized Topic Modelling«. *Computational and Mathematical Methods in Medicine* 2018: 22. Dostop 21. januar 2024. https://www.hindawi.com/journals/ cmmm/2018/2497471/.
130. Onan, Aytug. 2019. »Two-Stage Topic Extraction Model for Bibliometric Data Analysis Based on Word Embeddings and Clustering«. *Ieee Access* 7: 145614–145633. Dostop 21. januar 2024 https://ieeexplore.ieee.org/document/8861084.
131. Padilla, Thomas. 2019. »Responsible Operations: Data Science, Machine Learning, and AI in Libraries«. *OH: OCLC Research*. Dostop 1. januar 2024. https://www.oclc.org/research/publications/2019/oclcresearch-responsible-operations-data-science-machine-learning-ai.html.
132. Patel, Chirag in Mahesh Gadhavi. 2017. »A Model for Document Classification using Kernel Discriminant Analysis (KDA) and Semantic Analysis«. *International Journal of Advanced Research in Computer Science* 8(3). Dostop 21. januar 2024 https://www.ijarcs.info/index.php/Ijarcs/article/view/3097.
133. Peters, Matthew E., Waleed Ammar, Chandra Bhagavatula in Russell Power. 2017. »Semi-supervised sequence tagging with bidirectional language models«. V *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1756–1765. Vancouver: Assossiation for Computational Linguistics. Dostop 21. januar 2024 https://aclanthology.org/P17-1161/.
134. Ploj, Bojan. 2013. *Metoda mejnih parov za učenje umetnih nevronskih mrež*. Maribor: Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru.
135. Popovici, Bogdan Florin. 2021. »Is there a place for records systems in archival descruption«. *Moderna arhivistika* 2021(1). Dostop 15. januar 2024 http://www.pokarh-mb.si/uploaded/datoteke/2021/06%20Popovici_2021.pdf.
136. *ProQuest*. 2024. Dostop 1. januar 2024. https://www.proquest.com/.
137. *Python*. 2024. Dostop 14. januar 2024. https://www.python.org/.
138. Rajendran, Kartik, Pallavi Kasar, Shrushti Kharche in Amruta Aphale. 2023. »Text Summarization«. *International journal of novel research and development*. Dostop 15. januar 2024 https://www.ijnrd.org/papers/IJNRD2304724.pdf.
139. Rāts, Juris in Inguna Pede. 2020. »Document Capturing Automation: Case Study«. *Baltic Journal of Modern Computing* 8(2): 259–274. Dostop 15. januar 2024 https://www.bjmc.lu.lv/fileadmin/user_upload/lu_portal/projekti/bjmc/Contents/8_2_04_Rats.pdf.
140. RIC. 2023. *Records in Contexts – Conceptual Model*. Ottawa: International council on archives.
141. *Roberta-large*. 2023. Dostop 19. december 2023. https://huggingface.co/roberta-large.
142. Rothman, Denis. 2018. *Artificial Intelligence By Example*. Birmingham: Packt Publishing Ltd.

143. Rožanec, Alenka. 2017. *Baze podatkov*. Novo mesto: Fakulteta za upravljanje, poslovanje in informatiko.
144. Runardotter, Mari, Christina Mörtberg in Anita Mirijamdotter. 2011. »The Changing Nature of Archives: Whose Responsibility?«. *EJEG. Electronic Journal of E-Government* 9(1): 68–78. Dostop 15. januar 2024 https://academic-publishing.org/index.php/ejeg/article/download/545/508/541.
145. Russell, J. Stuart in Peter Norvig. 2010. *Artificial Intelligence A Modern Approach*. New Yersey: Pearson Education, Inc.
146. Salazar, Julian, Davis Liang, Toan Q. Nguyen in Katrin Krichoff. 2020. »Masked Language Model Scoring«. V *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* 2699–2712. Dostop 22. januar 2024 https://aclanthology.org/2020.acl-main.240/.
147. Sarker, Iqbal H. 2021. »Machine Learning: Algorithms, Real-World Applications and Research Directions«. *SN Computer Science* 2021(2): 160. Dostop 7. januar 2024 https://link.springer.com/article/10.1007/s42979-021-00592-x.
148. Schmidpeter, Rene in Christophe Funk. 2023. »Artificial Intelligence: Companion to a New Human "Measure"?«. V *Responsible Artificial Intelligence – Challenges for Sustainable Management,* ur. Rene Schmidpeter in Reinhard Altenburger. Switzerland Springer Nature AG.
149. *SCOPUS*. 2024. Dostop 1. januar 2024. https://www.scopus.com/.
150. Shafiq, Nida, Hamid Isma, Asif Muhammas, Nawaz Qamar, Aljuaid Hanan in Ali Hamid. 2023. Abstractive text summarization of low-resourced languages using deep learning. *PeerJ Computer Science* 9: e1176. Dostop 1. januar 2022. https://peerj.com/articles/cs-1176/.
151. Sharif, Omar. Mohammed Moshiul Hoque, A. S. M. Kayes, Raza Nowrozy in Iqbal H. Sarker. 2020. »Detecting Suspicious Texts Using Machine Learning Techniques«. *Applied Sciences-Basel* 10(18): 23. Dostop 7. januar 2024. https://www.mdpi.com/2076-3417/10/18/6527.
152. Siau, Keng in Weiyu Wang. 2020. »Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI«. *Journal of Database Management* 31(12). Dostop 10. januar 2024. https://www.researchgate.net/figure/AI-Ethics-Framework-of-building-ethical-AI_fig1_340115931.
153. *SL_core_news_lg*. 2023. Dostop 21. december 2023. https://spacy.io/models/sl#sl_core_news_lg.
154. Slovar slovenskega knjižnega jezika. 2020. »*Klasifikacija*«. Dostop 5. januar 2024. http://www.fran.si/iskanje?FilteredDictionaryIds=130&View=1&Query=klasifikacija.
155. *SPACY*. 2023. »EntityRecognizer«. Dostop 18. December 2023. https://spacy.io/api/entityrecognizer.
156. Stančić, Hrvoje. 2018. »Computational archival science«. *Moderna arhivistika* 2018(2). Dostop 15. januar 2024. http://www.pokarh-mb.si/uploaded/datoteke/2_2018_323-330_stancic.pdf.
157. Suleiman, Dina in Arafat Awajan. 2020. »Deep Learning Based Abstractive Text Summarization: Approaches, Datasets, Evaluation Measures, and Challenges«.

Mathematical Problems in Engineering 2020. Dostop 6. januar 2024. https://www.hindawi.com/journals/mpe/2020/9365340/.
158. Sun, Lin, Fule Ji, Kai Zhang in Chi Wang. 2019. »Multilayer ToI Detection Approach for Nested NER«. *Ieee Access* 7: 186600–186608. Dostop 6. januar 2024. https://ieeexplore.ieee.org/document/8937532.
159. Sun, Wencheng, Zhiping Cai, Yangyang Li, Fang Liu, Shengqun Fang in Guoyan Wang. 2018. »Data processing and text mining technologies on electronic medical records: A review«. *Journal of Healthcare Engineering* 2018. Dostop 14. januar 2024. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5911323/.
160. Suominen, Arho in Hannes Toivanen. 2016. »Map of Science with Topic Modeling: Comparison of Unsupervised Learning and Human-Assigned Subject Classification«. *Journal of the Association for Information Science and Technology* 67(10): 2464–2476. Dostop 14. januar 2024. https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.23596.
161. Štajner, Tadej, Tomaž Erjavec in Simon Krek. 2013. »Razpoznavanje imenskih entitet v slovenskem besedilu«. *Slovenščina 2.0*, 1(2): 58–81. Dostop 20. december 2021. http://slovenscina2.0.trojina.si/arhiv/2013/2/Slo2.0_2013_2_04.pdf.
162. Tabak, Feride Savaroglu in Vesile Evrim. 2020. »Event-based summarization of news articles«. *Turkish Journal of Electrical Engineering and Computer Sciences* 28(2): 850–864. Dostop 12. januar 2024. https://journals.tubitak.gov.tr/cgi/viewcontent.cgi?article=1387&context=elektrik.
163. Taleb, Ikbal, Mohamed Adel Serhani, Chafik Bouhaddioui in Rachida Dssouli. 2021. »Big data quality framework: a holistic approach to continuous quality management«. *Journal of Big Data* 8: 76. Dostop 23. januar 2024. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00468-0.
164. Telefonica. 2024. »A fit for purpose and borderless European Artificial Intelligence Regulation«. Dostop 10. januar 2024. https://www.telefonica.com/en/communication-room/blog/a-fit-for-purpose-and-borderless-european-artificial-intelligence-regulation/.
165. *Tensorflo*w. 2023. Dostop 1. januar 2023. https://www.tensorflow.org/.
166. *The Guardian*. 2022. Dostop 2. februar 2022. https://www.theguardian.com/.
167. *The Guardian Open Platform*. 2022. Dostop 02. februar 2022. https://open-platform.theguardian.com/.
168. Tian, Zhi, Weilin Huang, Tong He, Pan He in Yu Qiao. 2016. »Detecting Text in Natural Image with Connectionist Text Proposal Network«. V *14th European Conference on Computer Vision (ECCV)*. Dostop 15. september 2024. https://arxiv.org/pdf/1609.03605.pdf.
169. Trajanov, Dimitar, Gorgi Lazarev, Lubomir Chitkushev in Irena Vodenska. 2023. »Comparing the performance of ChatGPT and state-of-the-art climate NLP models on climate-related text classification tasks«. *E3S Web of Conferences*. Dostop 14. januar 2024. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00468-0.
170. TXT. 2023. »Text document«. Dostop 15. september 2023. https://docs.fileformat.com/word-processing/txt/.

171. *Uredba evropskega parlamenta in sveta o določitvi harmoniziranih pravil o umetni inteligenci – predlog*. 2022. Dostop 10. januar 2024. https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/en/pdf.
172. *Uredba o varstvu dokumentarnega in arhivskega gradiva ter arhivih*. Uradni list RS, št. 42/17.
173. Vinyals, Oriol, Alexander Toshev, Samy Bengio in Dumitru Erhan. 2015. »Show and Tell: A Neural Image Caption Generator«. V *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3156-3164. Dostop 15. januar 2024. https://ieeexplore.ieee.org/document/7298935.
174. Vollmer, Sebastian, Bilal A. Mateen, Gergo Bohner, Franz J. Király, Rayid Ghani, Pall Jonsson, Sarah Cumbers, Adrian Jonas, Katherine S. L. McAllister, Puja Myles, David Grainger, Mark Birse, Richard Branson, Karel G. M. Moons, Gary S. Collins, John P. A. Ioannidis, Chris Holmes in Harry Hemingway. 2020. »Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness«. *The BMJ* 368. Dostop 07. januar 2024. https://www.bmj.com/content/bmj/368/bmj.l6927.full.pdf.
175. Wagh, Kishor, Aishwarya Kulkarni, Pratiksha Pawar, Neha Kirange in Shraddha Kashid. (2017). »Bio Medical Named Entity Recognition Using Machine Learning Algorithms«. *International Journal of Advance Engineering and Research Development. Special Issue on Recent Trends in Data Engineering* 4(Special Issue 5). Dostop 09. januar 2024. https://www.researchgate.net/publication/347902664_Bio_Medical_Named_Entity_Recognition_Using_Machine_Learning_Algorithms.
176. Wang, Yanshan, Sunghwan Sohn, Sijia Liu, Feichen Shen, Liwei Wang, Elizabeth J. Atkinson, Shreyasee Amin in Hongfang Liu. 2019. »A clinical text classification paradigm using weak supervision and deep representation«. *Bmc Medical Informatics and Decision Making* 19: 13. Dostop 07. januar 2024. https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-018-0723-6.
177. *WEB OF SCIENCE*. 2024. Dostop 1. januar 2024. https://www.webofscience.com/.
178. Weston, Leigh, Vahe Tshitoyan, John Dagdelen, Olga Kononova, Kristin Persson, Gerbrand Ceder in Anubhav Jain. 2019. »Named Entity Recognition and Normalization Applied to Large-Scale Information Extraction from the Materials Science Literature«. *Journal of Chemical Information and Modeling* 59(9): 3692–3702. Dostop 08. januar 2024. https://chemrxiv.org/engage/chemrxiv/article-details/60c7422a9abda26972f8bf77.
179. Wooldridge, Michael. 2021. *A Brief History of Artifcial Intelligence – What It Is, Where We Are and Where We Are Going*. New York: Flatiron Books.
180. Wu, Hongping, Yuling Liu in Jingwen Wang. 2020. »Review of Text Classification Methods on Deep Learning«. *Cmc-Computers Materials & Continua* 63(3): 1309–1321. Dostop 07. januar 2024. https://www.techscience.com/cmc/v63n3/38877.
181. Xu, Mingbin, Hui Jiang in Sedtawut Watcharawittayakul. 2017. »A local detection approach for named entity recognition and mention detection«. V *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

Papers), 1237–1247. Vancouver: Assossiation for Computational Linguistics. Dostop 07. januar 2024. https://aclanthology.org/P17-1114/.

182. Yang, Zichao, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola in Eduard Hovy. 2016. »Hierarchical attention networks for document classification«. V *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1480–1489. San Diego: Assossiation for Computational Linguistics. Dostop 14. januar 2024. https://aclanthology.org/N16-1174/.
183. Zajšek, Boštjan in Miroslav Novak. 2021. »Arhivski strokovni izzivi dolgoročne hrambe elektronskih sporočil«. *Moderna arhivistika* 2021(2). Dostop 07. januar 2024. http://www.pokarh-mb.si/uploaded/datoteke/2021/07%20Zajsek_2021.pdf.
184. *Zakon o varstvu dokumentarnega in arhivskega gradiva ter arhivih (ZVDAGA)*. Uradni list RS, št. 30/06 (s poznejšimi spremembami in dopolnitvami).
185. Zannettou, Savvas, Michael Sirivianos, Jeremy Blackburn in Nicolas Kourtellis. 2019. »The Web of False Information: Rumors, Fake News, Hoaxes, Clickbait, and Various Other Shenanigans«. *Acm Journal of Data and Information Quality* 11(3): 37. Dostop 07. januar 2024. https://www.researchgate.net/publication/324435607_The_Web_of_False_Information_Rumors_Fake_News_Hoaxes_Clickbait_and_Various_Other_Shenanigans.
186. Zhao, Weizhong, James J Chen, Roger Perkins, Zhichao Liu, Weigong Ge, Yijun Ding in Wen Zou. 2015. »A heuristic approach to determine an appropriate number of topics in topic modeling«. *Bmc Bioinformatics* 16: 10. Dostop 07. januar 2024. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-16-S13-S8.
187. Zhou, Ran, Ruidan He, Xin Li, Lidong Bing, Erik Cambria, Luo Si in Chunyan Miao. 2022. »MELM: Data Augmentation with Masked Entity Language Modeling for Cross-lingual NER«. Dostop 1. januar 2022. https://arxiv.org/abs/2108.13655v1.
188. Žagar, Aleš in Marko Robnik Šikonja. 2021. Unsupervised Approach to Multilingual User Comments Summarization. V *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, 89–98. Dostop 1. januar 2024. https://aclanthology.org/2021.hackashop-1.13/ .

IZJAVA O AVTORSTVU

ALMA MATER EUROPAEA
ECM

07

IZJAVA O AVTORSKEM DELU IN ISTOVETNOSTI TISKANE IN ELEKTRONSKE VERZIJE ZAKLJUČNEGA DELA

Priimek in ime študenta	Miroslav Milovanović
Vpisna številka	31203023
Študijski program	Arhivske znanosti
Naslov zaključnega dela:	Model urejanja in popisovanja nestrukturiranih besedil z uporabo strojnega učenja
Naslov v angleščini:	Model of arrangement and description of unstructured texts using machine learning
Mentor:	izr. prof. dr. Miroslav Novak
Somentor:	/
Mentor iz podjetja:	/

S podpisom izjavljam da:
- Je predloženo zaključno delo z naslovom Model urejanja in popisovanja nestrukturiranih besedil z uporabo strojnega učenja izključno rezultat mojega lastnega raziskovalnega dela,
- Sem poskrbel/a da so dela in mnenja drugih avtorjev, ki jih uporabljam v predloženem delu navedena oz. citirana v skladu s fakultetnimi navodili,
- Se zavedam, da je plagiatorstvo – predstavljanje tujih del, bodisi v obliki citata, bodisi v obliki dobesednega parafraziranja, bodisi v grafični obliki, s katerim so tuje misli oziroma ideje predstavljene kot moje lastne, kaznivo po zakonu (Zakon o avtorskih in sorodnih pravicah, UrL RS št. 139/2006 s spremembami),
- V primeru kršitve zgoraj navedenega zakona prevzemam vso moralno, kazensko in odškodninsko odgovornost,

Podpisani-a Miroslav Milovanović izjavljam, da sem za potrebe arhiviranja oddal/a elektronsko verzijo zaključnega dela v Digitalno knjižnico. Zaključno delo sem izdelal-a sam-a ob pomoči mentorja. V skladu s 1. odstavkom 21. člena Zakona o avtorskih in sorodnih pravicah (Uradni list RS, št. 16/2007) dovoljujem, da se zgoraj navedeno zaključno delo objavi na portalu Digitalne knjižnice. Prav tako dovoljujem objavo osebnih podatkov vezanih na zaključek študija (ime, priimek, leto in kraj rojstva, datum diplomiranja, naslov diplomskega dela) na spletnih straneh in v publikacijah Alma Mater.

Tiskana verzija zaključnega dela je istovetna elektronski verziji, ki sem jo oddal/a za objavo v Digitalno knjižnico.

Datum in kraj:

28.03.2024, Maribor

Podpis študent/ke:

Slovenska ulica 17, 2000 Maribor, Slovenija
Tel: +386 2 250 19 99 / Fax: +386 2 250 19 98 / E: info@almamater.si
www.almamater.si

IZJAVA LEKTORJA

ALMA MATER EUROPAEA ECM

O6

POTRDILO O LEKTORIRANJU

Podpisani(a)
Ksenija Pečnik,

po izobrazbi (strokovni oz. znanstveni naslov)
profesorica slovenščine,

potrjujem, da sem lektoriral(a) zaključno delo študenta(ke)
Miroslava Milovanovića,

z naslovom:
Model urejanja in popisovanja nestrukturiranih besedil z uporabo strojnega učenja

Kraj: **Orehova vas**

Datum: **27. 3. 2024**

Podpis: 

Slovenska ulica 17, 2000 Maribor, Slovenija
Tel: +386 2 250 19 99 / Fax: +386 2 250 19 98 / E: info@almamater.si
www.almamater.si