Analiza IT firmi na osnovu podataka sa Joberty.rs platforme

Mitar Perović
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad
perovicmitar@uns.ac.rs

Alen Mujo
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad
mujo.r210.2021@uns.ac.rs

Milena Laketić
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad
laketic.milena@uns.ac.rs

Apstrakt

Zadovoljstvo zaposlenih je veoma važno za bilo kakvu organizaciju kako bi uspešno napredovala i ostvarivala svoje ciljeve. Kako bi se kreirala prijatna atmosfera koja podstiče produktivnost, neretko se politike firmi prilagođavaju zahtevima zaposlenih. Sa tim u vidu, od velikog značaja za kompanije je da imaju uvid u stavove i mišljenja kako svog kolektiva, tako i onih iz drugih firmi. Sa druge strane, onima koji traže posao ovakve informacije pomažu da otkriju prednosti i slabosti kompanija koje ih interesuju. U ovom radu kreirali smo skup podataka o srpskim IT firmama koristeći se recenzijama zaposlenih, bivših zaposlenih i kandidata kao i ostalim informacijama sa sajta Jobery.rs. Izvršene analize nad ovim skupom podeljene su u šest zadataka. Prvi od njih predstavlja klasterovanje kompanija na osnovu ocena aspekata poslovanja. Rezultati ove analize pokazali su da se kompanije iz ovog skupa mogu klasterovati u tri grupe na osnovu prosečne ocene, iako ona nije korišćena pri treniraju modela, kao i da aspekt "ocena benefita kompanije" ima značajan deo u podeli u klastere. Izvršena je analiza sentimenta komentara koji se tiču procesa selekcije upotrebom BERT modela. Zaključeno je da HR i tehnički intervju imaju podjenak uticaj na formiranje utiska o selekciji. Korišćenjem modela RandomForest istreniran je model koji vrši predikciju procenta zaposlenih koji smatraju da je plata fer sa rezultatom od 0.1 u kontekstu metrike MSE. Korišćenjem XGBClassifier modela je odrađena predikcija težine intervjua gde je model ostvario rezultat od 0.62 micro F1 metrike. Određeni su najznačajniji aspekati kompanije uz posmatranje ovog zadatka kao regresionog problema sa prosečnom ocenom kompanije kao ciljnom labelom. Rezultati ukazuju na to da su aspekti kao što su fleksibilnost i odnos sa kolegama značajniji u odnosu na korišćene tehnologije i platu. Na kraju, izvršena je ekstrakcija novih znanja na osnovu statističkih obrada podataka.

Ključne reči—it firme srbija; eksplorativna analiza; joberty; klasterovanje; značaj obeležja; regresija

I. Uvod

Kao ljudi, primarno socijalna bića, konstanto se oslanjamo na mišljenje drugih pri formiranju svojih stavova i donošenju odluka. Evolutivno smo naučeni da vrednujemo mišljenje većine, jer bi u suprotnom rizikovali izopštenje iz zajednice. Samim tim, nije iznenađujuća popularnost sajtova za recenziranje.

II. Pregled postojeće relevantne literature

Istraživanjem radova koji su se bavili sličnom tematikom upoznali smo se sa širokom primenom analize sentimenta u online recenzijama. Jedan od glavnih motiva bio je sprovođenje sličnih analiza nad skupom recenzija koje su napisane na srpskom jeziku i tiču se IT kompanija u Srbiji. Iako korišćeni metodi u ispod pomenutim radovima nisu direktno preneseni u ovaj rad, oni su u velikoj meri inspirisali sve zadatke koji se bave analizom aspekata kompanije.

A. Aspect based Sentiment Analysis of Employee's Review Experience [1]

Primarni cilj ovog rada bio je da se odredi ocena apsekata kompanije upotrebom analize sentimenta zasnovane na aspektima. Skup podataka sastoji se od recenzija zaposlenih sa sajta *Glassdor.com*. Prvi korak predstavlja određivanje ključnih reči koji se bazirao na njihovoj učestalosti. Zatim su iz recenzija izbačene sve reči koje se ne smatraju ključnim, odnosno samo su imenice klasifikovane kao prave ključne reči upotrebom *Stanford POS Tagger* modela. Ključne reči su

klasifikovane prema pet aspekata: poslovni balans (eng. *work balance*), kultura, mogućnosti za napredovanje, benefiti kompanije i menadžment. Rezultati aspekt orijentisane sentiment analize demonstrirali su koja od pomenutih kategorija predstavlja prednost odnosno slabost za pojedinačne kompanije.

B. *Aspect-Sentiment Embeddings for Company Profiling and Employee Opinion Mining* [2]

Odabir jedne od mnogih kompanija i organizacija kao i njihovo rangiranje u mnoštvu ponuda može postati težak zadatak. Podstaknuti ovim problemom autori u ovom radu kreirali su skup podataka na osnovu recenzija sa sajta *Glassdor.com* i izvršili aspekt orijentisanu sentiment analizu. Glavna ideja bila je projektovanje kompanija u 30-dimenzioni prostor gde svaka dimenzija predstavlja prosečnu ocenu aspekta kompanije. Aspekti su posmatrani kao osobine organizacija kao što su plata, poslovni balans, lokacija. Ova analiza doprinosi pogotovo onima koji traže posao da uspešno rangiraju firme koje ih interesuju. Svaka kompanija je na osnovu njenih recenzija korišćenjem Doc2vec pretvorena u 30 dimenzioni vektor (eng. *word embedding*) koji se koristio za analizu sentimenta i klasifikaciju. *SentiWordNet* je korišćen za kreiranje rečnika koji pojmove dodeljuje mogućim aspektima. Za klasifikator je korišćen ELM i SVM model.

C. Sentiment classification for employees reviews using regression vectorstochastic gradient descent classifier (RVSGDC) [3]

U ovom radu izvršena je klasifikacija recenzija zaposlenih na osnovu njihovog sentimenta. Korišćeni su podaci o šest kompanija: Facebook, Microsoft, Amazon, Netflix, Apple i Google koji su preuzeti sa sajta *Kaggle.com*. Koristeći *TextBlob* izdvojili su sentiment recenzija zaposlenih i klasifikovali ga kao pozitivan ili negativan. Nakon toga su iskoristili klasifikator koji predstavlja kombinaciju logističke regresije, SVM-a i stohastičkog gradijentnog spusta (eng. *Regression Vector-Stochastic Gradient Descent Classifier*, skraćeno RV-SGDC) za klasifikaciju sentimenata. Konačno rešenje su dobili većinskim glasanjem pomenutih modela.

III. Skup podataka

Podatke neophodne za realizaciju projekta smo prikupili direktno sa sajta *Jobery.rs*, raznim tehnikama povlačenja podataka (eng. *scraping*). Skup podataka je razdvojen na šest manjih skupova, od kojih svaki predstavlja jednu od stranica vezanih za pojedinačnu firmu. Na samom sajtu smo uspeli pronaći 389 validnih linkova ka stranicama firmi koje posluju u Srbiji. Prva stranica koje je skrejpovana predstavlja stranicu “iskustva”. Na njoj se nalaze ocene određenih aspekata poslovanja, od strane zaposlenih. Ukupno je ocenjeno deset aspekata, i to su sledeći:

- uslovi rada,
- radna atmosfera,
- fleksibilnost,
- odnos poslodavca,
- tim,
- projekti,
- tehnologije,
- zarada,
- benefiti,
- lični razvoj.

Pored navedenih aspekata poslovanja, prikupljena je i prosečna ocena kompanije, kao i procenat zaposlenih koji preporučuje kompaniju. Od ukupno 389 kompanija, četiri su bile bez ocena, te su uklonjene iz skupa podataka.

Sa iste stranice, prikupljeni su i komentari trenutnih i bivših zaposlenih. Svaki komentar ima odvojene sekcije za pozitivne, negativne strane i za projekte. 1495 komentara nemaju sekciju o opisu projekta kompanije. Svaki komentar sadrži i ocenu koju je zaposleni ostavio, kao i datum i vreme komentara.

Podaci na stranici “Plata i benefiti” sadrže informacije o broju zaposlenih koje je označilo prisustvo određenih benefita. Moguće je selektovati deset benefita od kojih neki imaju veliki broj nedostajućih vrednosti (predstavljeno u zagradi), i to:

- fleksibilno radno vreme (40),
- rad od kuće (40),
- plaćene obuke i treninzi (87),
- rad preko agencije (195),
- privatno osiguranje (98),
- zabavni i fitnes sadržaji (105),
- obezbeđena hrana (192),
- obezbeđen parking (168),
- deljenje profita (260),
- akcije kompanije (248)

Pored navedenih benefita, skrejpovan je i procenata zaposlenih koji misli da je plata fer, kao i procenat zaposlenih koji dobija godišnji bonus.

Sledeći podskup podataka se odnosi na plate za određene pozicije u firmi. Za svaku od navedenih pozicija, skrejpovane su informacije o minimalnoj, maksimalnoj, kao i prosečnoj plati za tu konkretnu poziicju. Problem sa ovim podskupom je bio nedostatak velikog broja vrednosti za minimalnu i maksimalnu platu (934 nedostajuće vrednosti po svakom atributu), te ove vrednosti nisu korišćene u nastavku rada.

Poslednja sekcija svake kompanije tiče se intervjua i procesa selekcije. Informacije koje su dostupne su:

- prosečna ocena procesa selekcije (od 1 do 5)
- težina intervjua (lak, srednje, težak) - izraženo procentualno
- trajanje procesa selekcije - izraženo u nedeljama

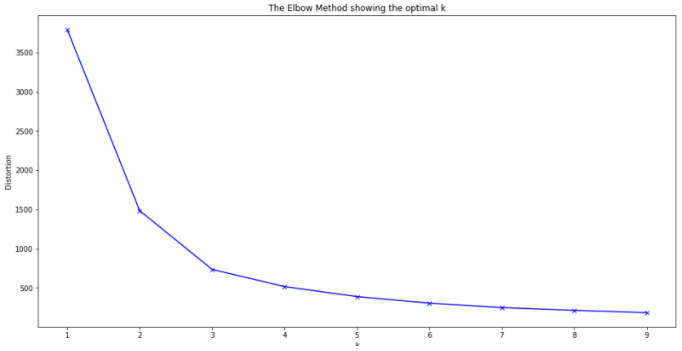
Na istoj stranici se nalaze i komentari od strane kandidata. Za poslednji podskup podataka su skrejpovani svi komentari sa ocenama od strane prethodnih kandidata. Pored ocene (od jedan do pet), nalaze se i dve tekstualne sekcije komentara. Prva se tiče samog HR (eng. *human resources*) intervjua, dok se sledeća odnosi na tehnički intervju. Na kraju svakog komentara se nalazi i status ponude, tačnije da li je korisnik prihvatio ponudu, odbio ponudu ili je nije dobio.

IV. Metod i rezultati

U ovom poglavlju bavićemo se konkretnom metodologijom primenjenom nad prethodno opisanim podacima. Sav rad je podeljen na šest zadataka, koji će biti opisani u nastavku poglavlja.

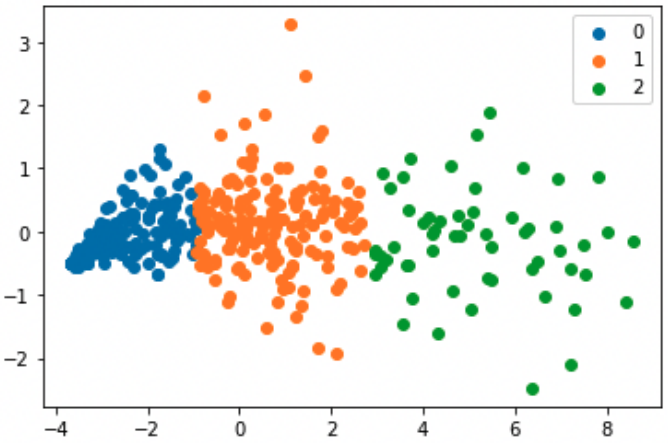
A. Nenadgledano klasterovanje nad ocenama aspekata kompanije

Prvi algoritam korišćen za klasterovanje jeste KMeans. Kako je u pitanju nenadgledano učenje, prvi problem je bio pronalazak optimalnog broja klastera. Najčešće korišćena metoda jeste tzv. "metoda lakta". Inicijalizovali smo KMeans algoritam sa različitim brojem klastera, varirajuću od jedan do deset. Za svaki model smo merili inerciju, odnosno prosečno kvadratno rastojanje od centroida klastera. Sa grafika ispod se može videti da je optimalan broj klastera tri, zato što daljim povećavanjem broj klastera, ne dobijamo značajno smanjenje inercije. Kako je ovo idealan scenario za metodu lakta, smatrali smo da ne treba isprobavati kompleksnije tehnike, poput funkcije siluete. Nastavili smo rad sa tri klastera.



Slika 3.1 Određivanje broja klastera metodom lakta

Obična implementacija u okviru *scikit-learn* biblioteke implementira KMeans++ algoritam, te nismo podešavali dodatno parametre. Odrađena je redukcija dimenzionalnosti na dve komponente, radi vizualizacije. Kao atributi su korišćene sve vrednosti aspekata kompanije, sem atributa prosečne ocene kompanije i procenta zaposlenih koji preporučuju datu kompaniju. Pomoću PCA algoritma je odrađena redukcija na dve dimenzije. Kumulativna objašnjena varijansa komponentama iznosi 87%. Nakon klasterovanja, prikazali smo podatke na grafiku 3.2, bojeći ih odgovarajućom labelom u zavisnosti od pripadanja klasteru.



Slika 3.2 Podaci grupisani u klastere primenom KMeans++ algoritma

Analiziranjem zelenog klastera i selekcijom podataka koje je model označio labelom dva, dobijamo 80 kompanija. Izlistavanjem prosečne ocene kompanije za date kompanije, dolazimo do uvida da su u proseku kompanije ocenjene ocenom 2.6. Minimalna vrednost je 1.2, dok je maksimalna 4.2. Očigledno se radi o klasteru najslabije ocenjenih kompanija. U ovaj klaster su upale četiri kompanije sa ocenom većom od 3.7. Iako za dobar deo aspekata

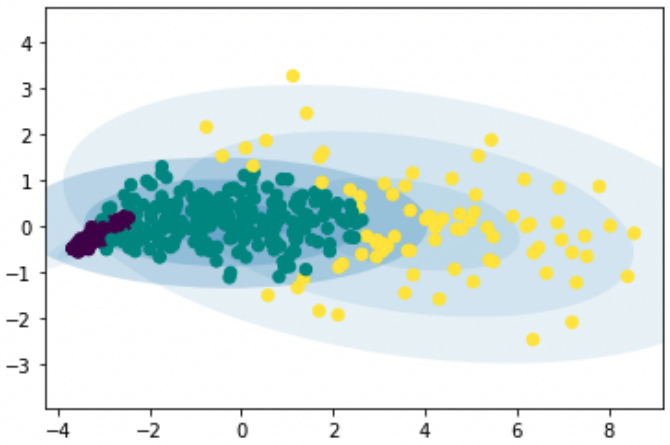
imaju ocene pet, za aspekt "ocena benefita kompanije" ocene su najniže. Klaster u sredini, narandžaste boje, dobijamo najveći broj kompanija, čak 228. Izlistavanjem prosečne ocene ovih kompanija, dolazimo do ocene od 3.97, čineći ih osrednje ocenjenim kompanijama. Ukupno 11 kompanija ima ocenu veću od 4.6, što bi ih moglo klasifikovati u najbolje ocenjene kompanije. Najbolje ocenjena kompanija od njih, kao najgore ocenjen aspekt ima "ocena benefita kompanija", što potencijalno ukazuje da je algoritam dao veliku važnost ovom atributu pri podeli u klastere. Najnižu srednju vrednost ocene, ovih 11 kompanija, imaju za aspekt "ocena projekata".

Poslednji klaster, obojen plavom bojom, čini 77 kompanija. Prosečna ocena kompanije ovog klastera iznosi 4.87, čineći ih najbolje ocenjenim kompanijama. Rezultati su dati u tabeli 3.1. Sa slike 3.2 i tabele 3.1 možemo zaključiti da je algoritam uspeo gotovo odlično da napravi distinkciju među prosečnom ocenom kompanije, uzevši u obzir samo ocene aspekata kompanije. Iz minimalne i maksimalne vrednosti ovog klastera se može videti da nema upadanja kompanija u pogrešan klaster, te ni samim tim, zanimljivih slučajeva za dalju analizu. Zaključak je da je algoritam iskonvergirao ka raspodeli na osnovu prosečne ocene, iako ona nije uključena u sam trening modela.

Rating/ Cluster	**Min**	**Max**	**Mean**	**Median**
Green	1.2	4.2	2.56	2.6
Orange	3.1	4.8	3.97	4
Blue	4.6	5	4.87	4.9

Tabela 3.1 Rezultati klasterovanja na osnovu prosečne ocene kompanije

Pored KMeans algoritma, odrađeno je klasterovanje pomoću Gausovih mešavina (eng. *Gaussian Mixture Models*). Korišćene su tri komponente, i rezultati su donekle slični sa KMeans algoritmom, uz par distinkcija. Klasterovani podaci prikazani su na grafiku 3.3.



Slika 3.3 Podaci grupisani u klastere primenom Gaussian Mixture Models

B. Predikcija procenta zaposlenih koji misle da je plata fer na osnovu benefita kompanije

Predikcija procenta zaposlenih koji smatraju da je plata fer je rađena nad skupom podataka o benefitima svake kompanije koji su opisani u poglavlju III. Prvi korak je bio pretvoriti broj korisnika koji potvrđuju postojanje određenog benefita za neku kompaniju u procenat deljenjem te vrednosti sa kolonom *review_count* koja predstavlja ukupan broj korisnika koji su ocenjivali postojanje benefita za tu kompaniju. Nakon toga je isprobano 7 različitih kombinacija kolona koje predstavljaju ulaz u model. Kolone koje imaju veliki broj nedostajućih vrednosti (više od dve trećine) su izbačene. Zadatak modela je da prediktuje procenat korisnika koji smatraju da je plata fer (kolona *considers fair salary*). Skup podataka je podeljen na trening i test skup u razmeri od 80:20. Trenirani su modeli *Support Vector Regression, XGBRegressor, Linear Regression* i *Random Forest Regressor*. Korišćena je unakrsna validacija nad 10 delova (eng. *folds*) dok su hiperparametri optimizovani pomoću mrežaste pretrage (eng. *grid search*). Modeli su evaluirani i poređeni korišćenjem MSE (eng. *mean squared error*) metrike. Rezultati treniranih modela su prikazani u tabeli 3.2..

Model	MSE
SupportVectorRegression	0.1245
XGBRegressor	0.1322
Linear Regression	0.1120

RandomForestRegressor	**0.1086**

Tabela 3.2 Rezultati treniranih modela

Najbolji rezultat od 0.1 je postigao *RandomForestRegressor* model sa parametrima *max_depth*=4 i *n_estimators*=128 i korišćenjem ulaza koji čine kolone *flex_hours, remote_work, paid_courses, insurance* što ukazuje na to da su ti benefiti najbolji pokazatelj da li zaposleni misle da je plata fer.

D. Analiza sentimenta komentara procesa selekcije

Za ovaj zadatak je iskorišćen podskup podataka koji se tiče komentara na sam proces selekcije u kompanijama. Sadrži ocenu koju je dao korisnik, kao i tekstulane komentare HR i tehničkog intervjua. Ideja zadatka je bilo ustanovljenje koji od dva intervjua ima veći uticaj na sveukupan utisak o procesu selekcije. Da bismo to odredili, odlučili smo se da računamo predikciju sentimenta samog komentara i poredimo je sa stvarnom ocenom koju je dao korisnik. Za ove potrebe korišćen je Transformer model iz HuggingFace bibliteke. Preciznije, korišćen je BERT model koji je treniran kao višejezični model za analizu sentimenta recenzija proizvoda od strane korisnika. Iako srpski jezik nije zvanično uključen u fine-tuning ovog modela, prvo smo ručno testirali kakve predikcije daje. U tabeli 3.3 se mogu videti rezultati za par primera.

Komentar	**Predikcija broja zvezdica**
Nista posebno iskreno. Okej	2
Promasio sam firmu, trebao sam zaobici	1
Jedna od boljih u Novom Sadu	5
Što se tiče samog prvog kruga tj. HR intervju-a, tu imam sve pohvale za devojku koja radi taj deo.\nOstali krugovi su neprijatni i veštački.	2

Tabela 3.3 Predikcija sentimenta komentara upotrebom BERT modela

Iz priloženog smo zaključili da možemo koristiti ovakav model za predikciju sentimenta komentara.

Glavna ideja je bila da računamo MSE (eng. mean squared error) za komentare HR dela, kao i za komentare tehničkog dela intervjua. Za istinitu labelu smo uzeli ocenu koju je dao korisnik, i pokušali smo da vidimo komentar kog dela intervjua je bliži njoj. Nakon prolaska kroz svih 1856 komentara, dobili smo rezultate, i to za MSE komentara HR dela iznosi 2.7384, dok MSE komentara tehničkog dela interjvua iznosi 2.7461. Iz priloženog vidimo da je sentiment manje više konzistentan za oba dela intervjua, uz neznatnu prednost HR dela kao većeg faktora na sveukupnu ocenu. Da bismo bili sigurni, uradili smo statističku analizu. Naime, čuvana je kvadrirana razlika između ocene korisnika i ocene modela, i to za oba dela intervjua. Na kraju smo imali dva dataseta kvadriranih grešaka i odlučili smo se Studentov T test. Nulta hipoteza (H0) glasi da su srednje vrednosti kvadriranih grešaka oba dela intervjua jednake, dok alternativna hipoteza glasi da nisu jednake. Nakon sprovedenog T Testa nad ova dva skupa podataka, dobili smo rezultat za p-vrednost 0.95. Kako je p-vrednost veća od 0.05, prihvatamo nultu hipotezu, te zaključujemo da ne postoji razlika u srednjoj vrednosti kvadriranih razlika sentimenta komentara u odnosu na ocenu korisnika. Zaključak je da oba dela ipak imaju podjednak uticaj na formiranje utiska o celokupnom procesu selekcije.

E. Predikcija labele težina intervjua

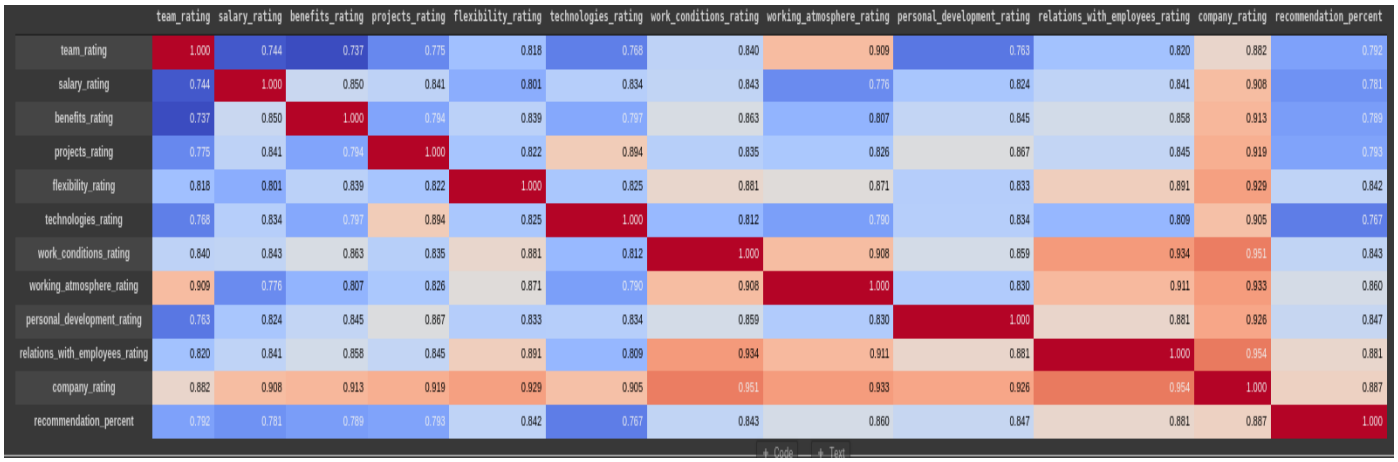
U skupu podataka o ocenama intervjua je za svaku kompaniju procenjena od strane korisnika težina intervjua (lak, srednje, težak) - izraženo procentualno. Cilj ovog zadatka je koristeći ocene procesa intervjua i statusa ponude prediktovati težinu intervjua. Kako postoje tri nivoa težine intervjua, za svaku kompaniju je težina intervjua ocenjena onom kategorijom koja ima najveću procentualnu vrednost i predstavljena *one hot encoding-om*. Status ponude je predstavljen *one hot* vektorom i ima tri kategorije: prihvatio, odbio i nije dobio ponudu. Skup podataka je podeljen na trening i test skup u razmeri od 80:20. Obzirom da je ovo *multi-class* problem klasifikacije sa tri klase, korišćen je *One-vs-Rest* pristup (problem se razdvaja na jedan binarni klasifikator za svaku od klasa u skupu podataka). Trenirani su modeli *Linear SVC, XGB Classifier* i *Random Forest Classifier.* Korišćena je unakrsna validacija nad 10 delova (eng. *folds*) dok su hiperparametri optimizovani pomoću mrežaste pretrage (eng. *grid search*). Modeli su evaluirani i poređeni korišćenjem *micro-F1* metrike. Rezultati su prikazani u tabeli 3.4.

Model	micro-F1
LinearSVC	0.6182
XGBClassifier	**0.6207**
RandomForestClassifier	0.6012

Tabela 3.4 Rezultati treniranih modela

Najbolji rezultat je ostvario *XGBClassifier* od 0.62 *micro-F1*. Predikcija je takođe rađena i korišćenjem *tf-idf* vektorizacije teksta koji predstavlja ocenu tehničkog i HR intervjua, ali je ostvareni rezultat bio lošiji.

F. Značaj aspekata kompanije



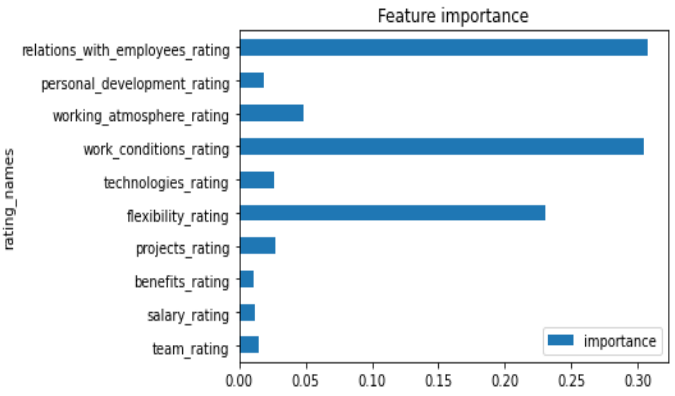
Slika 3.4 Matrica korelacija obeležja kompanija

Ovom problemu smo pristupili kao regresionom problemu. Tretirali smo prosečnu ocenu kompanija kao istinitu labelu, dok su nam ulazi u model bili ocene aspekata kompanije, poput procenta ocena plate, flekisiblnosti, odnosa sa drugim zaposlenima i ostali. Na slici 3.4 se može videti matrica korelacija za data obeležja sa ciljnom varijablom. Kako su vrednosti u rasponu od jedan do pet, korelacija je obično visoka. Vidimo da su top tri obeležja po jačini pozitivne korelacije: odnos sa drugim zaposlenima, uslovi rada i fleksibilnost. Kao regresioni model smo koristili Nasumične šume (eng. *Random Forest*, u nastavku). Kao kriterijum podele je korišćen *gini* nečistoća (eng. *impurity*). Hiperparametri su optimizovani pomoću mrežaste pretrage (eng. grid search), i to sledeći parametri:

- **n_estimators** (broj stabala, u rasponu 5 do 20)
- **max_depth** (maksimalna dubina stabla, u rasponu od dva do devet)

Korišćena je mrežasta pretraga sa unakrsnom validacijom. Unakrsna validacija je rađena nad pet delova (eng. folds), i na kraju je izvučen samo najbolji model. Metrika evaluacija je MSE (eng. mean squared error). Najbolji model postiže rezultat od

Na slici 3.5 su prikazani aspekti kompanije i njihova važnost. Vidimo da su top tri aspekta upravo ona koja imaju najjaču pozitivnu korelaciju sa ocenom kompanije. Rezultat ovog zadatka nam potencijalno ukazuje na prioritet zaposlenih pri vrednovanju svog radnog mesta. Neki od neočekivanih rezultat su da su projekti, tehnologije i plata mnogo manje važni od fleksibilnosti ili odnosa sa kolegama.



Slika 3.5 Značaj različitih aspekata kompanije pri predikciji ocene

G. Eksplorativna analiza

U ovom poglavlju bavili smo se zanimljivim upitima koji mogu biti korisni korisnicima sajta Joberty.rs. Prvi od upita odnosio se na kompanije u Srbiji koje uopšte imaju pozicije u okviru *Data Science/Machine Learning* sektora. Došli smo do spiska od dvadesetak kompanija:

Microsoft Development Center Serbia, SmartCat, Doob Innovation Studio, Endava, Fevo, Grid Dynamics, Htec, Jaggaer, Levi9 Technology Services, Merkle, Nordeus, ogranak Torchlight Technology Group, Ogury, Robert Bosch, Seven bridges, Synechron, Telenor, Telesign, Tymeshift, Umlaut, Vivify Ideas, Zuhlke Engineering

Sledeće pitanje odnosilo se na prosečan nivo plate po senioritetima. Uradili smo upite za top 10 kompanija po prosečnoj plati za juniore, mediore i seniore. U tabeli 3.5 možemo videti top kompanija po prosečno plati za juniorske pozicije.

Kompanija	Pozicija - Junior	Prosečna plat a (EUR)
Microsoft Development Center Serbia	Software Enfineer	2195
Decenter	Software Engineer	2027
Daon	Ostalo	2000
TomTom	Software Engineer	1800
Lotusflare	Software Engineer	1725
Zuhlke Engineering	Project Manager	1637
Quectel	Software Engineer	1633
Symphony	Software Developer	1600
Nordeus	Software Developer	1541

Tabela 3.5 Top deset kompanija sa najvišim prosečnim platama za juniorske pozicije

Preostala dva upita za najviše mediorske i seniorske plate takođe na prvom mestu imaju kompaniju Microsoft sa prosečnim platama za ove pozicije od 3908 i 800 EUR, redom.

Ono što bi moglo pomoći u potrazi za prvim poslom je i lista kompanija za koje je najveći broj korisnika ocenio da je težina intervjua laka odnosno teška. U tabeli 3.6 nalazi se spisak 5 kompanija čiji intervjui imaju najveću procentualnu vrednost za kategoriju lak. Upit je izvršen nad kompanijama koje imaju više ili jednako sa 20 korisničkih recenzija.

Kompanija	Lak intervju (%)	Broj recenzija
Team Sava	64	22
IGT	60	21
Ncr Corporation	57	22
Daon	52	21
Jaggaer	50	31

Tabela 3.6 Top 5 kompanija čiji intervjui su označeni labelom lak

V. Zaključak

U ovom radu opisana je primena raznih tehnika mašinskog učenja nad podacima prikupljenih sa sajta Joberty.rs. Kreiran je skup podataka o IT firmama u Srbiji. Izvšeno je nenadgledano klasterovanje kompanija na osnovu ocena njihovih aspekata (uslovi rada, tehnologije..), analiza sentimenta komentara o procesu selekcije i određivanje značaja svakog aspekta kompanije. Kreiran je model za regresioni problem: predikcija procenta zaposleni koji smatraju da je plata fer sa tačnošću 0.1 MSE. Predikcija težine intervjua je rešavana kao problem *multi class* klasifikacije i ostvaren je rezultat od 0.62 na micro F1 metrici koristeći *XGBClassifier*. Sprovedeno je nekoliko upita nad podacima koji mogu predstavljati izvor novih informacija.

Ove studije mogu poslužiti kao koristan alat za kompanije kako bi povećali zadovoljstvo zaposlenih, podigli moral i zagarantovali produktivnost. Sa druge strane, i oni koji traže posao moći će da iskoriste rezultate ovih istraživanja kako bi pronašli najbolje poslodavce u svom domenu interesovanja. Poslednje, ali ne i najmanje važno, kreirani skup podataka mogao bi da

posluži u nekim drugim istraživanjima koja se tiču IT tržišta u Srbiji.

Literatura

[1] N. Zata Dina, N. Juniarta, Aspect based Sentiment Analysis of Employee's Review Experience, *Journal of Information Systems Engineering and Business Intelligence*, *6*(1), 79–88

[2] R. Bajpai, D. Hazarika, K. Singh, S. Gorantla, E. Cambria, R. Zimmermann, Aspect-Sentiment Embeddings for Company Profiling and Employee Opinion Mining

[3] B. Gaye, D. Zhang, A. Wulamu, Sentiment classification for employees reviews using regression vectorstochastic gradient descent classifier (RVSGDC), *School of Computer and Communication Engineering, University of Science and Technology, Beijing, China*