Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.85
DOI: https://doi.org/10.24867/22BE22Perovic

АНАЛИЗА ПОДАТАКА СА ПЛАТФОРМЕ *JOBERTY.RS* ПОМОЋУ ТЕХНИКА МАШИНСКОГ УЧЕЊА

DATA ANALYSIS OF *JOBERTY.RS* DATA USING MACHINE LEARNING TECHNIQUES

Митар Перовић, *Факултет техничких наука, Нови Сад*

Област – ЕЛЕКТРОТЕХНИКА И РАЧУНАРСТВО

Кратак садржај – *Од великог значаја за компаније јесте да имају увид у ставове и мишљења како свог колектива, тако и оних из других фирми. Са друге стране, онима који траже посао овакве информације помажу да открију предности и слабости компанија које их интересују. У овом раду креиран је скуп података о српским ИТ фирмама користећи рецензије запослених, бивших запослених и кандидата, као и осталим информацијама са сајта Joberty.rs. Извршене анализе над овим скупом подељене су у три задатка. У првом задатку су одређени најзначајнији аспекти компаније уз посматрање овог задатка као регресионог проблема са просечном оценом компаније као циљном лабелом. Резултати указују на то да су аспекти као што су флексибилност и однос послодавца према запосленима значајнији у односу на коришћене технологије и плату. Извршена је анализа сентимента коментара који се тичу процеса селекције употребом БЕРТ модела. Закључено је да HR и технички интервју имају подједнак утицај на формирање утиска о процесу селекције. Коришћењем модела насумичних шума истрениран је модел који врши предикцију процента запослених који сматрају да је плата фер, на основу оцена бенефита које компанија нуди. Постигнут је резултат од 0.1 у контексту метрике MSE. Над истренираним моделом је одрађена анализа важности улазних обележја.*

Кључне речи: *joberty.rs, анализа задовољства запослених, машинско учење, ит фирме*

Abstract – *Employee satisfaction is essential for any organization to progress successfully and achieve its goals. Companies need to have insight into the attitudes and opinions of both their team and those from other companies. On the other hand, such information helps job seekers to discover the strengths and weaknesses of the companies they are interested in. In this paper, a data set of Serbian IT companies was created using reviews of employees, former employees, candidates, and other information from the Jobery.rs website. The analyzes performed on this set are divided into three tasks. In the first task, the most significant aspects of the company were determined while observing this task as a regression problem with the average rating of the company as the target label. The results indicate that the employer's flexibility and attitude towards the employees are more significant than used technologies and salary. Sentiment analysis of comments regarding the selection process was performed using the BERT model. It was concluded that HR and the technical interview have an equal influence on forming the impression about the selection process. Using the RandomForest model, a model was trained that performs prediction of the percentage of employees who think that the salary is fair, based on the evaluation of the benefits offered by the company. A score of 0.1 MSE was achieved. An analysis of the importance of the input features was performed on the trained model.*

Keywords *joberty.rs, data analysis, machine learning, it companies*

1. УВОД

Анализа задовољства запослених на основу рецензија са интернета постало је могуће захваљујући платформама за рецензирање. Као најпопуларнији репрезент ових платформи, истиче се Америчка компанија *Glassdoor.com*. На овој платоформи је од 2008. године објављено преко 45 милиона рецензија за око 830 хиљада фирми.

Овај рад се бави анализом рецензија запослених и извлачења различитих увида о важности компанијских бенефита, задовољству процесом селекције, као и груписању фирми на основу оцена. За потребе рада креиран је скуп података о српским ИТ фирмама користећи се рецензијама запослених, бивших запослених и кандидата, као и осталим информацијама са сајта *Joberty.rs*. Извршене анализе над овим скупом подељене су у три задатка.

У првом задатку су одређени најзначајнији аспекати компаније, уз посматрање овог задатка као регресионог проблема, где су улазна обележја оцене аспеката компаније, а просечна оцена компаније представља циљну лабелу. Резултати указују на то да су аспекти као што су флексибилност радног места и однос са колегама значајнији за задовољство запослених у односу на коришћене технологије и плату.

Други задатак је предикција процента запослених који сматрају да је плата фер на основу оцена бенефита

НАПОМЕНА:
Овај рад проистекао је из мастер рада чији ментор је била др Јелена Сливка, ванр. проф.

које компанија нуди. Искоришћен је *RandomForest* модел и постигнут је резултат од 0.1 просечне суме квадрата грешке (енг. *Mean Squared Error*, MSE). Након евалуирања модела, анализирани су бенефити компаније на основу важности улазних обележја.

У трећем задатку, извршена је анализа сентимента коментара који се тичу процеса селекције употребом БЕРТ модела. Закључено је да *HR* и технички интервју имају подједнак утицај на формирање утиска о процесу селекције.

У следећем поглављу описана је релевантна литература и еволуција техника за анализу задовољства запослених и битних аспеката компаније. Методологија је описана у поглављу 3 – описано је како су коришћени алгоритми и на који начин су евалуирани експерименти. Експерименти и постигнути резултати су описани у поглављу 4. На самом крају, у поглављу 5, укратко је описана методологија рада са најзначајнијим резултатима и потенцијалним побољшањима.

2. ПРЕТХОДНА РЕШЕЊА

У овом поглављу бавићемо се радовима сличне тематике, као и самом еволуцијом области и значајним резултатима и увидима. Нажалост, како су експерименти у овом раду издељени на више мањих задатака везаних за специфичан скуп података на српском језику, за већину задатака не постоји адекватна литература.

Као најважнији задатак, за који постоји релевантна литература, издвојио се задатак рангирања аспеката компаније по важности. Релевантни радови су тражени на основу тематике (анализа рецензија запослених) и издвојени су на основу године објаве и броја цитата. Радови који су узети у разматрање имају макар 10 цитата и објављени су 2016. године или касније.

Већина наведених радова се ослања на рецензије са сајта *Glassdoor.com,* као најсличнијој платформи са које су преузети подаци за овај рад. Превасходно питање које се поставља је да ли су ове рецензије валидне за озбиљнију анализу задовољства запослених. У раду [1], аутори су показали да су оцене са овог сајта валидно средство процене задовољства запослених. Јавно доступне податке о анализи задовољства запослених за оквирно 40 Америчких државних организација, аутори су упоредили са оценама на *Glassdoor*-у. Крајњи резултат показује да су рецензије добар, али не савршен, медијатор за анкете о задовољству запослених. Дошли су до закључка да нема индиција да корисници користе онлине рецензије као издувни вентил, већ да се држе што ближег искреног рецензирања.

Један од првих битнијих радова који је заснован на рецензијама са *Glassdoor*-а јесте [2]. Аутори су испитивали повезаност задовољства запослених са пословним резултатима компаније. У раду је пре тога показано да се мишљења потрошача и анализа сентимента онлине рецензија, твитова и блогова, могу користити за предикције обима продаје, као и цене деоница неке фирме [2]. Преузимањем 257,454 рецензија са *Glassdoor*-а, прикупили су скуп података. Након прикупљања, извршили су пречишћавање података, уклањање мање битних речи (енг. *stopwords*), као и стеминг (енг. *stemming*). Из рецензија су издвојили компанијске вредности које се највише спомињу, попут: интегритета, тимског рада, иновација, квалитета, заједнице, комуникације и награде. Коришћењем *bag-of-words* приступа, индексирали су све рецензије и извукли број појављивања речи за сваку категорију и кључне речи. На основу кључних речи су рецензије додељене некој од категорија (компанијске вредности). На основу броја појављивања ових вредности у рецензијама су направили регресиони модел. За сваку од фирми су прикупили податке о тржишној вредности и укупној вредности имовине компаније. На основу тих података су израчунали *Q фактор* (тржишна вредност / укупна вредност имовине) и употребили као зависну промењиву. Након уклопљеног регресионог модела, извршили су анализу важности фактора и дошли до закључка да су компанијске вредности тимски рад, поштовање и иновација у позитивној корелацији са циљном промењивом. Мана овог приступа јесте начин додељивања категорија свакој од рецензија и што нису коришћене софистициране технике анализе текста. Међутим, овај рад је отворио врата многим будућим радовима који користе јавно доступне податке са сајтова за рецензирање, наспрам класичних анкета о задовољству запослених.

У каснијем раду, аутори су издвојили десет најбоље оцењених компанија, као и три најгоре оцењене компаније из ИТ сектора на основу текстуалних рецензија са *Glassdoor.com-a* [3]. За анализу текста коришћен је *IBM Watson* [4], који на основу улазне реченице издваја пар кључних фраза. Затим су намапирали фразе на вредности компаније. Најчешће помињани аспекти су: друштвена вредност (задовољство запослених у раду са другима), интересантност (колико су изазовни задаци у оквиру посла, колико се користе нова решења, итд.), апликативна вредност (у којој мери су њихова знања и вештине примењене).

3. МЕТОД

У наредним поглављима изложени су скуп података и начин на који су спроведени експерименти за три експеримента.

3.1. Скуп података

Подаци су прикупљени са повлачењем (eng. *scraping*) са сајта *Joberty.rs.* Прикупљени су подаци о 389 компаније које послују у Србији. Скуп података је подељен у више подскупова, зависно са које странице су прикупљени подаци. Прикупљени су следећи подаци:

- Оцене запослених различитих аспеката компаније попут: услова рада, радне атмосфере, односа послодавца према запосленима, плата, пројекти, итд.

- Коментари о фирми - оцена и текстулани коментар запослених о самој компанији.
- Присутност бенефита које компанија нуди - рад од куће, плаћени курсеви, храна, итд. Поред бенефита, са исте странице прикупљен је и проценат запослених који мисле да је плата фер.
- Минималне, просечне и максималне плате за одређене позиције у оквиру фирме.
- Оцена процеса селекције, оцена тежине интервјуа, текстуални коментари о техничком и *HR* делу интервјуа.

3.2. Одређивање важности аспеката компаније

Проблем је моделован као регресиони проблем, а не класификациони, из разлога што бисмо дискретизацијом података потенцијално изгубили део информација. Циљна лабела је просечна оцена компаније, док су улаз у модел оцене аспеката компаније попут: оцене плате, флекисиблности (баланс између приватног и пословног дела живота), однос послодавца према запосленима и остали. За регресиони модел су трениране насумичне шуме. Критеријум поделе података у стаблима је гини нечистоћа (eng. *gini impurity*). Хиперпараметри су оптимизовани помоћу мрежасте претраге (eng. *grid search*), и то следећи:

- ***n_estimators*** (број стабала, у распону од 5 до 20, са кораком пет)
- ***max_depth*** (максимална дубина стабла, у распону од два до девет)

Коришћена је мрежаста претрага са унакрсном валидацијом. Унакрсна валидација је рађена над десет делова (енг. *folds*), и на крају је издвојен само најбољи модел. Метрика евалуација је *MSE*. Над најбољим моделом је извршена анализа важности обележја (аспеката компаније).

3.3. Предикција процента запослених који мисле да је плата фер

Предикција процента запослених који сматрају да је плата фер је вршена над скупом података о бенефитима сваке компаније. Први корак је био претворити број корисника који потврђују постојање одређеног бенефита за неку компанију у процентуалну вредност. Ово је урађено дељењем те вредности са колоном "*review_count*", која представља укупан број корисника који су оцењивали постојање бенефита за ту компанију. Након тога је испробано седам различитих комбинација колона које представљају улаз у модел. Колоне које имају велики број недостајућих вредности (више од две трећине) су избачене. Комбинације колона су прављене насумично.

Задатак модела је да предиктује проценат корисника који сматрају да је плата фер. Тренирани су модели *Support Vector Regressor*, *XGBoostRegressor*, линеарна регресија и насумичне шуме. Коришћена је унакрсна валидација над 10 делова док су хиперпараметри оптимизовани помоћу мрежасте претраге. Модели су евалуирани и поређени коришћењем *MSE* метрике. Најбољи резултат је остварен коришћењем модела насумичних шума. За модел насумичних шума је извршена претрага следећих хиперпараметара:

- ***n_estimators*** (број стабала, у распону од 4 до 128, са кораком степена двојке)
- ***max_depth*** (максимална дубина стабла, у распону од два до 32, са кораком степена двојке)

3.4. Анализа сентимента коментара о процесу селекције

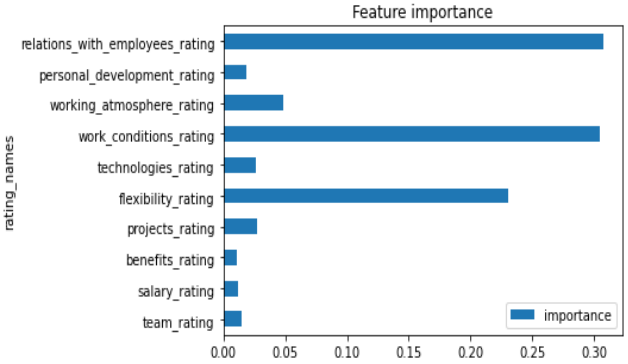
За овај задатак је искоришћен подскуп података који се тиче коментара на сам процес селекције у компанијама. Сваки коментар садржи оцену интервјуа коју је дао кандидат, као и текстуалне коментаре о *HR* и техничком делу интервјуа одвојено. Идеја задатка је установљење који од два дела интервјуа има већи утицај на свеукупан утисак о процесу селекције. Ради реализације задатка рачуната је оцена сентимента коментара за оба дела интервјуа и поређене су са коначном оценом целокупног процеса селекције. За ове потребе коришћен је трансформер модел из *HuggingFace* библиотеке [5]. Конкретније, коришћен је модел "*BERT Multilingual Uncased Sentiment*" из организације "*nlptown*". Након провлачења коментара кроз БЕРТ модел, као предикција је узета оцена од један до пет највеће вероватноће. Разлика је рачуната помоћу *MSE* метрике. Коментар о делу интервјуа који би мање одступао од циљне лабеле, би указивао на потенцијално већи утицај на свеукупан утисак о процесу селекције.

4. РЕЗУЛТАТИ И ДИСКУСИЈА

У овом поглављу су приказани добијени резултати за сва три експеримента и предочене потенцијалне мане и предложена побољшања.

4.1. Одређивање важности аспеката компаније

Након претраге хиперпараметара, најбољи модел постиже резултат од **0.0022** *MSE*. На слици 1 су приказани аспекти компаније и њихова важност. Види се да су три најважнија аспекта управо она која имају најјачу позитивну корелацију са оценом компаније. Резултат овог задатка потенцијално указује на приоритет запослених при вредновању свог радног места. Неки од неочекиваних резултата су места приоритета пројеката, коришћених технологија, као и плата.



Слика 1 *Значај различитих аспеката компаније при предикцији оцене*

Најважнији аспект које компаније треба да негују јесте управо однос према запосленима. Услови рада

су доста широк појам, и запосленима то може бити примарна асоцијација на оцену компаније свеукупно. Добијени резултати могу се додатно поредити са резултатима линеарне регресије и других модела.

4.2. Предикција процента запослених који мисле да је плата фер

Најбољи резултат од **0.1** *MSE* постигнут је помоћу *RandomForestRegressor* модела са хиперпараметрима *max_depth*=4 и *n_estimators*=128. Колоне подскупа података које дају најбољи резултат су „флексибилно радно време", „рад на даљину", „плаћени курсеви", „добијање бонуса" и „осигурање". Закључак је да су дати бенефити најбољи показатељи да ли запослени мисле да је плата фер. Како се најбоље показао модел насумичних шума, над тиме истренираним моделом је одрађена и анализа важности улазних обележја. Као најзначајнији бенефит се показало флексибилно радно време. Након њега следи „рад на даљину", који има сличну важност са бенефитима „плаћени курсеви и обуке", „приватно осигурање" и „добијања бонуса".

4.3. Анализа сентимента коментара о процесу селекције

Након проласка кроз свих 1856 коментара, добијени су резултати да мера *MSE* коментара *HR* дела износи 2.7384, док *MSE* коментара техничког дела интервјуа износи 2.4461. Из приложеног се може већ закључити да је сентимент мање-више конзистентан за оба дела интервјуа, уз незнатну предност *HR* дела као већег фактора на свеукупну оцену.

Недостатак овог приступа лежи у чињеници да не постоји пре-тренирани трансформер модел за овај задатак на српском језику, те у случају појављивања таквог модела, експеримент би требало поновити.

5. ЗАКЉУЧАК

У овом раду описана је примена разних техника машинског учења над подацима прикупљених са сајта *Joberty.rs*. Креиран је скуп података о рецензијама ИТ фирми у Србији. Екстракција знања из рецензија први пут урађена над подацима фирми у Србији, по узору на најпознатију алтернативу - *Glassdoor.com*.

У првом задатку је вршена анализа важности аспеката компанија. Проблем је моделован као регресиони, помоћу модела насумичних шума. За циљну лабелу је коришћена оцена компаније, а улаз у модел су оцене аспеката компаније попут флексибилности, односа послодавца према запосленима, занимљивост пројеката, итд. Резултати указују да је запосленима најважнији аспект односа послодавца према запосленима. Уз овај аспект, најбитнији су општи услови рада и флексибилност, односно баланс приватног и пословног живота. Добијени резултати су доста слични са радовима који су анализе вршили над подацима са *Glassdoor.com* сајта. Проблем представља другачије фразирање сличних аспеката и недовољно објашњење шта тачно представља оцењивани аспект. Резултате експеримента би потенцијално требало поредити са резултатима линеарни регресије или других модела.

Други задатак моделује предикцију проценту запослених који мисле да је плата фер. Улаз у модел су оцене бенефита које компанија нуди попут плаћених курсева, осигурања, итд. Најбољи резултат постигнут је тренирањем *RandomForestRegressor* модела од 0.10 *MSE*. На основу истренираног модела, извршена је анализа важности бенефита компаније. Флексибилно радно време је убедљиво најважнији бенефит, а за њим следе рад на даљину, добијање бонуса, плаћени курсеви и осигурање, наведени сходно важности.

У следећем задатку је спроведена анализа сентимента коментара. Наиме помоћу пре-тренираног БЕРТ модела је извршена анализа коментара о техничком делу процеса селекције, и *HR* делу процеса селекције. Добијена оцена сентимента је поређена са свеукупном оценом процеса селекције коју је кандидат доделио. Поређење је вршено помоћу *MSE* метрике. Резултати указују да не постоји значајна разлика у оценама једног дела интервјуа у односу на оцену целог процеса селекције. Непостојање трансформер модела истренираног на корпусу српског језика, за задатак анализе сентимента, представља проблем и доводи у питање валидност резултата.

Ове студије могу послужити као користан алат за компаније како би повећали задовољство запослених. Такође, креирани скуп података могао би да послужи у неким другим истраживањима која се тичу ИТ тржишта у Србији.

6. ЛИТЕРАТУРА

[1] Landers, Richard N., Robert C. Brusso, and Elena M. Auer. "Crowdsourcing job satisfaction data: Examining the construct validity of Glassdoor. com ratings." Personnel Assessment and Decisions 5, no. 3 (2019): 6.

[2] Luo, Ning, Yilu Zhou, and John Shon. "Employee satisfaction and corporate performance: Mining employee reviews on glassdoor. com." (2016).

[3] Jung, Yeonjae, and Yongmoo Suh. "Mining the voice of employees: A text mining approach to identifying and analyzing job satisfaction factors from online employee reviews." Decision Support Systems 123 (2019): 113074.

[4] High, Rob. "The era of cognitive systems: An inside look at IBM Watson and how it works." *IBM Corporation, Redbooks* 1 (2012): 16.

[5] https://huggingface.co/docs/transformers/index [приступљено: 11.10.2022.]

Кратка биографија:

Митар Перовић рођен је 1998. године у Подгорици. Основне академске студије завршио је 2021. године на Факултету техничих наука у Новом Саду, на ком брани и мастер рад 2022. године из области **Електротехнике и рачунарства** контакт: perovicmitar@uns.ac.rs