UDC: 519.72:007 Info M str. 4-11

UPOREDNA ANALIZA METODA ZA KATEGORIZACIJU TEKSTA
SEMANTIC ANALYSIS OF TEXT

Ana Bojanić, Zoran Đurić

REZIME: Kategorizacija tekstualnih dokumenata upotrebom metoda mašinskog učenja postala je jedna od osnovnih tehnika ekstrakcije i sumarizacije korisnih informacija sadržanih u njima. U radu je opisan proces pripreme tekstualnih dokumenata i analizirani su pristupi nadgledanog i nenadgledanog mašinskog učenja za njihovu kategorizaciju. Evaluirano je pet algoritama na pet standardnih skupova podataka za kategorizaciju teksta. Za većinu posmatranih algoritama, na svim skupovima podataka, preciznost i odziv se kreću u rasponu 70-90%. U pogledu predefinisanih metrika, algoritmi nadgledanog učenja pokazuju bolje rezultate na četiri skupa podataka, dok pristup nenadgledanog učenja daje bolje rezultate na jednom skupu podataka. U radu je naglašena i osnovna prednost pristupa nenadgledanog učenja u odnosu na algoritme nadgledanog učenja i dati su neki od mogućih prijedloga za dalja istraživanja u ovoj oblasti.

KLJUČNE REČI: Kategorizacija teksta, Nadgledano i nenadgledano mašinsko učenje, Mašinsko razumijevanje prirodnih jezika

ABSTRACT: Text categorization using machine learning methods has become one of the key techniques for extracting and summarization of valuable information from text documents. In this paper text pre-processing steps are described, and supervised and unsupervised machine learning approaches for text categorization are analyzed. Five algorithms are evaluated on five standard datasets for text categorization. For majority of applied algorithms, on all datasets, achieved precision and recall are in range 70-90%. In terms of predefined metrics, supervised algorithms perform better on four datasets, while unsupervised approach shows better results on one dataset. Also, main advantage of unsupervised approach comparing to those supervised is emphasized and some possible suggestions for further research in this area are given.

KEY WORDS: Text categorization, Supervised and unsupervised machine learning, Natural Language Processing

1. UVOD

Posljednjih nekoliko godina svjedoci smo hiperprodukcije tekstualnih podataka u elektronskom obliku, koja sa sobom nosi i velike izazove, prije svega skladištenja i organizacije podataka, zatim pretraživanja, a na kraju i njihove efektivne upotrebe. Analiza ogromne količine podataka koji se svakodnevno generišu postala je neophodna za pravovremeno donošenje poslovnih odluka. Kompleksnost obrade i ekstrakcije korisnih informacija iz ogromne količine „sirovih“ podataka uslovila je primjenu metoda mašinskog učenja.

Od samih početaka istraživanja u oblasti mašinskog razumijevanja prirodnih jezika rješava se isti fundamentalni problem: kako predstaviti semantiku jezika na način koji je mašinski interpretabilan i tačan? Kao posljedica kompleksnosti prirodnih jezika mnogobrojni su izazovi u rješavanju ovog problema, a neki od najvećih su dvosmislenost, prenesena značenja, upotreba ironije i sarkazma, tipografske greške, kolokvijalni izrazi i sl. Potrebni su multidisciplinarni pristupi u njegovom rješavanju, koji adresiraju različite jezičke nivoe, od morfoloških, preko sintaksnih do semantičkih, a česte su i upotrebe postojećih *web* baziranih leksikografskih baza i enciklopedija. Najjednostavniji pristupi tretiraju tekst samo kao skup izolavanih riječi bez poretka (eng. *bag-of-words*), dok oni napredniji uzimaju u obzir i korelacije gradivnih elemenata teksta, semantiku i kontekst.

Vremenom su identifikovane i osnovne klase problema koje se rješavaju semantičkom analizom teksta, od kojih su neke: sumarizacija teksta, identifikovanje ključnih riječi i entiteta u tekstu, određivanje kategorija teksta i dominantnih emocija u tekstu.

Savremeni pristupi u rješavanju svih navedenih klasa problema sve više teže adaptivnim i dinamičkim modelima, „svjesnim“ konteksta (eng. *context-aware*) i namjere (eng. *intent-driven*) sadržaja teksta. Naime, smatra se da se budućnost mašinskog razumijevanja prirodnih jezika ogleda u paradigmama koje emuliraju biološke i lingvističke koncepte i načine na koje ljudski mozak procesira prirodne jezike, uzimajući u obzir semantičke atribute koji nisu eksplicitno prisutni u tekstu [1]. Tako se, npr. koncepti fazi (eng. *fuzzy*) logike primjenjuju za sumarizaciju sadržaja, reprezentaciju sadržanog znanja i identifikovanje značenja riječi [2,3,4]. Zatim, upotreba neuronskih mreža sve češće postaje neizostavan dio u rješavanju problema reprezentacije riječi [5], kategorizacije teksta [6,7] i detekcije emocija u tekstu [8,9].

Predmet ovog rada je kategorizacija teksta. Obrađeni su, testirani i upoređeni neki od najčešće korištenih algoritama nadgledanog mašinskog učenja za kategorizaciju teksta, kao i jedan pristup nenadgledanog mašinskog učenja, zasnovan na upotrebi nekih poznatih algoritama iz oblasti mašinskog učenja za obradu teksta i postojećih leksikografskih baza engleskog jezika. Ovi pristupi testirani su na pet standardnih skupova podataka za kategorizaciju teksta i evaluirani i upoređeni u pogledu predefinisanih metrika.

Rad je organizovan u šest poglavlja. U drugom poglavlju opisan je proces kategorizacije teksta, koji podrazumijeva nekoliko koraka: inicijalnu obradu tekstualnih dokumenata (eng. *pre-processing*), ekstrakciju/selekciju atributa (eng. *features*), odabir odgovarajućeg algoritma za klasifikaciju i na kraju evaluaciju rezultata. U trećem poglavlju je dat pregled nekih od najčešće korištenih algoritama nadgledanog učenja iz oblasti mašinskog učenja za kategorizaciju teksta i opisan je primi-

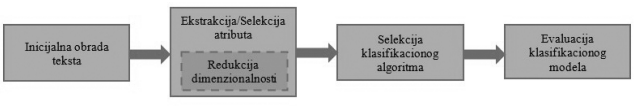
jenjeni pristup nenadgledanog učenja. U četvrtom poglavlju opisani su eksperimentalni skupovi podataka za primjenu ovih pristupa i njihove osnovne karakteristike, kao i proces selekcije optimalnih vrijednosti parametara za odgovarajuće algoritme. U petom poglavlju su prikazani dobijeni rezultati. Šesto poglavlje sadrži zaključke i prijedloge za dalja istraživanja u ovoj oblasti.

2. KATEGORIZACIJA TEKSTA

Kategorizacija teksta podrazumijeva određivanje kategorije teksta, iz skupa predefinisanih kategorija. Odnosno, ako je d_i dokument iz skupa svih dokumenata $D = \{d_1, d_2,...,d_n\}$ i $C = \{c_1, c_2,...,c_m\}$ skup predefinisanih kategorija, onda kategorizacija teksta predstavlja dodijeljivanje jedne ili više kategorija $C_j \subseteq C$ dokumentu d_i. Predmet ovog rada je određivanje jedne kategorije teksta, iz skupa predefinisanih kategorija (eng. *single label classification*) [10,11].

Kategorizacija teksta je jedan od najčešće rješavanih problema u oblasti semantičke analize teksta. Neki od velikog broja slučajeva upotrebe su *web* pretraživanje, filtriranje informacija, sentiment analiza teksta, klasifikacija dokumenata [7,12].

Kategorizacija teksta može da bude nadgledana (eng. *supervised*), nenadgledana (eng. *unsupervised*) i polu-nadgledana (eng. *semi-supervised*) [13]. Modeli nadgledanog mašinskog učenja pokazuju dobre rezultate u ovoj oblasti, ali zahtjevaju manuelno određivanje kategorija velikog broja tekstova za obučavanje modela, pa su često neekonomični, a samim tim i neprihvatljivi. Proces labelisanja tekstova se dodatno usložnjava postojanjem velikog broja kategorija u predefinisanom skupu. Rješenja iz domena polu-nadgledanog mašinskog učenja su manje zahtjevna od onih iz domena nadgledanog mašinskog učenja u pogledu potrebnih resursa za anotaciju korpusa dokumenata za obučavanje modela, ali se to najčešće odražava na performanse takvih rješenja. Iz tih razloga, sve češće, rješenja za kategorizaciju tekstualnih dokumenata se oslanjaju na pristupe nenadgledanog mašinskog učenja. Proces kategorizacije tekstualnih dokumenata može da se dekomponuje na nekoliko faza, koje su prikazane na slici 1 i opisane u nastavku [14].



Slika 1. *Proces kategorizacije tekstualnih dokumenata*

U prvoj i drugoj fazi ovog procesa „sirovi" tekstualni podaci, koji su u opštem slučaju nestruktuirani podaci, se konvertuju u struktuirani ulazni prostor podataka za obučavanje klasifikacionog algoritma.

2.1 *Inicijalna obrada teksta (eng.* pre-processing*)*

Ovo je početni korak u procesu kategorizacije tekstualnih dokumenata i obuhvata nekoliko tehnika koje mogu da se kombinuju na različite načine, u zavisnosti od prirode tekstova u dokumentima koji se obrađuju. Neki, već skoro standardni, tok primjene ovih tehnika obuhvata: tokenizaciju teksta, uklanjanje neinformativnih tokena (eng. *stopwords*), identifikovanje vrsta riječi u rečenicama (eng. *part-of-speech*) i različitih entiteta u tekstu (eng. *named entity recognition*), definisanje načina upotrebe malih i velikih slova u rečenicama, definisanje načina upotrebe skraćenica, te riječi i fraza iz neformalnog jezika, uklanjanje tipografskih grešaka, upotreba *n-gram* konstrukcija, te upotreba tehnika za morfološku normalizaciju.

Čitav proces počinje tokenizacijom, čime se tekst predstavlja kao skup karaktera, riječi i fraza, jednim nazivom tokena, koji imaju neko značenje [14].

Zatim, iz dokumenata mogu da se uklone tokeni koji nemaju veliki značaj i doprinos procesu klasifikacije, a podrazumijevaju najčešće interpunkcijske znakove, specijalne karaktere i riječi koje se veoma često pojavljuju u prirodnom jeziku, pri čemu ne nose neko specifično značenje (eng. *stopwords*) [14].

Dodatno, tekst može da se isfiltrira po različitim kriterijumima, kao što su: zadržavanje/uklanjanje samo nekih vrsta riječi (npr. imenica, glagola, pridjeva), zatim identifikovanje entiteta, kao što su osobe, institucije, događaji, različite vremenske i geografske odrednice, i njihova upotreba na specifične načine.

Takođe, različite konvencije upotrebe malih i velikih slova u rečenicama su jedan od problema koji se rješavaju u ovom koraku. Jedno od čestih rješenja je upotreba svih malih slova u rečenicama [14].

Upotreba skraćenica, kao i riječi i fraza iz neformalnog jezika su takođe problemi koji se adresiraju u ovoj fazi procesa kategorizacije teksta. U nekim pristupima ovakvi oblici se eliminišu, dok se u drugim, različitim metodama, konvertuju u formalne oblike [14].

I tipografske greške, posebno karakteristične za tekstove generisane na različitim socijalnim mrežama, mogu negativno da utiču na proces klasifikacije. Razvijaju se različite tehnike za njihovo uklanjanje. Neke od njih su heš-bazirane (eng. *hashing-based*) i kontekstno-osjetljive (eng. *context-sensitive*) tehnike, kao i upotreba *Trie* i *Damerau-Levenshtein distance* bigram-a [14].

Česta je i upotreba *n-gram* pristupa za generisanje dodatnih atributa teksta. *N-gram* je skup *n* riječi koje se pojavljuju zajedno i u istom redoslijedu u tekstu. Naime, osim reprezentacije teksta kao skupa samo pojedinačnih riječi u njemu, mogu da se koriste i ovakve konstrukcije kao njegovi atributi. Najčešće su to skupovi od dvije (*2-gram*) ili tri (*3-gram*) uzastopne riječi [14].

Takođe, u cilju naglašavanja značenja riječi u njenom osnovnom obliku, veoma često se koriste tehnike za morfološku normalizaciju. Dvije osnovne su korijenovanje (eng. *stemming*) i lematizacija (eng. *lemmatization*). Korijenovanje je postupak koji uklanjanjem afiksa iz različitih oblika riječi, na osnovu predefinisanih pravila, pokušava da identifikuje korijen riječi zajednički za sve oblike, koji sam po sebi ne mora da predstavlja riječ iz rječnika. Lematizacijom se riječ svodi

na kanonski oblik, tj. lemu, koji i dalje predstavlja riječ iz rječnika, pri čemu koristi i informacije o vrsti riječi u rečenicama, pomoću procesa za identifikovanje vrsta riječi u rečenicama (eng. *part-of-speech*) [14].

2.2 *Ekstrakcija/selekcija atributa*

Nakon inicijalnog „prečišćavanja" originalnog teksta u dokumentima, mogu se primijeniti formalne metode za ekstrakciju/selekciju atributa na ulazni skup podataka. Neke od najčešće upotrebljavanih su tehnike bazirane na težinskim faktorima riječi (eng. *weighted-word*), kao što su TF (*Term Frequency*) i TF-IDF (*Term Frequency-Inverse Document Frequency*), kao i tehnike predstavljanja riječi vektorom i njenim ugrađivanjem u vektorski prostor (eng. *word embeddings*) [14], od kojih su neke Word2Vec [15,16], FastText [12,17] i GloVe [18].

Nakon ovog koraka tekstualni dokumenti su predstavljeni kao dokument vektori, odnosno ulazni podaci za obučavanje odgovarajućeg klasifikacionog algoritma. Dodatno, mogu se primijeniti i metode za redukciju dimenzionalnosti ulaznog prostora podataka, a neke od najpoznatijih su: analiza glavnih komponenti (eng. *Principal Component Analysis - PCA*), linearna diskriminaciona analiza (eng. *Linear Discriminant Analysis - LDA*), nenegativna faktorizacija matrice (eng. *Non-Negative Matrix Factorization - NMF*) i vektori t-distribuiranih slučajnih susjeda (eng. *t-Distributed Stochastic Neighbor Embedding* - t-SNE) [14].

Sljedeći koraci u procesu kategorizacije dokumenata su odabir odgovarajućeg klasifikacionog algoritma i evaluacija rezultata na testnom skupu podataka.

2.3 *Selekcija klasifikacionog algoritma*

Za odabir najboljeg klasifikacionog algoritma neophodno je poznavanje samih algoritama, njihovih prednosti i nedostataka, najboljih slučajeva upotrebe, kao i osnovnih karakteristika skupa podataka za obučavanje i testiranje algoritma. Neki od najpoznatijih algoritama za klasifikaciju teksta su: naivni Bajesov klasifikator (eng. *Naïve Bayes Classifier - NBC*), metoda najbližeg centroida (eng. *Rocchio Classifier - RC),* metoda k najbližih susjeda (eng. *K-Nearest Neighbor - KNN*), metoda potpornih vektora (eng. *Support Vector Machine - SVM*), stablo odlučivanja (eng. *Decision Tree - DT*) i šuma slučajnih stabala (eng. *Random Forest - RF*) [8,10,11,14,19,21]. U novije vrijeme koriste se i pristupi dubokog učenja (eng. *deep learning*) koji modeluju kompleksne nelinearne relacije između podataka. Dodatno, nakon izbora algoritma, metodama hiperoptimizacije parametara mogu da se odrede i optimalne vrijednosti parametara algoritma, za dati slučaj upotrebe. U trećem poglavlju su opisani neki od najpoznatijih algoritama za klasifikaciju teksta.

2.4 EVALUACIJA MODELA

Posljednji korak u procesu kategorizacije teksta je evaluacija generisanog modela. Postoje mnoge metrike za evaluaciju rezultata klasifikacije, a najčešće korištene su preciznost (eng. *precision*), odziv (eng. *recall*), tačnost (eng. *accuracy*) i F_β-mjera (F_β-*measure*).

Preciznost predstavlja odnos broja primjeraka za koje je prediktovana ispravna kategorija (TP) i broja svih primjeraka za koje je prediktovana ta kategorija (TP + FP).

Odziv predstavlja odnos broja primjeraka za koje je izvršena predikcija ispravne kategorije (TP) i ukupnog broja primjeraka iz te kategorije (TP + FN).

Ta*čnost* predstavlja odnos broja svih ispravnih predikcija (TP + TN) i broja svih predikcija u testnom skupu podataka (TP + FP + FN + TN).

F_β-mjera predstavlja kombinaciju metrika preciznosti i odziva i definiše se na sljedeći način:

$$F_\beta = \frac{(1 + \beta^2)(precision \times recall)}{\beta^2 \times precision + recall}$$

Za najčešće upotrebljavanu vrijednost parametra $\beta = 1$, jednačina postaje:

$$F_1 = \frac{2TP}{2TP + FP + FN}$$

3. PREGLED NAJČEŠĆE KORIŠTENIH ALGORITAMA ZA KATEGORIZACIJU TEKSTA

U nastavku su opisani neki od najčešće korištenih algoritama nadgledanog mašinskog učenja za kategorizaciju teksta, koji su ujedno testirani i evaluirani u pogledu predefinisanih metrika u ovom radu, na skupovima podataka opisanim u četvrtom poglavlju.

3.1 *Naivni Bajesov klasifikator*

Naivni Bajesov klasifikator je jedan od najčešće korištenih klasifikacionih algoritama, još od pedesetih godina prošlog vijeka. Zasnovan je na Bajesovoj teoremi i pripada grupi probabilističkih algoritama za klasifikaciju [10,14]. Osnovna pretpostavka ovog generativnog modela je da su svi atributi skupa podataka za obučavanje modela nezavisni jedni od drugih, odakle i naziv „*Naïve*" [10]. Za dokument *d* iz korpusa dokumenata rezultat klasifikacije je kategorija iz skupa predefinisanih kategorija sa najvećom *a posteriori* vjerovatnoćom. Bez obzira na to što se radi o relativno jednostavnom algoritmu, njegove performanse u pogledu klasifikacije su uporedive sa mnogo kompleksnijim klasifikacionim algoritmima. Jedno od osnovnih ograničenja jeste upotreba ovog algoritma na skupovima podataka u kojima su atributi međusobno izraženo korelisani [20].

3.2. *Metoda potpornih vektora*

Originalna verzija metode potpornih vektora datira iz 1963. godine, a razvili su je Vapnik i Chervonenkis. Adaptiranu verziju za klasifikaciju podataka koji nisu linearno separabilni

razvio je Boser početkom 1990-tih godina [14]. Algoritam mapira ulazni prostor podataka za obučavanje u novi vektorski prostor veće dimenzionalnosti, u kojem podaci iz trening skupa mogu biti linearno separabilni, a zatim pronalazi optimalno separabilnu hiper-ravan za razdvajanje različitih klasa podataka u tom prostoru. Za mapiranje ulaznih podataka koji nisu linearno separabilni u novi višedimenzionalni vektorski prostor koristi se kernel trik, odn. kernel funkcija [19,22].

3.3 *Stablo odlučivanja*

Stablo odlučivanja kao induktivnu tehniku mašinskog učenja za rješavanje problema klasifikacije je predstavio D. Morgan, a razvio J.R. Quinlan 1980-tih godina. Osnovna ideja ove tehnike je hijerarhijska dekompozicija prostora podataka za treniranje, kreiranjem stabla baziranog na atributima podataka iz trening skupa. Unutrašnji čvorovi stabla predstavljaju atribute podataka, grane predstavljaju težine, a listovi klase podataka iz predefinisanog skupa. Klasifikacija primjerka iz skupa podataka se odvija prolaskom kroz stablo od korijena, preko odgovarajućih čvorova sve do krajnjih listova koji predstavljaju prediktovanu klasu. Ovaj algoritam je zbog brzine, kako obučavanja, tako i predikcije, popularan za rješavanje problema klasifikacije, ali s druge strane je osjetljiv i na male perturbacije u podacima i može lako dovesti do pretreniranja (eng. *overfitting*) [10,14].

3.4 *Šuma slučajnih stabala*

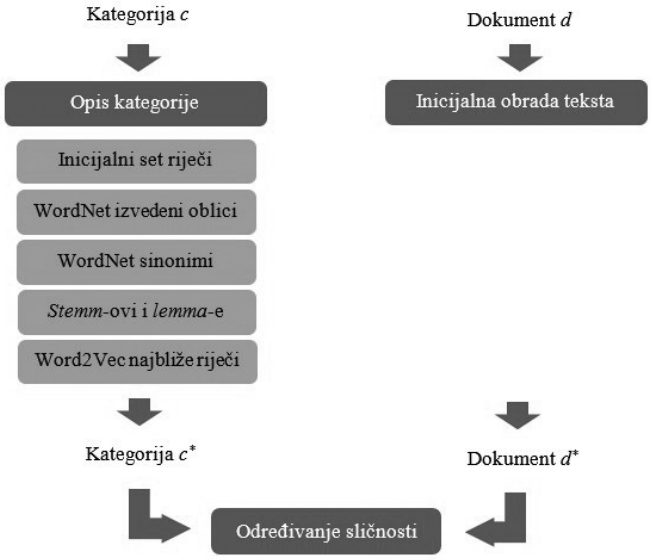
Šuma slučajnih stabala je ansambl metoda mašinskog učenja za klasifikaciju podataka. Predstavio ju je T. Kam Ho 1995. godine, a L. Breiman unaprijedio 1999. godine. Osnovni princip ove metode je paralelno generisanje n nasumičnih (eng. *random*) stabala (eng. *decision trees*). Nakon obučavanja svih stabala, predikcije se određuju metodom glasanja. Jedna od osnovnih prednosti ovog algoritma je brzina obučavanja modela u poređenju sa nekim drugim tehnikama dubokog učenja, ali je proces predikcije znatno sporiji. Smanjivanje vremenske kompleksnosti faze predikcije može da se postigne smanjivanjem broja stabala [14].

3.5 *Jedan pristup nenadgledanog mašinskog učenja za kategorizaciju tekstualnih dokumenata*

U ovom radu primijenjen je i jedan pristup nenadgledanog mašinskog učenja za kategorizaciju tekstualnih dokumenata, po uzoru na pristup iz 2019. godine, koji su Haj-Yahia, Sieg i Deleris [23] razvili i primijenili u domenu upravljanja operativnim rizicima u bankarskom sektoru, odnosno za kategorizaciju širokog spektra incidenata u identifikovane klase rizika. U ovom pristupu se problem kategorizacije teksta modeluje kao problem određivanja sličnosti dva skupa riječi. Prvi skup riječi predstavlja dokument koji se kategorizuje, a drugi skup čine riječi koje opisuju neku kategoriju iz skupa predefinisanih kategorija. Izabranom metodom za određivanje sličnosti tekstualnih dokumenata određuje se kategorija, iz skupa predefinisanih kategorija, najsličnija datom dokumentu iz korpusa dokumenata i proglašava se kategorijom tog dokumenta. Dokumenti iz korpusa se inicijalno obrađuju tehnikama za inicijalnu obradu teksta, a ključne riječi koje opisuju kategorije se određuju u nekoliko koraka.

U ovom radu primijenjen je modifikovani pristup u odnosu na pomenuti. Razlike su prije svega u koracima inicijalne obrade dokumenata, kao i koracima opisivanja kategorija. Pored toga, primijenjena je i dodatna metoda za određivanje sličnosti tekstualnih dokumenata. Uspjeh ovakvog pristupa prevashodno zavisi od opisa predefinisanih kategorija, pa je akcenat čitavog pristupa upravo na tom koraku.

Na slici 2 su prikazani osnovni koraci primijenjenog pristupa nenadgledanog mašinskog učenja u ovom radu.



Slika 2. *Koraci primijenjenog pristupa nenadgledanog učenja*

3.5.1 Inicijalna obrada dokumenata

U ovom koraku se iz teksta uklanjaju neinformativne riječi, koje nemaju veliki doprinos procesu kategorizacije. Nakon tokenizacije teksta, uklanjaju se: interpunkcijski znakovi, specijalni karakteri, numerički karakteri i riječi iz predefinisane liste stop-riječi (eng. *stopwords list*) (bazirane na listama stop-riječi iz Python biblioteka NLTK[1] v3.5 i spaCy[2] v2.3.1). Zatim se, procesom lematizacije, riječi u tekstu mijenjaju njihovim lemama.

Na kraju, sve riječi u tekstu se konvertuju u oblike sastavljene od malih slova.

3.5.2 Opis kategorija

Opisivanje kategorija je proces koji se sastoji od nekoliko koraka. U svakom koraku dodaju se nove riječi koje dodatno opisuju kategoriju, na sljedeći način.

1 https://www.nltk.org/
2 https://spacy.io/

Najprije se, u prvom koraku, za svaku kategoriju iz predefinisanog skupa kategorija definiše inicijani set od nekoliko riječi koje je najbolje opisuju. Ovaj korak može da bude izvršen manuelno ili automatizovano. Automatizovani pristup podrazumijeva da se iz neke *web* dostupne baze dokumenata, kao što je Wikipedia, izdvajaju riječi koje se najčeše pojavljuju u kontekstu neke kategorije.

Naredna tri koraka se oslanjaju na eksploataciju WordNet baze.

3.5.2.1 WordNet

WordNet je leksička baza engleskog jezika, u kojoj su imenice, glagoli, pridjevi i prilozi grupisani u skupove sinonima - riječi sličnih po značenju i međusobno zamjenjivih u mnogim kontekstima. Ovi skupovi riječi nazivaju se *synset*-i i međusobno mogu da budu u nekoliko vrsta relacija [24,25].

U drugom koraku se za svaku riječ iz inicijanog seta riječi, upotrebom WordNet baze, dodaju sve riječi koje predstavljaju neki od izvedenih oblika date riječi (npr. za imenicu *technology* to su oblici *technologist* i *technological*), uz ograničenje da se svi dodati oblici nalaze u vokabularu korpusa dokumenata.

U trećem koraku se za svaku riječ iz inicijalnog seta riječi dodaju svi njeni sinonimi iz WordNet baze, opet uz ograničenje da se svi sinonimi nalaze u vokabularu korpusa dokumenata.

Zatim se, u četvrtom koraku, za sve riječi definisane u prethodnim koracima, u vokabularu korpusa dokumenata pronalaze sve riječi koje imaju isti korijen (eng. *stemm*) ili lemu (eng. *lemma*), upotrebom Porter stemera i WordNet lematajzera iz NLTK v3.5 Python biblioteke.

Peti korak podrazumijeva upotrebu Word2Vec modela, s ciljem da se za riječi iz inicijalnog skupa riječi pronađu semantički slične riječi iz *embedding* prostora.

3.5.2.2 Word2Vec

Word2Vec je jedan od najčešće korištenih *word embedding* algoritama, zasnovan na upotrebi dvoslojne neuronske mreže, na način da se maksimizuje vjerovatnoća da su riječi prediktovane iz konteksta i obrnuto. Postoje dvije arhitekture Word2Vec algoritma: *Continuous Bag-Of-Words* (CBOW) i *Skip-Gram* (SG). CBOW arhitektura prediktuje riječ na osnovu njenog konteksta definisanog kontekstnim prozorom, dok SG arhitektura za zadatu riječ prediktuje riječi u njenom kontekstnom prozoru [15].

Pristup korišten u ovom radu podrazumijeva da se za svaku riječ iz inicijalnog skupa riječi dodaje deset najbližih riječi iz generisanog *embedding* prostora.

Setovi riječi generisani kroz prethodne korake, za svaku kategoriju iz skupa predefinisanih kategorija, predstavljaju njihove opise i dalje se koriste kao reprezentativni dokumenti za svaku od kategorija i porede metodama za određivanje sličnosti tekstualnih dokumenata sa dokumentima iz testnog skupa dokumenata.

3.5.3 Određivanje sličnosti

Nakon što su dokumenti obrađeni i opisi kategorija definisani na prethodno opisane načine, u posljednjem koraku primijenjenog pristupa nenadgledanog učenja, određuje se sličnost svakog od dokumenata iz testnog skupa sa opisima svake od kategorija iz predefinisanog skupa kategorija. Na taj način najsličnija kategorija postaje prediktovana kategorija dokumenta iz testnog skupa.

Testirane su dvije metode za određivanje sličnosti tekstualnih dokumenata. Prva metoda podrazumijeva generisanje i upotrebu dokument vektora latentne semantičke analize (eng. *Latent Semantic Analysis* - *LSA*), a druga testirana metoda za mjerenje sličnosti tekstualnih dokumenata zasniva se na reprezentaciji dokumenta kao srednjeg Word2Vec vektora svih njegovih tokena.

3.5.3.1 LSA

LSA je tehnika za kreiranje vektorskih reprezentacija tekstova i primarno se koristi za poređenje i određivanje sličnosti tekstualnih dokumenata. Ova tehnika se zasniva na upotrebi vektora koji nose informacije o frekventnosti pojavljivanja riječi u tekstu i tehnike dekompozicije singularne vrijednosti (eng. *Singular Value Decomposition* - *SVD*). Tehnikom dekompozicije singularne vrijednosti se generiše vektorski prostor u kojem su dimenzije sortirane od najviše do najmanje značajnih. Ignorisanjem manje značajnih dimenzija redukuje se dimenzionalnost vektorskih reprezentacija teksta [26].

U oba slučaja pomoću kosinusne sličnosti (eng. *cosine similarity*) računa se distanca između vektora dokumenata, odnosno njihova sličnost.

Opisani pristup je primjenjiv i na skupovima podataka na drugim prirodnim jezicima, uz uslov postojanja sličnih leksikografskih baza WordNet-u i odgovarajućih algoritama za morfološku normalizaciju.

4. EKSPERIMENTALNI USLOVI

4.1 Skupovi podataka (eng. *datasets*)

Opisani pristupi nadgledanog i nenadgledanog mašinskog učenja testirani su i evaluirani na pet standardnih skupova podataka za kategorizaciju teksta, čije su osnovne karakteristike prikazane u tabeli 1. Vokabular i prosječan broj tokena u dokumentu se odnose na skupove podataka nakon faze inicijalne obrade dokumenata.

Skup podataka	Trening skup podataka	Testni skup podataka	Kategorije	Vokabular	Prosječan broj tokena u dokumentu
5AbstractGroup	4,996	1,250	5	23,250	115
20NewsGroup	11,314	7,532	20	131,606	138
Google-Snippets	10,060	2,280	8	27,145	17
AG's Corpus	120,000	7,600	4	64,797	23
News Category Dataset	44,919	11,230	8	47,636	16

Tabela 1. *Karakteristike skupova podataka za evaluaciju klasifikacionih metoda*

5AbstractGroup[3] je kolekcija akademskih radova iz pet različitih oblasti: biznis, vještačka inteligencija, sociologija, transport i pravo, prikupljena sa *Web of Science*[4] domena. Iz svakog rada su izdvojeni sažetak i naslov i korišteni kao doku-

3 https://github.com/qianliu0708/5AbstractsGroup

4 https://login.webofknowledge.com/

ment. Skup podataka sadrži 6,246 dokumenata, od čega 4,996 dokumenata pripada skupu za treniranje, a 1,250 dokumenata skupu za testiranje klasifikacionih pristupa.

20NewsGroup[5] je kolekcija 18,846 dokumenata, ravnomjerno raspoređenih u 20 različitih kategorija. Neke od kategorija su domenski veoma bliske (npr. comp.sys.ibm.pc.hardware i comp.sys.mac.hardware). 11,314 dokumenata pripada skupu za treniranje, a 7,532 dokumenta skupu za testiranje klasifikacionih pristupa.

Google-Snippets[6] je skup dokumenata koji sadrže rezultate *web* pretraživanja vezanih za 8 različitih domena, od kojih su neki biznis, kompjuteri i inženjerstvo. Skup podataka sadrži 12,340 dokumenata, od čega 10,060 dokumenata pripada skupu za treniranje, a 2,280 dokumenata skupu za testiranje klasifikacionih pristupa.

AG's Corpus of news articles[7] je kolekcija više od milion novinskih članaka, prikupljenih iz više od 2,000 izvora. U radu je korištena verzija skupa podataka koju su kreirali Zhang, Zhao i LeCun 2015. godine, koja sadrži dokumente iz četiri najveće kategorije, sastavljene od polja naslova i opisa iz originalnih dokumenata [6]. Svaka kategorija sadrži po 30,000 primjeraka za treniranje i 1,900 primjeraka za testiranje, pa je ukupan broj primjeraka u skupu podataka za treniranje 120,000, a u testnom skupu 7,600, odnosno ukupan broj dokumenata u korpusu 127,600.

News Category Dataset[8] je kolekcija članaka objavljenih na *HuffPos*[9] domenu, u periodu od 2012. do 2018. godine. Originalni skup podataka se sastoji od preko 200,000 članaka, raspoređenih u 41 kategoriju, sastavljenih od labele kategorije, naslova, autora, linka, kratkog opisa i datuma publikovanja [27]. U ovom radu korišten je podskup ovog skupa podataka, koji sadrži 56,149 dokumenata iz 8 kategorija, sastavljenih od labele kategorije, naslova i kratkog opisa, od čega 44,919 dokumenata pripada skupu za treniranje, a 11,230 dokumenata skupu za testiranje klasifikacionih pristupa.

4.2 *Selekcija atributa i najboljih vrijednosti parametara*

U svim primijenjenim pristupima za kategorizaciju teksta, testirane su višestruke vrijednosti pojedinih parametara, kao i različite metode za mjerenje sličnosti tekstualnih dokumenata, i evaluirani rezultati u pogledu predefinisanih metrika. Metrike koje su korištene za evaluaciju rezultata kategorizacije su preciznost, odziv i F1 metrika.

4.2.1 Nadgledana kategorizacija

U ovom radu su korišteni sljedeći klasifikacioni algoritmi nadgledanog mašinskog učenja: naivni Bajesov klasifikator, metoda potpornih vektora, stabla odlučivanja i šuma slučajnih stabala. Nakon faze inicijalne obrade dokumenata, dokumenti su mapirani u vektorski ulazni prostor za obučavanje algoritama. Kao vektorske reprezentacije dokumenata korišteni su srednji Word2Vec vektori svih tokena u dokumentu. Za generisanje srednjih vektora korišteni su Word2Vec modeli koji su pokazali najbolje rezultate u nenadgledanoj kategorizaciji i opisani su u poglavlju 4.2.2. U slučaju upotrebe naivnog Bajesovog klasifikatora, vrijednosti atributa, odnosno vrijednosti u srednjim Word2Vec vektorima su skalirane u opseg od 0 do 1, po svakoj dimenziji.

Hiperoptimizacijom parametara, pristupom sistematske pretrage parametara po mreži vrijednosti (eng. *grid search*) [28], izabrani su skupovi optimalnih parametara za svaki od primijenjenih algoritama za klasifikaciju, za svaki skup podataka. Dodatno, sa tehnikom sistematske pretrage parametara po mreži vrijednosti primijenjena je i tehnika unakrsne validacije (eng. *cross-validation*) [29], a kao metrika za evaluaciju svake iteracije sistematske pretrage parametara po mreži vrijednosti korištena je F1 metrika. Skupovi vrijednosti parametara koji su evaluirani hiperoptimizacijom su u nastavku opisani za svaki algoritam. Za testiranje svih pomenutih algoritama korištene su Python implementacije iz biblioteke scikit-learn[10] v0.23.1.

Za naivni Bajesov klasifikator, evaluirane su različite vrijednosti *additive smoothing* parametra *alpha*, u opsegu vrijednosti od 0 do 1. *Alpha* je parametar koji određuje vrstu *smoothing*-a atributa, odnosno za vrijednosti manje od 1 Lidstone *smoothing*, a vrijednost 1 Laplace *smoothing* atributa.

Za metodu potpornih vektora, evaluirane su različite vrijednosti *penalty* parametra C, kojim se reguliše količina šuma u podacima, u opsegu vrijednosti od 0 do 5.

Za stabla odlučivanja podešavani su sljedeći parametri: *max_features*, koji određuje maksimalan broj atributa koji se uzimaju u obzir pri podjeli čvora, sa vrijednostima jednakim broju atributa (150 u slučaju korištenih srednjih Word2Vec vektora za reprezentaciju dokumenata), polovini broja atributa i kvadratnom korijenu broja atributa, *min_samples_split* parametar, koji definiše minimalan broj uzoraka potrebnih za podjelu unutrašnjeg čvora, sa vrijednostima 2, 5 i 10, *min_samples_leaf* parametar, koji definiše minimalan broj uzoraka neophodnih na listovima stabla, sa vrijednostima 1, 5 i 10, te *max_depth* parametar, koji određuje maksimalnu dubinu stabla, sa vrijednostima 3, 50 i vrijednošću sa kojom svi listovi stabla sadrže manje od *min_samples_split* uzoraka.

Za šume slučajnih stabala podešavani su isti parametri, sa istim vrijednostima kao i za stabla odlučivanja, uz dodatni parametar koji definiše broj stabala u šumi stabala, sa vrijednostima 10, 30 i 100.

4.2.2 Kategorizacija pristupom nenadgledanog učenja

U primijenjenom pristupu nenadgledanog mašinskog učenja testirani su različiti brojevi riječi po kategoriji u inicijalnom koraku procesa generisanja opisa kategorija, različite Word2Vec arhitekture, kao i različite metode za mjerenje sličnosti tekstualnih dokumenata.

5 http://qwone.com/~jason/20Newsgroups/

6 http://jwebpro.sourceforge.net/data-web-snippets.tar.gz

7 http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html

8 https://www.researchgate.net/publication/332141218_News_Category_Dataset

9 https://www.huffpost.com/

10 https://scikit-learn.org/stable/

Brojevi riječi testirani u inicijalnom koraku procesa generisanja opisa kategorija su deset, petnaest i dvadeset riječi po kategoriji. U tabeli 2 su prikazani prosječni brojevi generisanih riječi po kategoriji, nakon čitavog procesa opisivanja kategorija, u zavisnosti od broja riječi u inicijalnom koraku, za svaki skup podataka.

Inicijalni broj riječi po kategoriji	5AbstractGroup	20NewsGroup	Google-Snippets	AG's Corpus	News Category Dataset
10	190-280	140-490	140-240	160-390	150-400
15	240-430	200-650	220-350	250-530	260-550
20	360-540	250-910	260-430	400-650	270-760

Tabela 2. *Brojevi generisanih riječi po kategoriji u zavisnosti od inicijalnog broja riječi*

Word2Vec modeli su obučavani na cijelim korpusima dokumenata, nakon faze inicijalne obrade dokumenata. Testirane su obe Word2Vec arhitekture, CBOW i SG. Za obe arhitekture generisana su po četiri modela, za svaki skup podataka, sa sljedećim kombinacijama parametara: dimenzionalnost izlaznog vektora 150 sa dužinama kontekstnog prozora 5 i 10 i dimenzionalnost izlaznog vektora 300 sa dužinama kontekstnog prozora 5 i 10.

Na osnovu izgenerisanih modela generisane su i najbliže riječi riječima iz inicijalnog seta riječi u procesu opisivanja kategorija, kao i srednji vektori svih tokena u dokumentu, koji su korišteni za računanje sličnosti dokumenata.

U pristupu za mjerenje sličnosti dokumenata pomoću dokument vektora latentne semantičke analize testirani su vektorski prostori dimenzionalnosti 150 i 300, generisani na osnovu ulaznog TF-IDF (*Term Frequency-Inverse Document Frequency*) vektorskog prostora.

Rezultati svih primijenjenih pristupa, sa najboljim parametrima, u pogledu predefinisanih metrika, za sve testne skupove podataka, dati su u petom poglavlju.

5. REZULTATI I DISKUSIJA

U tabeli 3 su prikazani rezultati najboljih iteracija svake od testiranih metoda za kategorizaciju teksta, odnosno preciznost, odziv i F1 metrike, izražene u procentima, za svaki skup podataka.

Metoda	5AbstractGroup			20NewsGroup			Google-Snippets		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Naïve Bayes	81.5145	80.8000	80.8730	79.8719	78.9697	79.1390	83.7082	82.1053	82.1326
SVM	86.5371	86.5600	86.5397	84.8243	84.8513	84.5796	81.1128	79.3860	79.4593
Decision Tree	70.3091	70.0000	70.0557	47.1257	46.6144	46.8022	55.7935	53.5526	53.0327
Random Forest	84.0864	83.7600	83.8164	80.1470	80.0584	79.6878	82.1067	78.2456	77.8166
Unsupervised	79.0375	76.8800	77.1629	73.7340	58.1386	59.4787	85.2047	84.0351	84.2052

Metoda	AG's Corpus			News Category Dataset		
	Prec.	Rec.	F1	Prec.	Rec.	F1
Naïve Bayes	84.9592	85.0658	84.9749	76.0160	75.3339	75.5569
SVM	89.0170	89.0526	89.0281	80.8980	80.9617	80.8849
Decision Tree	82.5368	82.4737	82.4992	51.1608	51.1843	51.1265
Random Forest	89.3394	89.3553	89.3376	76.4655	75.7257	75.3681
Unsupervised	83.4975	83.6316	83.4866	75.1052	71.7275	72.1134

Tabela 3. *Rezultati testiranih pristupa za kategorizaciju teksta - preciznost, odziv i F1 metrike (%)*

Iz prikazanih rezultata se vidi da najbolje rezultate u pogledu svih metrika, na skupovima podataka 5AbstractGroup, 20NewsGroup i News Category Dataset, daje metoda potpornih vektora. U slučaju Google-Snippets skupa podataka najbolje rezultate daje primijenjeni pristup nenadgledanog učenja, dok na AG's Corpus skupu podataka šuma slučajnih stabala pokazuje najbolje performanse. Najlošiji rezultati su dobijeni primjenom stabala odlučivanja, na svim skupovima podataka, te se na osnovu toga može zaključiti da se stabla odlučivanja ne preporučuju za rješavanje problema klasifikacije teksta.

Najveće razlike između najboljeg algoritma nadgledanog učenja i primijenjenog pristupa nenadgledanog učenja, a u korist algoritma nadgledanog učenja, su primjetne na 20NewsGroup skupu podataka, dominantno u pogledu odziva (26.7127%), a posljedično i F1 metrike (25.1009%). Na ostalim skupovima podataka razlike se kreću u opsegu od 5% do 10%, u korist algoritma nadgledanog učenja, osim u slučaju Google-Snippets skupa podataka, gdje pristup nenadgledanog učenja daje bolje rezultate za 4-5% u pogledu svih korištenih metrika.

6. ZAKLJUČAK

Kategorizacija teksta je oblast mašinskog razumijevanja prirodnih jezika koja, u opštem slučaju, može da ima primjenu u skoro svakom automatskom upravljanju tekstualnim sadržajima, odnosno pronalaženju različitih obrazaca (eng. *patterns*) i izdvajanju relevantnog sadržaja iz istih. U radu je opisan proces kategorizacije tekstualnih dokumenata, od inicijalne obrade teksta u dokumentima, zatim ekstrakcije/selekcije atributa i obučavanja izabranog algoritma za klasifikaciju, te testiranja i evaluacije rezultata klasifikacije. Dat je pregled najčešće korištenih algoritama nadgledanog učenja za kategorizaciju teksta, kao i opis jednog pristupa nenadgledanog učenja, zasnovanog na upotrebi Word2Vec algoritma i WordNet leksičke baze engleskog jezika. Opisani pristupi su testirani na pet poznatih skupova podataka za kategorizaciju tekstualnih dokumenata. U radu je opisan i proces selekcije vrijednosti pojedinih parametara u svim pomenutim pristupima.

Algoritmi nadgledanog učenja pokazuju na četiri skupa podataka bolje rezultate u pogledu predefinisanih metrika, dok pristup nenadgledanog učenja daje bolje rezultate na jednom skupu podataka. Međutim, za razliku od algoritama nadgledanog učenja, upotreba pristupa nenadgledanog učenja ne zahtjeva labelisanje skupa podataka za obučavanje, čime se drastično smanjuju potrebni resursi za realizaciju čitavog procesa kategorizacije tekstualnih dokumenata. Labelisanje korpusa dokumenata je vremenski i finansijski zahtjevan proces, često i neprihvatljiv, posebno u slučajevima jako velikog broja dokumenata u skupu za obučavanje, kada pristupi nenadgledanog učenja postaju jedino rješenje.

Dalja istraživanja u primijenjenom pristupu nenadgledanog učenja bi se, prije svega, odnosila na unapređenje procesa opisivanja kategorija. Samo neke od mogućnosti u tom pravcu su, na primjer, upotreba Word2Vec modela obučavanih na velikim skupovima podataka, kao što je Wikipedia, te upotreba drugih *word embedding* algoritama, kao što su Glove [18], FastText [12,17], BERT [30], ELMO [31] i sl. Takođe, i početni korak definisanja inicijalnih riječi u kategorijama bi mogao da se oslanja na upotrebu različitih dostupnih baza dokumenata. Dodatno, jedan od narednih pravaca istraživanja mogle bi biti i različite metode za mjerenje sličnosti tekstualnih dokumenata.

LITERATURA

[1] Erik Cambria, Bebo White, „Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article] “, IEEE Computational Intelligence Magazine, vol. 9, no. 2, pp. 48-57, doi: 10.1109/MCI.2014.2307227, 2014.

[2] Janusz Kacprzyk, Slawomir Zadrozny, „Computing With Words Is an Implementable Paradigm: Fuzzy Queries, Linguistic Data Summaries, and Natural-Language Generation“, IEEE Transactions on Fuzzy Systems, vol. 18, no. 3, pp. 461-472, doi: 10.1109/TFUZZ.2010.2040480, June 2010.

[3] Lien-Fu Lai, Chao-Chin Wu, Pei-Ying Lin and Liang-Tsung Huang, „Developing a fuzzy search engine based on fuzzy ontology and semantic search“, IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011), Taipei, Taiwan, pp. 2684-2689, doi: 10.1109/FUZZY.2011.6007378, 2011.

[4] Abe Kazemzadeh, Sungbok Lee and Shrikanth Narayanan, „Fuzzy Logic Models for the Meaning of Emotion Words“, IEEE Computational Intelligence Magazine, vol. 8, no. 2, pp. 34-49, doi: 10.1109/MCI.2013.2247824, May 2013.

[5] Minh-Thang Luong, Richard Socher, Christopher D. Manning, „Better Word Representations with Recursive Neural Networks for Morphology“, CoNLL-2013, 104, 2013.

[6] Xiang Zhang, Junbo Zhao, Yann LeCun, „Character-level Convolutional Networks for Text Classification”, Advances in neural information processing systems, pp. 649–657, 2015.

[7] Siwei Lai, Liheng Xu, Kang Liu, Jun Zhao, „Recurrent Convolutional Neural Networks for Text Classification”, Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, pp. 2267-2273, 2015.

[8] Shad Akhtar, Deepanway Ghosal, Asif Ekbal, Pushpak Bhattacharyya and Sadao Kurohashi, „All-in One: Emotion, Sentiment and Intensity Prediction using a Multi-task Ensemble Framework“, IEEE Transactions on Affective Computing, doi: 10.1109/TAFFC.2019.2926724, 2019.

[9] Cícero Nogueira dos Santos, Maíra Gatti, „Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts“, Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pp. 69–78, Dublin, Ireland, August 23-29 2014.

[10] Ahmed Aliwy, Esraa Hussein, „Comparative Study of Five Text Classification Algorithms with their Improvements“, International Journal of Applied Engineering Research, Vol. 12 (14), pp. 4309-4319, 2017.

[11] Emmanouil Ikonomakis, Sotiris Kotsiantis, V. Tampakas, „Text Classification Using Machine Learning Techniques”. WSEAS transactions on computers. Vol. 4. pp. 966-974, 2005.

[12] Armand Joulin, Edouard Grave, Piotr Bojanowski, Tomas Mikolov, „Bag of Tricks for Efficient Text Classification”, Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Vol. 2, pp. 427–431, 2017.

[13] Shweta Dharmadhikari, Maya Ingle, Parag Kulkarni, „Empirical Studies On Machine Learning Based Text Classification Algorithms“, Advanced Computing: An International Journal, Vol. 2 (6), pp. 161-169, 2011.

[14] Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, Donald Brown, „Text Classification Algorithms: A Survey”, Information 2019, 10(4), 150, 2019.

[15] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, „Efficient estimation of word representations in vector space”, Proceedings of the International Conference on Learning Representations, 2013.

[16] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean, „Distributed Representations of Words and Phrases and their Compositionality”, Advances in Neural Information Processing Systems 26, 2013.

[17] Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov, „Enriching Word Vectors with Subword Information“, Transactions of the Association for Computational Linguistics, Vol. 5, pp. 135–146, 2017.

[18] Jeffrey Pennington, Richard Socher, Christopher D. Manning, „Glove: Global Vectors for Word Representation”, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Vol. 14, pp. 1532–1543, 2014.

[19] Anuradha Patra, Divakar Singh, „A Survey Report on Text Classification with Different Term Weighting Methods and Comparison between Classification Algorithms“, International Journal of Computer Applications, Vol. 75 (7), 2013.

[20] Vandana Korde, „Text Classification and Classifiers: A Survey”, International Journal of Artificial Intelligence & Applications, Vol. 3, pp. 85-99, 2012.

[21] Vikas K. Vijayan, K. R. Bindu and Latha Parameswaran, „A comprehensive study of text classification algorithms“, International Conference on Advances in Computing, Communications and Informatics (ICACCI), Udupi, pp. 1109-1113, doi: 10.1109/ICACCI.2017.8125990, 2017.

[22] Ranko Petrović, Miloš Pavlović, Branka Stojanović, Snežana Puzović, „Prepoznavanje emocija sa slika lica primenom multi-senzorskih sistema”, Infom, br. 70, pp. 38-45, 2020.

[23] Zied Haj-Yahia, Adrien Sieg, Léa A. Deleris, „Towards Unsupervised Text Classification Leveraging Experts and Word Embeddings“, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 371–379 Florence, Italy, July 28 - August 2, 2019.

[24] Christiane Fellbaum, „Wordnet: An on-line lexical database”, 1998.

[25] Christiane Fellbaum, „WordNet and wordnets”, In: Brown, Keith et al. (eds.), Encyclopedia of Language and Linguistics, Second Edition, Oxford: Elsevier, pp. 665-670, 2005.

[26] P. Hastings, „Latent Semantic Analysis”, 2004.

[27] Rishabh Misra, “News Category Dataset”, 2018.

[28] Iwan Syarif, A. Prugel-Bennett, Gary Wills, „SVM Parameter Optimization using Grid Search and Genetic Algorithm to Improve Classification Performance”. TELKOMNIKA (Telecommunication Computing Electronics and Control), Vol.14 (4), pp. 1502-1509, 2016.

[29] Sebastian Raschka, „Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning“, Technical Report, 2018.

[30] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, „Bert: Pre-training of deep bidirectional transformers for language understanding”, Proceedings of NAACL-HLT 2019, pp. 4171–4186, 2018.

[31] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer, „Deep contextualized word representations”, Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1, pp. 2227–2237, 2018.

Ana Bojanić, dipl. ing., Bravo Systems d.o.o. Banja Luka, RS, BiH / Elektrotehnički fakultet, Univerzitet u Banjoj Luci, RS, BiH
Kontakt: ana.bojanic@bravosystems.com
Oblast interesovanja: data science, data engineering, mašinsko učenje, objektno-orijentisano programiranje i modelovanje, Internet programiranje

Zoran Đurić, Elektrotehnički fakultet, Univerzitet u Banjoj Luci, RS, BiH
Kontakt: zoran.djuric@etf.unibl.org
Oblast interesovanja: sigurnost, kriptografija, PKI, platni sistemi i protokoli, formalna verifikacija, mašinsko učenje, data science, objektno-orijentisano programiranje i modelovanje, Internet programiranje, razvoj mobilnih aplikacija, XML-bazirana međuoperativnost, web servisi, računarske mreže, penetration testing, sistem integracija