Predikcija popularnosti oglasa za posao na Linkedin društvenoj mreži

Dejan Kurdulija
Elektronsko poslovanje
Računarstvo i automatika
Fakultet tehničkih nauka
Univerzitet u Novom Sadu

Nemanja Malinović
Elektronsko poslovanje
Računarstvo i automatika
Fakultet tehničkih nauka
Univerzitet u Novom Sadu

Zdravko Milinković
Elektronsko poslovanje
Računarstvo i automatika
Fakultet tehničkih nauka
Univerzitet u Novom Sadu

***Abstract*—Sa porastom popularnosti društvenih mreža i internet stranica za promovisanje i reklamiranja otvorenih poslovnih pozicija, predviđanje popularnosti oglasa za posao postalo je važno za kompanije koje nude poslovne prilike. U ovom radu istražujemo različite faktore koji utiču na popularnost oglasa za posao na *Linkedin* društvenoj mreži, kao što su broj prijavljenih na oglas, broj pregleda oglasa. Analiziramo efikasnost različitih tehnika mašinskog učenja, kao što su klasifikacioni modeli i neuronske mreže. Naši rezultati ukazuju da modeli mašinskog učenja mogu predvideti popularnost oglasa za posao na *Linkedin* društvenoj mreži sa visokom tačnošću.**

Keywords—Oglas za posao na Linkedin mreži, mašinsko učenje, neuronske mreže

I. UVOD

Linkedin društvena mreža je osnovana u decembru 2002. godine, a pokrenuta je u maju 2003. godine. Od svog nastanka je postala ozbiljna društvena mreža koju prvenstveno koriste poslovni ljudi kako bi postavljali informacije o svojim dostignućima, poslovima, kompanijama, kao i oglase za posao. Osim toga, *Linkedin* nudi i praktične funkcije kao što su pretraga prethodno navedenih stavki, personalizovani preporučeni sadržaj i mogućnost pretplate na omiljene stranice i profile. Korisnicima *Linkedin* profila koji su u potrazi za poslom je na ovaj način mnogo olakšana potraga za poslom.

U ovom kontektsu, oglašavanje poslova na ovoj mreži nije samo deo rutinskog zapošljavanja, već ključni element strategije privlačenja talenta. Oglasi za posao na ovoj platformi imaju potencijal da dopru do široke publike visoko kvalifikovanih profesionalaca iz različitih industrija i sektora, obezbeđujući poslodavcima priliku da angažuju najbolje talente. Međutim, kako bi oglasi bili efikasni, potrebno je razumeti dinamiku Linkedin platforme, kao i faktore koji utiču na njihovu popularnost među ciljnom publikom. To obuhvata širok spektar elemenata, uključujući sadržaj oglasa, ciljanje publike, vreme objavljivanja, kao i algoritamske faktore koji utiču na vidljivost oglasa u *Linkedin feed-u* korisnika.

Popularan i trending oglas za posao nisu ista stvar, iako se ovi termini često koriste kao slični. Popularan oglas za posao je onaj oglas za posao koji ima velik broj pregleda, komentara, velik broj apliciranih i drugih interakcija od strane korisnika. S druge strane, trending oglas za posao je oglas koji se nedavno pojavio i naglo počeo da dobija velik broj pregleda i broj apliciranih u kratkom vremenskom periodu. Popularan oglas može biti već dugo dostupan *na Linkedin-u* i imati veliku publiku koja ga posećuje, dok trending oglas predstavlja novu pojavu koju su ljudi tek otkrili.

U ovom radu predstavljamo studiju o predviđanju popularnosti *Linkedin* oglasa za posao korišćenjem tehnika mašinskog učenja. Analiziramo različite karakteristike oglasa za posao i koristimo ih za obuku modela koji može predvideti popularnost, a zatim evaluiramo performance našeg modela. Prva faza realizacije rešenja je nalaženje ili formiranje skupa podataka. Korišćeni skup podataka sadrži 33000 evidentirana posla. Svaki pojedinačni oglas sadrži 27 vrednih atributa uključujući i naslov, broj apliciranih, broj pregleda, opis posla, platu, lokaciju, tip rada i slično. Sa toliko podataka, potencijal za istraživanje ovog skupa je ogroman. Kroz detaljnu analizu podataka o aktivnostima korisnika, kao i karakteristika oglasa, planiramo da identifikujemo obrasce i trendove koji ukazuju na visoko angažovane oglase.Pre obučavanja modela vršen je inženjering svojstava i eksplorativna analiza po različitim tipovima podataka koje skup podataka sadrži. Za predikciju korišćeni su klasifikacioni modeli koji daju ocenu popularnosti smeštanjem oglasa za posao u jednu od dve klase broja pregleda i broja apliciranih dobijenih diskretizacijom. Pomoću *XGboost* algoritma je urađena analiza numeričkih podataka i predikcija popularnosti na osnovu njih.

Kroz ovaj rad, ne samo da želimo da pružimo uvid u kompleksnost procesa oglašavanja na *Linkedin-u* , već i da ponudimo konkretne smernice koje će poslodavcima pomoći da optimizuju svoje kampanje za oglašavanje poslova i postignu bolje rezultate u privlačenju talenata. Naš cilj je da doprinesemo efikasnijem procesu zapošljavanja i boljem povezivanju poslodavaca i potencijalnih zaposlenih na ovoj globalnoj profesionalnoj mreži.

Navedeni koraci implementirani su korišćenjem programskog jezika Python.

U drugom poglavlju dat je kratak opis tuđih istraživanja koja se bave sličnim problemima i metodama koje su koristili. Treće poglavlje sadrži detaljniji opis skupa podataka. U četvrtom poglavlju je već spomenuta eksplorativna analiza. Peto poglavlje sadrži detaljan opis korišćenih metodologija. Opis treniranja i testiranja odabranih modela nalazi se u poglavljima VI i VII.

II. ISTRAŽIVANJE

A. Predikcija popularnosti

Predviđanje popularnosti društvenih medija ima za cilj da automatski predvidi buduću popularnost (npr. klikovi, pregledi, lajkovi, itd.) onlajn objava putem mnoštva podataka društvenih mreža sa javnih platformi. To je ključni problem za učenje i predviđanje društvenih medija i jedan od najizazovnijih problema u ovoj oblasti. Sa stalno promenljivim interesima korisnika i pažnjom javnosti na platformama društvenih medija, kako tačno predvideti popularnost postaje izazovnije nego ranije. Ovaj zadatak je dragocen za dobavljače sadržaja, trgovce ili potrošače u nizu aplikacija u stvarnom svetu, uključujući multimedijalno oglašavanje, sistem preporuka ili analizu trendova.

U radu [1] cilj je predikcija predikcija popularnosti *Instagram* objave. Autori istražuju ovu temu kao ključan izazov u društvenim medijima, sa fokusom na postizanje tačnog klasifikovanja popularnosti objava. Rad koristi Gradient boosting kao osnovni model za predviđanje popularnosti. Problemu je pristupljeno kao zadatku klasifikacije, koristeći relevantne karakteristike objava za treniranje modela. Rad koristi skup podataka sa Instagram mreže za treniranje i evaluaciju modela. Skup podataka sadrži informacije o objavama kao što su slika, tekst, vreme objavljivanja, broj lajkova, broj komentara i sl. Evaluacija se vrši kroz određene metrike koje omogućavaju procenu tačnosti predloženog modela. Metrike, kao što su tačnost, preciznost, odziv ili F1 mera, koriste se za ocenu performansi modela. Ovo uključuje tačnost predikcija, analizu relevantnih metrika i upoređivanje sa drugim pristupima ili modelima.

Takođe su korišćeni algoritmi poput SGD (*Stohastic Gradient Descent*), neuronska mreža (MLPC), *Decision Trees i Random Trees (*Stabla za odlučivanje*)*. Zbog nebalansiranih klasa dodeljene su im težine i primenjena je tehnika višenivoovskog binarnog okvira. Ovim tehnikama osigurana je raspoređena klasifikacija gde predikcija više nije naklonjena određenim klasama. Vizualizacija rezultata je urađena pomoću heat mape, a evaluacija je izvršena pomoću F1 mere, gde su upoređeni svi korišćeni modeli.

Najbolje rezultate daje XGBoost na čije bi štelovanje i rezultate dali najviše pažnje. Rezultati rada daju nezadovoljavajuće rezultate za određene klase što je izazvano nebalansiranim podacima. Ovo možda može biti popravljeno sampling tehnikom. Relevantnost ovog rada za naš projekat ogleda se u tome što pruža uvide u sličan problem i pristup rešavanju istog.

Rad [2] se fokusira na proučavanje faktora koji doprinose popularnosti onlajn sadržaja. Analizirane su dve platforme: *YouTube* i *Digg.* Autori su imali za cilj da razviju prediktivni model koji bi mogao da proceni popularnost sadržaja pre nego što postane popularan. Ispitivali su različite karakteristike, kao što su broj pregleda, komentara, ocean i vremenski zavisni faktori, da bi identifikovali indikatore uspešnosti sadržaja. Pored toga, istraživali su ulogu spoljnih faktora, kao što su uticaj društvenih mreža korisnika i vreme postavljanja sadržaja. Sakupljanje podataka sa *YouTube* platforme su započeli 21 Aprila 2008. godine. Izabrali su 7146 video snimaka iz liste "most recently uploaded" kako ne bi imali bajas na neke kategorije video snimaka. I onda su, svakodnevno narednih 30 dana, ažurirali podatke o broju pregleda tih video snimaka. Podaci o video snimcima su podeljeni u dve grupe, u trening skup i u testni skup podatakam raspodela je 50% 50% odnosno u svaki skup po 3573 video snimka.

Korišćena su tri modela za predikciju popularnosti video sadržaja. Prvi je *linear regression on a logarithmic scale; least-squares absolute error*. A drugi model je *constant scaling model; relative squared error.* Treći korišćen model je *growth profile model.* Iako su se sva tri modela pokazala odlična u predikciji popularnosti video snimaka na *YouTube*. Najbolje predikcije daje drugi model odnosno *constant scaling model*. Autori su predložili prediktivni model koji može da proceni buduću popularnost datog sadržaja na osnovu njegovih ranih indikatora popularnosti. Otkrili su da rana merenja popularnosti zaista ukazuju na budući uspeh, omogućavajući im da predvide buduću popularnost sadržaja sa razumnom tačnošću.

Cilj istraživanja u radu [3] je unapređenje razumevanja poslova na Linkedin platformi kroz primenu dubokog prenosa znanja. Integracija ovih modela u Linkedin omogućava kontinuirano poboljšanje razumevanja poslova. Korišćeni su neobrađeni oglasi za posao sa širokim spektrom informacija poput broja prijavljenih, broja apliciranih ka oi detalji o poslovima koji se nalaze u ponudi poput lokacije, informacije o platama, način rada (rad od kuće, hibridno, iz kancelarije) i sl. Evaluacija se sprovodi kroz *Feedback loop* sa korisnicima, uključujući postupak prikupljanja povratnih informacija od postavljača oglasa. Na kraju je prikazano kako Linkedin pruža sugestje za evaluaciona pitanja i prikuplja povratne informacije od osoba koje postavljaju oglase za posao. Model standardizacije evaluacionih pitanja koristi opis posla koji je pružio korisnik koji postavlja oglas kao ulaz i izlazni podatak je lista važnih evaluacionih pitanja izdvojenih iz oglasa za posao. Prvo se primenjje fino podešen model dubokih prosečnih mreža na rečenicama u oglasu za posao kako bi se generisala lista kandidata za evaluaciona pitanja. Agregirana pitanja iz rečenica u oglasu za posao se uzimaju i izrđuje se bogat skup karakteristika. Takoše se karakteristike ne uređuju samo iz same tekstualne sadržine, već i iz statistika zabeleženih na celokupnom tržištu poslova, kao što su skorovi međusobnih informacija između industrija poslova i tipova pitanja. Na kraju, model na osnovu stable koje poboljsava *(gradient boosted tree based model)* rangira sva pitanja i vraća ih korisncima, gde imaju opcije da prihvate ili odbace. Petlja povratnih informacija se formira kada se iterativno ažurira model koristeći ažurirane podatke o povratnim informacija i ponovo se postavlja u produkciju.

Rezultati istraživanja u radu [3] su poboljšavanje zadovoljstva korisnika, povećanje broja prijava za posao i efikasnost u preporučivanju poslova i oglasa za posao. Rad je relevantan za naš projekat predikcije popularnosti oglasa za posao jer istražuje efikasne metode razumevanja neobrađenih oglasa.

III. OPIS SKUPA PODATAKA

Skup podataka koji se koristi je *Linkedin Job Postings – 2023* sa Kaggle-a koji se ažurira redovno i može se preuzeti sa [4]. Za ovaj projekat je poslednji put preuzet 25.01.2024. kako bi imali konzistentan skup podataka.

A. Kolone odabranog skupa podataka

- *Job_id* (Identifikator posla koji je definisala Linkedin društvena mreža)
- *Company_id* (Identifikator kompanije koja je vezana za oglas za posao)
- *Title* (Naziv posla)
- *Description* (Opis posla)
- *Max_salary* (Maksimalna vrednost plate za odgovarajući posao)
- *Med_salary* (Srednja vrednost plate za odgovarajući posao)
- *Min_salary* (Mininalna vrednost plate za odgovarajući posao)
- *Pay_period* (Plaćeno vreme koje je izraženo u satima, mesecima, ili na godišnjem nivou)
- *Formatted_work_type* (Tip posla koji može biti *fulltime, parttime ili* contract)
- *Location* (Lokacija posla)
- *Applies* (Broj aplikacija koje su podnete)
- *Original_listed_time* (Prvobitno vreme kada je posao objavljen)
- *Remote_allowed* (Da li posao dozvoljava rad od kuće)
- *Views* (Broj koji predstavlja Koliko puta je jedan oglas bio pregledan)
- *Job_posting_url* (URL link do oglasa za posao na platformi Linkedin)
- *Application_url* (URL link na kom aplikacije mogu biti podnete)
- *Application_type* (Tip procesa apliciranja, *offsite, complex/simple onsite*)
- *Expiry* (Datum isteka ili vreme isteka oglasa za posao)
- *Closed_time* (Vreme zatvaranja oglasa za posao)
- *Formatted_experience_level* (Nivo poslovnog iskustva, *entry, associate, executive,* itd.)
- *Skills_desc* (Opis koji specificira potrebne veštine za odgovarajuću poslovnu poziciju)
- *Listed_time* (Vreme kada je oglas okačen)
- *Posting_domain* (Domen vebsajta sa prijavom)
- *Sponsored* (Da li je posao sponzorisan ili promovisan)
- *Work_type* (Tip posla vezan za oglas)
- *Currency* (Valuta u kojoj je izražena plata)
- *Compensation_type* (Tip naknade za posao)
- *Scraped*

Ciljne kolone su *views i applies.* Dodatne kolone koje su uzete u razmatranje su *location, job_id, description.*

IV. EKSPLORATIVNA ANALIZA

U ovom radu vršimo eksplorativnu analizu podataka kako bismo istražili osnovne karakteristike i uzorke u našem skupu podataka. Predstavlja ključan korak u procesu istraživanja podataka, jer omogućava sticanje uvida u strukturu i raspodelu I međusobne veze podataka pre nego što se primene složenije analize. U nastaku, prikazujemo rezultate naše eksplorativne analize kao i implikacije koje potiču iz tih nalaza.

TABELA 1: NAJZASTUPLJENIJI POSLOVI

Posao	Broj apliciranih
Sales manager	5892
Web developer	6765
Fullstack developer	9765
Personal banker	2890

Nakon analize ustanovljeno je da se zastupljenost odredjenih poslova menjala tokom godina, pa tako imamo različita stanja:

- *Godina 2021:*
  1. *Sales manager*
  2. *Fullstack developer*
  3. *Personal banker*

- *Godina 2022:*
    1. *Fullstack developer*
    2. *Sales manager*
    3. *Web developer*
- *Godina 2023:*
    1. *Fullstack developer*
    2. *Web developer*
    3. *Sales manager*

A. *Kolone sa numeričkim vrednostima*

TABELA 2: KORELACIJA

	view_count
applies_count	0.8579
companies_count	0.5983

Iz tabele se vidi da je korelacija između broja apliciranih i broja pregleda dosta veća od korelacije broja kompanija sa brojem pregleda. Međutim, nijedna od dobijenih vrednosti nije zanemarljiva i ima svoj doprinos u procesu treniranja modela, pa nijedna numerička kolona nije izbačena.

V. METODOLOGIJA

Metodologija rešenja opisanog u ovom radu realizovana je kroz sledeće korake: prikupljanje podataka, eksplorativna analiza podataka, inženjering svojstava i pretprocesiranje podataka, obučavanje i optimizacija neuronskom mrežom. U nastavku su detaljno opisane sve prethodno navedene faze metodologije.

A. *Prikupljanje podataka*

Ovaj korak obuhvata pronalaženje osnovnog skupa podataka i njegovu početnu analizu.

B. *Inženjering svojstava*

Inženjeringom svojstava koristimo informacije iz postojećeg skupa kako bi došli do polja *companies_count* koje nam je značilo kao pomoćna promenljiva u daljim analizama. Spajanjem identifikatora sa imenima se dobila preglednija eksplorativna analiza.

C. *Pretprocesiranje podataka*

Pretproceriranje podataka transformiše podatke u format koji se lakše i efikasnije obrađuje u rudarenju podataka, mašinskom učenju i drugim zadacima nauke o podacima. Tehnike se generalno koriste u najranijim fazama mašinskog učenja I razvoja veštačke inteligencije kako bi se osigurali tačni rezultati.

Pretprocesiranje u ovom radu se sastoji iz više koraka koji nam omogućuju lakši rad sa podacima i koji dozvoljavaju da model ima prikladne podatke kao ulaze.

Koraci su sledeći:

- pretvaranje identifikatora posla u njegovo ime
- pretvaranje podataka kada je objavljen oglas u podatak
- koliko je poslu trebalo da uđe u popularniji posao
- zamena outlier-a sa granicama koristeći *upper/lower bound* tehniku
- zamena nedostajućih vrednosti sa medijan vrednošću
- kodiranje lokacija pomoću *One hot encoding* tehnike
- prepoznavanje popularnijih naslova i kodiranje pomoću *One hot encoding* tehnike
- normalizacija numeričkih podataka

Isprobana je integracija lokacija i naslova kao numerički i tekstualni podatak. U obe situacije lokacije su doveli do pada performansa svih modela, tako da smo odlučili tu kolonu da odbacimo.

D. *Obučavanje i optimizacija klasifikacionog modela*

Razlog za postojanje klasifikacije oglasa za posao po broju pregleda je potreba za relativnom merom popularnosti u odnosu na druge sadržaje koji su popularniji. Broj pregleda dobijen regresijom je apsolutna mera, koja kanalima često ne znači puno, jer ne znaju da li je taj rezultat bolji ili lošiji od ostalih u smislu popularnosti na dnevnom nivou. Zbog ovoga, u ovom istraživanju realizovana je i klasifikacija oglasa za posao po broju pregleda, u cilju davanja relativne ocene popularnosti oglasa za posao.

Da bi klasifikacija oglasa za posao bila moguća, neophodno je kreiranje kategoričkog atributa koji predstavlja diskretne vrednosti broja pregleda. Ovo je postignuto diskretizacijom kontinualne vrednosti broja pregleda u dve grupe balansirane po veličini. Ideja diskretizacije u dve kategorije je davanje ocene popularnosti oglasa za posao, tako da postoje nepopularni i popularni oglasi za posao. Diskretizacija se sprovodi dinamički, tako da se uvek dobijaju balansirane kategorije popularnosti oglasa za posao.

Po uzoru na postojeću literaturu korišćeni su sledeći klasifikacioni modeli:

- Support Vector klasifikator,
- Random Forest ansambl klasifikacionih stabala,
- Klasifikaciona stabla u Extreme Gradient Boosting konfiguraciji

Svi navedeni klasifikacioni modeli obučavani su and istim skupom podataka, koristeći isto pretprocesiranje. Ovde su korišćeni samo numerički atributi što znači da su naslov i opis bili izbačeni iz skupa podataka. Na ovaj način dobijeni su podaci u obučavajućem skupu.

Zastupljenost kategorija popularnosti u obučavajućem skupu je dobro izbalansirana: po 12289 u obe kategorije. Optimizacija parametara realizovana je upotrebom unakrsne validacije. Za evaluaciju performansi modela korišćene su metrike tačnost, F-mera (*F1 score*), AUC (*Area under the ROC Curve*) i MAE (*Mean absolute error*).

SVM je korišćen kao početni model kako bi videli da li ovaj problem predikcije može da se reši prostijim modelom i da li je SVM kriva dovoljna kako bi razdvojili popularne od nepopularnih oglasa za posao. Parametri koji su pomoću unakrsne validacije optimizovani su kernel, C i gamma. Našli smo da je rbf kernel najpogodniji za dati skup podataka. SVM daje najgore rezultate od izabranih klasifikacionih modela, ali I ne toliko gore od ostalih kako ne bi poslužio kao validan model za ovaj problem.

Sledeći korišćen model je *Random forest* ansambl klasifikacionih stabala. Ovaj model je pogodan pogotovo za ovaj skup podataka gde postoji visoka dimenzionalnost atributa, gde imamo kategorička svojstva i gde nam njegova nasumičnost omogućava robusnost na šum koji se često javlja pri većem broju atributa. Parametri koji su optimizovani su broj stabla i maksimalna dubina stable. Njihove vrednosti su 415 I 17 respektivno. *Random Forest* iznenađujuće daje najbolje rezultate od 3 izabrana klasifikatora.

Ispitan je I model ansambla klasifikacionih stabala u *Extreme Gradient Boosting konfiguraciji.* Ovaj model kao i *Random Forest* podržava visoku dimenzionalnost što se tiče atributa, a uglavnom je korišćen zbog njegovih superiornih performansi u odnosu na druge klasifikatore. Kod ovog modela smo najviše pažnje obratili na štelovanje parametara, a optimizovani parametri su: broj stabala, stopa učenja, poduzorak, maksimalna dubina stabla, procenat kolona za izgradnju stabla i minimalna težina potomka. Vrednosti dobijene unakrsnom validacijom su sledeće: 314, 0.036, 0.715, 13, 0.93 i 2. Model je pokazao zadovoljavajuće rezultate, ali i dalje zaostaje za *Random Forest* klasifikatorom koji se u ovoj situaciji pokazao kao bolje rešenje.

Poslednji model je višeslojna mreža perceptrona koji u suštini predstavljaju granu kombinovane neuronske mreže i služi kako bi se video uticaj tekstualnih atributa na performanse modela.

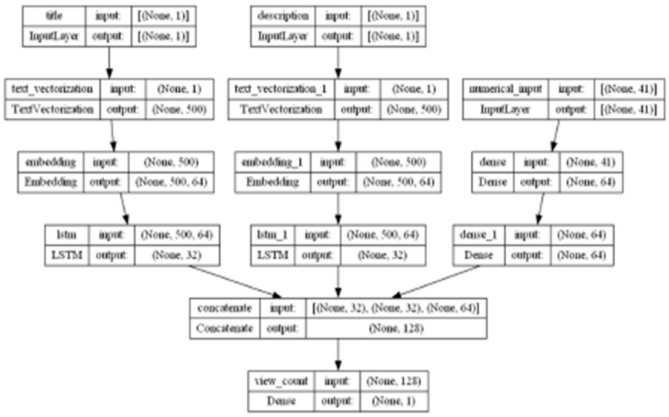
E. Arhitektura neuronske mreže

Kako bi se predvidela popularnost oglasa za posao, može se primeniti neuronska mreža, koja je sposobna da nauči složene obrasce u podacima i da ih koristi za predviđanje ciljne promenljive – popularnosti oglasa za posao.

Za implementaciju neuronske mreže, koristimo *Keras,* otvoreni programski okvir za razvoj dubokih neuronskih mreža napisan u *Python-u.*

Pored numeričkih podataka, u ovom modelu imamo mogućnost da integrišemo i tekstualne podatke kao što su opis i deskripcija oglasa za posao. Ovo će nam pokazati koliki uticaj mogu imati ovi podaci i koliko *clickbait* kultura sa naslovom i pogađanje ključnih reči opisom može da utiče na popularnost oglasa za posao.

Mreža se sastoji od tri grane koja prima određene ulaze, koji prolaze kroz nekoliko slojeva, koji su specifični po grani, i na kraju se konkateniraju kako bi dobili izlaz, tj. predikciju modela. Ulaz u prvoj grani je naslov, u drugoj opis, a u trecoj svi ostali podaci koji su predstavljeni numerikom.



SLIKA 1. ARHITEKTURA KOMBINOVANE NEURONSKE MREŽE

U prvoj grani kao ulaz imamo naslov oglasa za posao koji se prosleđuje na sloj vektorizacije. Ovaj sloj pretvara tekst u vector koji predstavlja lakši format za obradu. Vektor se dalje prosleđuje na *Embedding sloj* , koji u suštini predstavlja mapiranje na vektore realnih vrednosti i pomaže u smanjivanju dimenzionalnosti vektora. Ovim mapiranjem dobijamo smislenije vektore koji sadrže više značenja nego kada bi ulazi bili tekstualni. Model se dalje sastoji od LSTM (*Long short-term memorz networks*) rekurentne neuronske mreže koja ima sposobnost da uhvati dugoročne zavisnosti u sekvencijalnim podacima, poput teksta. U NLP-u (*Natural Language Processing*) to znači da LSTM može efikasno modelirati kontekst i odnose između reči u rečenici ili dokumenti. Identičan tok ima i druga grana gde se nalazi informacija o opisu oglasa za posao. U poslednjoj grani se nalaze numerički podaci, gde su naredni slojevi višeslojni perceptroni.

Pomoću ove grane unosimo uticaj statističkih , kao i kodiranih kategoričkih podataka, gde opštost višeslojnih perceptrona može najbolje predstaviti tu zavisnost. Ove grane se na kraju konkateniraju kako bi nam dale klasifikaciju popularnosti oglasa za posao. Na izlazu koristimo sigmoidnu funkciju aktivacije zbog klasifikacionog problema koji je u skupu [0,1], a za funkcije gubitka se koriste binarna i kategorijska unakrsna entropija, gde je veća težina postavljena na binarnu sa obzirom na prirodu problema.

Ovaj model za klasifikaciju daje zadovoljavajuće rezultate, a najviše je bio fokus da se vidi uticaj naslova i opisa na popularnost oglasa za posaosnia. Važno je napomenuti da se ovaj model trenira znatno duže od ostalih zbog obrade prirodnog jezika koji zahteva višeslojne rekurentne mreže u našem slučaju. Zbog dugog vremena treniranja parametri su optimizovani ručno kako bi dali najbolje performanse, radije nego unakrsnom validacijom.

VI. EVALUACIJA

U ovom poglavlju dat je pregled performansi modela korišćenih za klasifikaciju, kao i prikaz performansi pojedinačnih elemenata kombinovane neuronske mreže.

Klasifikacioni modeli testirani su na test skupu podataka, izdvojenom u početnoj fazi projekta i njega čini 20% od celog skupa podataka. Validacioni skup nije bio potreban pošto biblioteka sa kojom se radila unakrsna validacija ima ugrađeno izdvajanje i evaluiranje validacionih podataka. Izvršen je i pregled prošlih eksperimenata, kako bi se videla eksperimentacija sa podacima i kako su oni uticali na rezultate.

Klasifikacioni modeli evaluirani su upotrebom metrika tačnost, AUC, MAE i F-mera. Metrike su izabrane u skladu sa problemom, imajući u obzir da je cilj binarna klasifikacija, kao i da je broj pozitivnih i negativnih klasa dobro izbalansiran u datom skupu podataka. Korišćena je i matrica konfuzije, da bi se uočilo na koim klasama model više greši. Pored matrice konfuzije izvršena je i dodatna analiza grešaka kako bi mogli da uočimo zašto naši modeli greše i kako možemo da poboljšamo njihove performanse. Urađena je analiza i lokacija utiče na klasifikacioni model. Tabela 3 daje uvid i pregled vrednosti metrika za sve korišćene klasifikacione modele u fazi kada nisu integrisane lokacije kao atribut modela.

TABELA 3: EVALUACIJA BEZ LOKACIJE

Klasifikacioni modeli	Metrike [%]			
	Tačnost	AUC	MAE	F1
SVM	86.22	86.18	13.78	85.39
Random Forest	87.33	87.30	12.51	87.01
XGBoost	87.12	87.09	40.84	86.62
MLP	85.43	85.40	14.51	84.97

Primećujemo da najgore rezultate daje prosta mreža višeslojnih perceptrona. Ovo je očekivano s obzirom da je ona samo deo kombinovane mreže. Najbolje se pokazao *Random Forest* model.

U tabeli 4 vidimo rezultate sa integrisanim lokacijama u klasifikacione modele. Lokacije su integrisane tako što su analizirani sve jedinstvene lokacije i nađeno top 20 lokacija. Ove lokacije su nakon toga kodirane pomoću *One Hot Encoding* tehnike.

TABELA IV: EVALUACIJA SA LOKACIJAMA

Klasifikacioni modeli	Metrike [%]			
	Tačnost	AUC	MAE	F1
SVM	86.05	86.02	13.95	85.25
Random Forest	87.33	87.30	12.67	86.80
XGBoost	86.80	86.78	40.94	86.33
MLP	85.70	85.69	14.30	85.43

Kao što vidimo svi rezultati osim MLP su lošiji sa integracijom lokacija u model. Ovo je zbog velike raznovrsnosti I velikom broja jedinstvenih lokacija, gde se najpopularnije lokacije ne ponavljaju u tolikoj frekvenciji da bi poboljšalje performanse modela. Lokacije onda postaju šum i izazivaju pad u rezultatima. Zbog toga smo odlučili da odbacimo lokacije kao numeričke podatke.

Do sad smo za sve eksperimente koristili uobičajene parameter koji dolaze sa modelom. Na kraju and najboljim modelom vršimo optimizaciju parametara, gde su rezultati tih modela prikazani u tabeli 5.

TABELA 5: EVALUACIJA BEZ LOKACIJA I SA OPTIMIZOVANIM PARAMETRIMA

Klasifikacioni modeli	Metrike [%]			
	Tačnost	AUC	MAE	F1
SVM	86.56	86.53	13.43	85.88
Random Forest	87.60	87.57	12.40	87.13
XGBoost	87.52	87.47	40.75	87.01
MLP	85.77	85.73	14.23	84.95

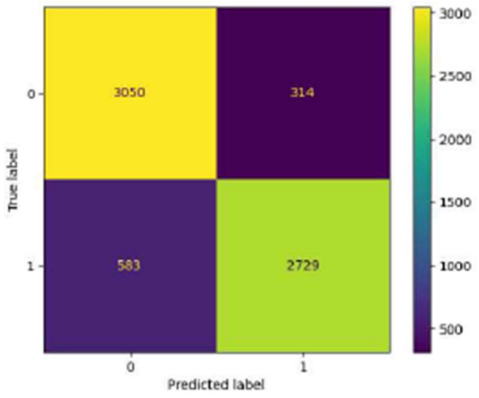
Na kraju imamo tabelu 6 evaluacije kombinovane neuronske mreže. Evaluacija je urađena samo na finalnom modelu zbog dužine treniranja modela. Vidimo kako se u odnosu na višeslojnu mrežu perceptrona povećavaju mere evaluacije. Možemo da zaključimo da tekstualni atributi imaju uticaj na predikciju popularnosti i da se sa većom skalabilnošću mreže mogu još poboljšati rezultati.

TABELA 6: EVALUACIJA BEZ LOKACIJA I SA OPTIMIZOVANIM PARAMETRIMA

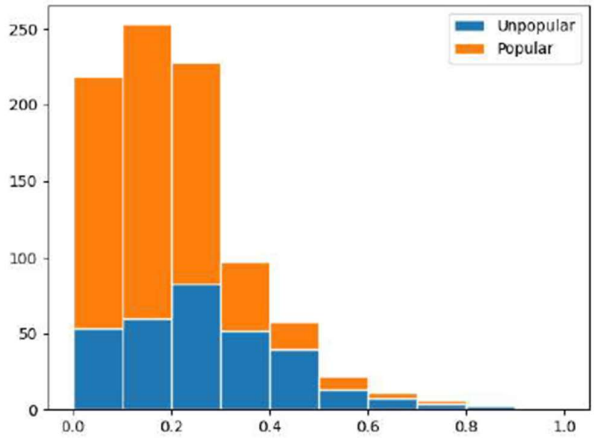
Klasifikacioni modeli	Metrike [%]			
	Tačnost	AUC	MAE	F1
Kombinovana neuronska mreža	85.85	85.83	14.30	85.43
MLP	85.77	85.73	14.23	84.95

Matrica konfuzije SVM je prikazana na slici 2. Matrice modela su u sličnim odnosima kao i ova. Vidimo da se najviše grešaka javlja kod oglasa za posao koji su popularni a model ih prepoznaje kao nepopularne.

Ovo se može objasniti činjenicom da je teže klasifikovati populrne oglase za posao koji imaju manje cifre kad se tiče najvažnijih atributa za klasifikaciju. Ovo se može primetiti na slici 3. gde imamo histogram broja apliciranih kod pogrešnih predikcija, i gde vidimo da model greši kod malih broja apliciranih i to najviše kod popularnih oglasa za posao.



SLIKA 2. MATRICA KONFUZIJE SVM MODELA



SLIKA 3. HISTOGRAM APLIKACIJA KOD POGREŠNIH PREDIKCIJA U ODNOSU NA POPULARNE I NEPOPULARNE OGLASE ZA POSAO

VII. ZAKLJUČAK

U ovom naučnom radu bavili smo se predikcijom popularnosti oglasa za posao na *Linkedin* društvenoj mreži. Metodologija se sastojala od nekoliko koraka, uključujući prikupljanje podataka, eksplorativnu analizu podataka, inženjering svojstva , obučavanje i optimizaciju klasifikacionih modela, analizu teksta rekurentnom neuronskom mrežom i formiranje kompleksne neuronske mreže koja kombinuje sve navedene tipove ulaza za klasifikaciju oglasa za posao.

Prikupljanje podataka je obuhvatilo pronalaženje odgovarajućeg skupa podataka i inicijalnu analizu istog kako bi se ostvarila ideja našeg istraživanja.

Inženjering svojstva i pretprocesiranje podataka su bili važan korak u obradi podataka. Podaci su pretvoreni u format koji je lakši za obradu, uključujući pretvaranje identifikatora u imena, zamenu outlier-a, zamenu nedostajućih vrednosti i normalizaciju numeričkih podataka.

Obučavanje i optimizacija klasifikacionog modela su izvršeni korišćenjem različitih modela, uključujući *Support Vector* klasifikator, *Random Forest* ansambl klasifikacionih stabala, ansambl klasifikacionih stabala u *Extreme Gradient Boosting* konfiguraciji i višeslojnu mrežu perceptrona. Svi modeli su obučeni nad istim skupom podataka I korišćen je isti podskup atributa. Performanse modela su evaluirane kroz različite metrike, uključujući tačnost, F-meru, AUC i MAE.

Eksperimenti su pokazali da *Random Forest* ansambl klasifikacionih stabala daje najbolje rezultate u klasifikaciji popularnosti oglasa za posao. Ovaj model je posebno pogodan za visokodimenzionalne podatke I pokazao se robustnim u prisustvu šuma u podacima. *Support Vector* klasifikator je pružio slabije rezultate u poređenju sa ostalim modelima, dok je ansambl klasifikacionih stabala u *Extreme Gradient Boosting* konfiguraciji dao zadovoljavajuće rezultate, ali nešto slabije od *Random Forest-a.*

Višeslojna mreža perceptrona je istražena kako bi se procenio uticaj tekstualnih atributa na performanse modela pri poređenju sa kombinovanom neuronskom mrežom.

Kombinovana neuronska mreža je takođe primenjena za analizu popularnosti oglasa za posao. Arhitektura mreže sadržala je tri grane koje su primale različite ulaze, uključujući naslov, opis i numeričke podatke pri čemu je LSTM rekurentna neuronska mreća korišćena za obradu naslova i opisa oglasa za posao. Ovaj model je takođe pokazao zadovoljavajuće rezultate, ali je vreme treniranja bilo duže u odnosu na ostale modele. Ova arhitektura je omogućila modelu da nauči složene obrasce u podacima i da sa visokom tačnošću klasifikuje oglase za posao.

Kao moguća proširenja i unapređenja rešenja bilo bi dodavanje analize plate na određenim poslovnim pozicijama jer i taj podatak ima veliki uticaj na broj pregleda i na popularnost oglasa za posao. To bi moglo da se postigne uz pomoć neuronske mreže.

References

[1] "Popularity prediciton of Instagram Posts"
Avaliable: https://www.mdpi.com/2078-2489/11/9/453

[2] "Predicting the Popularitz of Online Content" by Gabor Szabo and Bernardo A. Huberman.
Avaliable: https://www.researchgate.net/publication/23417017_Predicting_the_Popularity_of_Online_Content

[3] Shan Li, Baoxu Shi, Jaewong Yang "Deep Job Understanding at Linkedin"
Avaliable: https://arxiv.org/pdf/2006.12425.pdf

[4] Linkedin dataset
Avaliable: https://www.kaggle.com/datasets/arshkon/linkedin-job-postings/